SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

001

ERGODIC PROPERTIES OF LINEAR DYNAMICAL SYSTEMS*

RUSSELL A. JOHNSON†, KENNETH J. PALMER‡ AND GEORGE R. SELL§

Abstract. The Multiplicative Ergodic theorem, which gives information about the dynamical structure of a cocycle Φ, or a linear skew product flow π, over a suitable base space $\mathbf{M}$, asserts that for every invariant probability measure μ on $\mathbf{M}$ there is a measurable decomposition of the vector bundle over $\mathbf{M}$ into invariant measurable subbundles, and that every solution with initial conditions in any of these subbundles has strong Lyapunov exponents. These exponents depend on the measure μ, and when μ is ergodic, they are constant (almost everywhere) on $\mathbf{M}$ and form a finite set meas $\Sigma(\mu)$. The dynamical spectrum dyn Σ consists of those values $\lambda \in \mathbf{R}$ for which the shifted flow π_λ fails to have an exponential dichotomy over $\mathbf{M}$. The Spectral theorem for linear skew product flows states that when $\mathbf{M}$ is compact and dynamically connected then dyn Σ is the finite union of k disjoint compact intervals and the vector bundle over $\mathbf{M}$ is the sum of k continuous invariant subbundles. We show that

$$\text{Boundary dyn } \Sigma \subseteq \bigcup_{\mu} \text{meas } \Sigma(\mu) \subseteq \text{dyn } \Sigma$$

where the union above is over all ergodic measures μ on $\mathbf{M}$. Also we show that the measurable invariant subbundles which arise in the Multiplicative Ergodic theorem form a refinement of the continuous invariant subbundles described in the Spectral theorem. A new proof of the Multiplicative Ergodic theorem is presented here. This proof is a substantial simplification over other arguments. Applications of the theory of Lyapunov exponents to "spiral" systems, products of "random" matrices, stochastic differential equations, and the almost periodic Schrödinger operator are included.

Key words. ergodic properties, linear dynamical system, Lyapunov exponents

AMS(MOS) subject classifications. 34C35, 58F11, 58F19

1. Introduction. Nearly two decades ago Oseledec (1968) published his proof of the Multiplicative Ergodic theorem. This theorem, which is one of the milestones in the study of ergodic properties of dynamical systems, has had far-reaching applications, including its role in the work of Margulis (1975) on arithmeticity in Lie groups, in the theory of Pesin (1977) on Bernoullian substructures for diffeomorphisms, in the theory of Katok (1980) on entropy and periodic points, in the study of Kotani (1982) on spectral measures for Schrödinger operators, in the work of Constantin and Foias (1983) on attractors in the Navier–Stokes equations, and in the study of Novikov (1975) and Millionscikov (1978) on almost reducible systems with almost periodic coefficients. As a testimony to the importance of this theorem one finds several alternative proofs including the contemporaneous paper of Millionscikov (1968), and those of Raghunathan (1979), Ruelle (1979) and Crauel (1981), as well as the anticipatory paper of Liao (1966).

The Multiplicative Ergodic theorem gives information about the dynamical structure of a cocycle Φ, or a linear skew product flow π, over a suitable base space $\mathbf{M}$. In typical applications the base space $\mathbf{M}$ is either an attractor, a compact invariant set, or the space of coefficients for a diffeomorphism, a differential equation, or a vector field. This theorem asserts that for every invariant probability measure μ on $\mathbf{M}$ there is a measurable decomposition of the vector bundle over $\mathbf{M}$ into invariant measurable

* Received by the editors September 4, 1985, and in revised form January 10, 1986. This research was supported in part by a grant from the National Science Foundation.

† Department of Mathematics, University of Southern California, Los Angeles, California 90089.

‡ Department of Mathematics and Computer Science, University of Miami, Coral Gables, Florida 33124.

§ Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

subbundles, and that every solution with initial conditions in any of these subbundles has strong Lyapunov exponents. These exponents, or growth rates, depend on the measure μ, and when μ is ergodic, they are constant (almost everywhere) on $\mathbf{M}$ and form a finite set meas $\Sigma(\mu)$, the measurable (Millionscikov-Oseledec) spectrum.

The main objective in this paper is to study the connection between the measurable spectrum meas $\Sigma(\mu)$ and the dynamical spectrum dyn Σ introduced by Sacker and Sell (1975), (1978), (1980). (Also see Daletskii and Krein (1974), as well as Selgrade (1975).) The dynamical spectrum dyn Σ consists of those values $\lambda \in \mathbf{R}$ for which the shifted flow π_λ fails to have an exponential dichotomy over $\mathbf{M}$. It follows from the Spectral theorem for linear skew-product flows, Sacker and Sell (1978), that the dynamical spectrum is the finite union of disjoint compact intervals when $\mathbf{M}$ is compact and dynamically connected.

The dynamical spectrum and the theory of exponential dichotomies are central concepts in wide-ranging branches of analysis including the perturbation theories for invariant manifolds (see Sacker (1969), Fenichel (1971) and Hirsch, Pugh and Shub (1977), the bifurcation theories of Chenciner and Iooss (1979) and Sell (1979), the characterization of the spectrum of Schrödinger operators in Johnson (1982), linearization theories near invariant manifolds in Sell (1984), the study of transversal homoclinic orbits in Palmer (1984), as well as the study of inertial manifolds for dissipative systems in Foias, Sell and Temam (1985).

It is important therefore to understand the connection between these two spectral concepts. We will show, in § 8, that

$$\text{boundary dyn}\,\Sigma \subseteq \bigcup_{\mu} \text{meas}\,\Sigma(\mu) \subseteq \text{dyn}\,\Sigma, \tag{1.1}$$

where the union above is over all ergodic measures μ on $\mathbf{M}$. We actually derive much more than (1.1). We show that the measurable invariant subbundles which arise in the Multiplicative Ergodic theorem form a refinement of the continuous invariant subbundles described in the Spectral theorem. The relationship (1.1) also leads to good methods for computing the Lyapunov exponents and the continuous spectral bundles (see Perry (1986)).

Another objective is to show that the cocycle Φ, itself, has a strong Lyapunov exponent (almost everywhere) and that this agrees with max meas $\Sigma(\mu)$. Although simple, this fact is very important because it forms the foundation for deriving an approximation theory which leads to the numerical evaluation of the measurable and dynamical spectra. The approximation theory and the related numerical coding is described in the University of Minnesota Ph.D. thesis of David Perry (1986).

While doing this investigation we discovered a new proof of the Multiplicative Ergodic theorem. Since our proof is a substantial simplification over other arguments, we present it here. In addition to this simplification, our proof has some interesting geometrical features which may be useful elsewhere. While our proof of the Multiplicative Ergodic theorem is restricted to cocycles over a compact base space $\mathbf{M}$, we will see that this includes practically every application. Among other things, our theory applies to linear stochastic differential equations with bounded measurable coefficients, as well as to the linearized flow near an attractor in a nonlinear dynamical system.

A final objective of this paper is the presentation of several applications of these spectral theories. One of these theories, the theory of Lyapunov exponents for "spiral" systems, is central to any numerical investigation of Lyapunov exponents. Other applications include products of "random" matrices, stochastic differential equations, and the almost periodic Schrödinger operator.

This paper is organized as follows: In § 2 we present the statements of the main theorems in this paper. Section 3 is concerned with a number of technical details which shall be used in the proofs of our theorems. One may wish to skip this on the first reading. In § 4 we present the basic triangularization method as it applies to linear skew product flows. Section 5 is concerned with a brief review of some basic facts about invariant measures, and in § 6 we present our proof of the Multiplicative Ergodic theorem. The ergodic properties of the induced flow on the projective bundle are presented in § 7. In § 8 we derive (1.1) which describes the connection between the measurable and the dynamical spectra. In § 9 we study the theory of wedge-product flows and show how this can be used to compute the measurable spectrum, and in § 10 we present the applications discussed above. The paper concludes with an Appendix which contains some comments on related geometric properties of linear skew product flows.

2. Statement of main theorems. Let **M** be a compact Hausdorff space and let **T** denote either the integers **Z** or the reals **R**. Assume that $\theta \cdot t$ is a flow on **M**, i.e. the mapping $(\theta, t) \to \theta \cdot t$ of $\mathbf{M} \times \mathbf{T}$ into **M** is continuous and satisfies $\theta \cdot 0 = \theta$, and $\theta \cdot (t+s) = (\theta \cdot t) \cdot s$. The Krylov-Bogoliubov theorem (see Nemytskii and Stepanov (1960)) assures us that there is an invariant probability measure μ on **M**. This means that $\mu(A \cdot t) = \mu(A)$ for all Borel sets $A \subseteq \mathbf{M}$ and all $t \in \mathbf{T}$, where $A \cdot t = \{\theta \cdot t \colon \theta \in A\}$. The invariant measure μ is *ergodic* if $\mu(A \triangle A \cdot t) = 0$ for all $t \in \mathbf{T}$ implies that $\mu(A) = 0$ or $\mu(A) = 1$. Recall that $A \triangle B = (A \backslash B) \cup (B \backslash A)$ is the symmetric difference. For an integer $m \geqq 1$ let $\mathcal{GL}(m)$ denote the group of all isomorphisms on $\mathbf{R}^m$, i.e., the group of nonsingular $(m \times m)$ matrices with entries in **R**. A *cocycle* on **M** is a continuous mapping $\Phi \cdot \mathbf{M} \times \mathbf{T} \to \mathcal{GL}(m)$ that satisfies

$$\Phi(\theta, t+s) = \Phi(\theta \cdot t, s)\Phi(\theta, t) \tag{2.1}$$

for all $\theta \in \mathbf{M}$ and $s, t \in \mathbf{T}$. We note that Φ is a cocycle on **M** if and only if

$$\pi(x, \theta, t) := (\Phi(\theta, t)x, \theta \cdot t) \tag{2.2}$$

is a linear skew product flow on $R^m \times \mathbf{M}$.

If $\mathbf{T} = \mathbf{R}$ we shall say that the flow π is *smooth* provided the mapping

$$A\colon \theta \to \frac{d}{dt}\Phi(\theta, t)|_{t=0}$$

exists and is continuous. In this case the cocycle $\Phi(\theta, t)$ is simply the fundamental matrix solution of

$$x' = A(\theta \cdot t)x \qquad (x \in \mathbf{R}^m) \tag{2.3}$$

that satisfies $\Phi(\theta, 0) = I$. This is the prototypical example of a cocycle.

Let Φ and Ψ be two cocycles on **M** with range in $\mathcal{GL}(m)$. We shall say that Φ and Ψ are *cohomologous* if there is a continuous mapping $F\colon \mathbf{M} \to \mathcal{GL}(m)$ that satisfies

$$\Phi(\theta, t) = F(\theta \cdot t)\Psi(\theta, t)F(\theta)^{-1} \tag{2.4}$$

for all $\theta \in \mathbf{M}$, $t \in \mathbf{T}$, where (-1) denotes the matrix inverse.

Let Φ be a cocycle on $\mathbf{M}$. Let $x \in \mathbf{R}^m$ $(x \neq 0)$ and $\theta \in \mathbf{M}$ and define the four *Lyapunov exponents* $\lambda_i^{\pm}(x, \theta)$, $\lambda_s^{\pm}(x, \theta)$ by

$$\lambda_s^{\pm}(x, \theta) = \limsup_{t \to \pm\infty} \frac{1}{t} \log |\Phi(\theta, t)x|, \qquad \lambda_i^{\pm}(x, \theta) := \liminf_{t \to \pm\infty} \frac{1}{t} \log |\Phi(\theta, t)x|.$$

If it happens that the following two limits exist *and* are equal

$$\lim_{t \to +\infty} \frac{1}{t} \log |\Phi(\theta, t)x| = \lim_{t \to -\infty} \frac{1}{t} \log |\Phi(\theta, t)x|, \tag{2.5}$$

then we shall denote the common value as $\lambda(x, \theta)$. *In the future when we write the symbol* $\lambda(x, \theta)$ *this should be interpreted as an assertion that both limits in* (2.5) *exist and* $\lambda(x, \theta)$ *is the common value.* In this case one says that (x, θ) has a *strong* Lyapunov exponent.

Let Φ and Ψ be two cohomologous cocycles on $\mathbf{M}$ that satisfy (2.4), and let $x = F(\theta)y$. Then one has

$$\limsup_{t \to \pm\infty} \frac{1}{t} \log |\Phi(\theta, t)x| = \limsup_{t \to \pm\infty} \frac{1}{t} \log |\Psi(\theta, t)y|$$

and

$$\lim_{|t| \to \infty} \frac{1}{t} \log |\Phi(\theta, t)x| = \lim_{|t| \to \infty} \frac{1}{t} \log |\Psi(\theta, t)y|.$$

In other words, cohomologous cocycles have the same Lyapunov exponents.

For $0 \leqq k \leqq m$ let $\mathscr{G}(m, k)$ denote the Grassman manifold of k-planes in $\mathbf{R}^m$, and let $\mathscr{G}(m) = \bigcup_{k=0}^m \mathscr{G}(m, k)$ denote the disjoint union of these compact manifolds. For $k \in \{1, \cdots, m\}$ we shall let $N(k)$ denote those vectors $\vec{m} = (m_1, \cdots, m_k)$ with $1 \leqq m_i$ and $m_1 + \cdots + m_k = m$.

The first two theorems are statements of the Multiplicative Ergodic theorem.

THEOREM 2.1. *Let* $\mathbf{M}$ *be a compact Hausdorff space with a flow* $\theta \cdot t$ *and let* μ *be an invariant probability measure on* $\mathbf{M}$. *Let* ϕ *denote a cocycle on* $\mathbf{M}$. *Then there exist*:

(i) *an invariant set* $\mathbf{M}_\mu \subseteq \mathbf{M}$ *with* $\mu(\mathbf{M}_\mu) = 1$;

(ii) *a measurable decomposition* $\mathbf{M}_\mu = \bigcup \mathbf{M}_\mu(p)$ *where each* $\mathbf{M}_\mu(p)$ *is invariant and the union is taken over all pairs* $p = (k, \vec{m})$ *where* $1 \leqq k \leqq m$, *and* $\vec{m} \in N(k)$;

(iii) *measurable mappings* $\lambda_1, \cdots, \lambda_k : \mathbf{M}_\mu(p) \to \mathbf{R}$, *where*

$$\lambda_1(\theta) < \lambda_2(\theta) < \cdots < \lambda_k(\theta)$$

for $\theta \in \mathbf{M}_\mu(p)$; *and*

(iv) *measurable mappings* $\mathscr{W}_i : \mathbf{M}_\mu(p) \to \mathscr{G}(m, m_i)$, $1 \leqq i \leqq k$, *where*

$$\vec{m} = (m_1, \cdots, m_k),$$

such that for $\theta \in \mathbf{M}_\mu(p)$ *one has*:

(v) $\{\mathscr{W}_1(\theta), \cdots, \mathscr{W}_k(\theta)\}$ *is linearly independent*;

(vi) $R^m = \mathscr{W}_1(\theta) + \cdots + \mathscr{W}_k(\theta)$;

(vii) *if* $x \in \mathscr{W}_i(\theta)$, $x \neq 0$, *then* $\lambda(x, \theta) = \lambda_i(\theta)$ *for* $1 \leqq i \leqq k$.

If, in addition, μ *is an ergodic measure, then precisely one* $\mathbf{M}_\mu(p)$ *has positive measure, and the mappings* $\lambda_i : \mathbf{M}_\mu \to \mathbf{R}$ *are constant,* $1 \leqq i \leqq k$.

The results in the last theorem extend readily to linear skew product flows on arbitrary vector bundles, Sacker and Sell (1978).

THEOREM 2.2. *Let $\mathscr{E}$ be a vector bundle over a compact Hausdorff space* $\mathbf{M}$ *and let π be a linear skew product flow on $\mathscr{E}$. Let μ be an invariant probability measure on* $\mathbf{M}$. *Then the conclusions of Theorem* 2.1 *remain valid where $\mathscr{W}_i$, $1 \leqq i \leqq k$, now assume values in the appropriate Grassman bundles over* $\mathbf{M}$.

The *measurable spectrum* meas $\Sigma(\mu)$ is defined to be the collection $\{\lambda_1, \cdots, \lambda_k\}$ when μ is ergodic. The numbers $m_1, \cdots, m_k$ are the *multiplicities* of the spectral values $\lambda_1, \cdots, \lambda_k$. When μ is not ergodic, then the spectrum is meas $\Sigma(\mu, \theta) = \{\lambda_1(\theta), \cdots, \lambda_k(\theta)\}$ and the multiplicities $\{m_1, \cdots, m_k\}$ depend on $\theta \in \mathbf{M}_\mu$ and $\theta \in \mathbf{M}_\mu(p)$. For an ergodic measure μ, the *measurable bundle* associated with a spectral value λ_i, $1 \leqq i \leqq k$ is

$$\mathscr{W}_i = \{(x, \theta) \colon x \in \mathscr{W}_i(\theta), \theta \in \mathbf{M}_\mu\}.$$

If μ is not ergodic, then the measurable bundles are defined similarly on each of the invariant sets $\mathbf{M}_\mu(p)$.

The next theorem compares the measurable spectrum and the measurable bundles with the dynamical (or continuous) spectrum and associated continuous spectral subbundles arising in the theory of exponential dichotomies in linear skew product flows; see Sacker and Sell (1978), (1980).

Let $\pi(x, \theta, t)$ be a linear skew product flow on $\mathbf{R}^m \times \mathbf{M}$, where $\mathbf{M}$ is a compact, connected space, and for $\lambda \in \mathbf{R}$ let

$$\pi_\lambda(x, \theta, t) := (\Phi_\lambda(\theta, t)x, \theta \cdot t) \tag{2.6}$$

be the shifted flow, where $\Phi_\lambda(\theta, t) := e^{-\lambda t}\Phi(\theta, t)$. Recall that π_λ has an *exponential dichotomy over* $\mathbf{M}$ if there is a (continuous) projector $P(x, \theta) = (P(\theta)x, \theta)$ on $\mathbf{R}^m$ and constants $K \geqq 1$, $\alpha > 0$ such that

$$|\Phi_\lambda(\theta, t)P(\theta)\Phi_\lambda^{-1}(\theta, s)| \leqq K e^{-\alpha(t-s)}, \qquad s \leqq t,$$

$$|\Phi_\lambda(\theta, t)[I - P(\theta)]\Phi_\lambda^{-1}(\theta, s)| \leqq K e^{-\alpha(s-t)}, \qquad t \leqq s$$

for all $\theta \in \mathbf{M}$ and $s, t \in \mathbf{T}$. The set $\lambda \in \mathbf{R}$ for which π_λ fails to have an exponential dichotomy over $\mathbf{M}$ is defined to be dyn Σ, the dynamical spectrum. The Spectral theorem (Sacker and Sell (1978)) assures us that dyn $\Sigma = \bigcup_{i=1}^k [a_i, b_i]$ is the union of k-nonoverlapping compact intervals, where $1 \leqq k \leqq m$. Also corresponding to each spectral interval $[a_i, b_i]$ there is an invariant spectral subbundle $\mathscr{V}_i$ of $\mathbf{R}^m \times \mathbf{M}$ with dim $\mathscr{V}_i(\theta) \geqq 1$, where $\mathscr{V}_i(\theta) = \{x \in \mathbf{R}^m \colon (x, \theta) \in V_i\}$, $1 \leqq i \leqq k$, the spaces $\{\mathscr{V}_1(\theta), \cdots, \mathscr{V}_k(\theta)\}$ are linearly independent and $\mathbf{R}^m \times \mathbf{M} = \mathscr{V}_1 + \cdots + \mathscr{V}_k$ (as a Whitney sum). The boundary of dyn Σ is the finite collection of end points $\{a_1, \cdots, a_k, b_1, \cdots, b_k\}$. The next theorem, which describes the connection between the measurable and dynamical spectra, is proved in § 7.

THEOREM 2.3. *Let π be a linear skew product flow on $\mathbf{R}^m \times \mathbf{M}$ where $\mathbf{M}$ is compact and connected, and let* dyn Σ *denote the dynamical spectrum of π. Then one has*

$$\text{boundary dyn}\,\Sigma \subseteq \bigcup_\mu \text{meas}\,\Sigma(\mu) \subseteq \text{dyn}\,\Sigma \tag{2.7}$$

where the union is either over all invariant probability measures μ on $\mathbf{M}$ *or over all ergodic measures on* $\mathbf{M}$. *Let μ be a given invariant probability measure on* $\mathbf{M}$ *and let* meas $\Sigma(\mu, \theta) = \{\lambda_1(\theta), \cdots, \lambda_k(\theta)\}$ *be the measurable spectrum for $\theta \in \mathbf{M}_\mu(p)$. Then for each λ_j there is precisely one spectral interval $[a_i, b_i]$ with $\lambda_j(\theta) \in [a_i, b_i]$ for all $\theta \in \mathbf{M}_\mu(p)$. Also the associated measurable bundle $\mathscr{W}_j(\theta)$ satisfies $\mathscr{W}_j(\theta) \subseteq \mathscr{V}_i(\theta)$ for all $\theta \in \mathbf{M}_\mu(p)$.*

Finally one has $\mathcal{V}_i(\theta)=\sum \mathcal{W}_j(\theta)$ *for all* $\theta\in \mathbf{M}_\mu(p)$ *where the summation is over all j with* $\lambda_j(\theta)\in[a_i, b_i]$.

If $\mathbf{T}=\mathbf{Z}$, then the last theorem is valid when $\mathbf{M}$ is compact and "dynamically connected," where the latter means that $\mathbf{M}$ cannot be written as the union of two disjoint nonempty closed invariant sets. Also, as in the spirit of Theorem 2.2, we note that Theorem 2.3 extends to linear skew product flows on general vector bundles.

Our next theorem is concerned directly with the problem of computing the measurable spectrum meas $\Sigma(\mu, \theta)$. The point is that one is able to do this without computing the basis elements $e_1, \cdots, e_m$. The key idea here is the notion of a wedge product, cf. Matshushima (1972). For $1\leqq k\leqq m$ let $\Lambda^k\mathbf{R}^m$ denote the vector space generated by all k-fold wedge products $x_1\wedge\cdots\wedge x_k$ where $x_i\in\mathbf{R}^m$, $1\leqq i\leqq k$. Recall that the wedge product $x_1\wedge\cdots\wedge x_k$ is linear in each factor and antisymmetric, i.e. $x\wedge y=-y\wedge x$. If $L:\mathbf{R}^m\to\mathbf{R}^m$ is linear, then this induces a linear mapping Λ^kL on $\Lambda^k\mathbf{R}^m$ by the formula

$$\Lambda^kL(x_1\wedge\cdots\wedge x_k):=Lx_1\wedge\cdots\wedge Lx_k.$$

Since one has $\Lambda^k(LM)=(\Lambda^kL)(\Lambda^kM)$, we see that if $\Phi(\theta, t)$ is a cocycle on $\mathbf{M}$ then $\Lambda^k\Phi(\theta, t)$ is also cocycle, for $1\leqq k\leqq m$.

In the statement of the next theorem reference will be made to the notation of Theorem 2.1. In particular for $\theta\in\mathbf{M}_\mu(p)$ the growth rates

$$\lambda_1(\theta)<\lambda_2(\theta)<\cdots<\lambda_k(\theta)$$

with multiplicities $m_1, \cdots, m_k$ will be rewritten in the form

$$\gamma_1(\theta)\leqq\gamma_2(\theta)\leqq\cdots\leqq\gamma_m(\theta) \tag{2.8}$$

where $\lambda_i(\theta)$ is repeated m_i-times in (2.8), $1\leqq i\leqq k$.

THEOREM 2.4. *Let* $\mathbf{M}$ *be a compact Hausdorff space with a flow* $\theta\cdot t$ *and let* μ *be an invariant probability measure on* $\mathbf{M}$. *Let* Φ *denote a cocycle on* $\mathbf{M}$ *and adopt the conclusions and notation of Theorem* 2.1. *Let* $\gamma_1, \cdots, \gamma_m$ *satisfy* (2.8) *for* $\theta\in\mathbf{M}_\mu(p)$. *Then for all* $\theta\in\mathbf{M}_\mu(p)$ *one has*:

(i) $\lim_{t\to+\infty}(1/t)\log|\Phi(\theta, t)|=\gamma_m(\theta)$,

(ii) $\lim_{t\to+\infty}(1/t)\log|\Lambda^k\Phi(\theta, t)|=\gamma_{m+1-k}(\theta)+\cdots+\gamma_m(\theta)$, *for* $2\leqq k\leqq m$,

(iii) $\lim_{t\to-\infty}(1/t)\log|\Phi(\theta, t)|=\gamma_1(\theta)$,

(iv) $\lim_{t\to-\infty}(1/t)\log|\Lambda^k\Phi(\theta, t)|=\gamma_1(\theta)+\cdots+\gamma_k(\theta)$, *for* $2\leqq k\leqq m$.

The last theorem extends to linear skew product flows on a vector bundle $\mathscr{E}$ over a compact Hausdorff space $\mathbf{M}$. In this case the wedge product of vectors $x_1, \cdots, x_k\in\mathscr{E}(\theta)$ forms a new bundle $\Lambda^k\mathscr{E}$, $1\leqq k\leqq m$ over $\mathbf{M}$. Also the flow π on $\mathscr{E}$ induces a flow $\Lambda^k\pi$ on $\Lambda^k\mathscr{E}$. This extension is a direct consequence of the proof of the last theorem together with Lemma 3.4 below. We will omit the details.

Remark 2.1. For simplicity of exposition we have formulated these theorems for cocycles with values in $\mathscr{GL}(m, \mathbf{R})$. The theorems are valid for cocycles with values in $\mathscr{GL}(m, \mathbf{C})$, and the proofs we give below extend with only trivial modifications.

3. Some technicalities. Before we turn our attention to the proofs of the main theorems, we need to dispense with some technical details which will enable us to simplify our arguments. We begin with a proof of the following facts:

1. One can assume, without loss of generality, that the base space $\mathbf{M}$ is a compact metric space instead of a compact Hausdorff space. (This fact simplifies substantially some of the measure theoretic considerations.)

2. If $\mathbf{T}=\mathbf{R}$ one can assume that the cocycle Φ is the fundamental solution matrix of an ordinary differential equation on $\mathbf{M}$ with continuous coefficients. We call such a cocycle *smooth.*

3. A linear skew product flow on an arbitrary vector bundle $\mathscr{E}$ over a compact Hausdorff space $\mathbf{M}$ can be imbedded into a linear skew product flow on $\mathbf{R}^m\times\mathbf{M}$ for some $m\geqq 1$.

The argument in each of these three cases is based on the same principle, viz. one can show that the given flow is cohomologous to the desired flow. The resulting cohomology preserves all the desired properties of our main theorems. In particular if Φ_1 and Φ_2 are two cohomologous cocycles on a compact Hausdorff space $\mathbf{M}$ that satisfy

$$F(\theta\cdot t)\Phi_1(\theta,t)=\Phi_2(\theta,t)F(\theta)$$

where $F:\mathbf{M}\to\mathscr{GL}(m)$ is continuous, then as noted above Φ_1 and Φ_2 have the same collection of Lyapunov exponents. Furthermore (x,θ) has a strong Lyapunov exponent for Φ_1 if and only if $(F(\theta)x,\theta)$ has a strong Lyapunov exponent for Φ_2. Thus F preserves the measurable spectrum meas $\Sigma(\mu,\theta)$ and it maps the measurable bundles of Φ_1 onto those of Φ_2. In addition F preserves the dynamical spectrum, and it sets up a one-to-one correspondence between the continuous spectral subbundles.

The situation is, in fact, more general. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two compact Hausdorff spaces and let $f:\mathbf{M}_1\to\mathbf{M}_2$ be a flow epimorphism. Next let Φ_i be a cocycle on $\mathbf{M}_i$ $(i=1,2)$ and let $F:\mathbf{M}_1\to\mathscr{GL}(m)$ satisfy

$$F(\theta_1\cdot t)\Phi_1(\theta_1,t)=\Phi_2(f(\theta_1),t)F(\theta_1).$$

Then Φ_1 and Φ_2 have the same measurable and dynamical spectra and F sets up a one-to-one correspondence between the associated spectral bundles.

Our first step is to show that we can replace a compact Hausdorff base space $\mathbf{M}_1$ with a compact metric space $\mathbf{M}_2$. We use an argument of Ellis (1969).

LEMMA 3.1. *Let Φ_1 be a cocycle over a compact Hausdorff space $\mathbf{M}_1$ with a flow $\theta_1\cdot t$. Then there is* (i) *a compact metric space $\mathbf{M}_2$ with a flow $\theta_2\cdot t$,* (ii) *a flow epimorphism $f:\mathbf{M}_1\to\mathbf{M}_2$ and* (iii) *a cocycle Φ_2 over $\mathbf{M}_2$ such that $\Phi_1(\theta_1,t)=\Phi_2(f(\theta_1),t)$.*

Proof. Let $\{t_n\}$ be a countable dense subset of $\mathbf{T}$. Let $\mathscr{A}$ be the closed subalgebra of $\mathscr{C}(\mathbf{M}_1,\mathbf{R})$ generated by all functions of the form $\{\theta_1\to\phi_{ij}(\theta_1,t_n)\}$ where ϕ_{ij} are the components of Φ. Then $\mathscr{A}$ is a separable subalgebra of $\mathscr{C}(\mathbf{M}_1,\mathbf{R})$. Since $\mathscr{A}$ is closed it contains all mappings $\{\theta_1\to\phi_{ij}(\theta_1,\tau)\}$ where $\tau\in T$; in fact $\mathscr{A}$ is also the closed subalgebra generated by all such mappings for $\tau\in T$. Because of the cocycle identity (2.1) we see that $\mathscr{A}$ is invariant, in the sense that if $g\in\mathscr{A}$ then $g_\tau\in\mathscr{A}$, where $g_\tau(\theta_1)=g(\theta_1\cdot\tau)$. The Stone theorem, cf. Hewitt and Ross (1963, pp. 483-484), says that $\mathscr{A}=\mathscr{C}(\mathbf{M}_2,\mathbf{R})$, where $\mathbf{M}_2$ is the maximal ideal space of $\mathscr{A}$. Since $\mathscr{A}$ is separable, $\mathbf{M}_2$ is a compact metric space. Recall that $\mathbf{M}_2$ can be realized as the space of equivalence classes $[\theta_1]$ where $\theta_1\sim\tilde{\theta}_1$ provided $\phi_{ij}(\theta_1,t_n)=\phi_{ij}(\tilde{\theta}_1,t_n)$ for all i,j and all t_n. Note that if $\theta_1\sim\tilde{\theta}_1$ then $\theta_1\cdot\tau\sim\tilde{\theta}_1\cdot\tau$ for all $\tau\in T$. Consequently a flow on $\mathbf{M}_2$ is given by $[\theta_1]\cdot\tau=[\theta_1\cdot\tau]$. Also the mapping $f(\theta_1):=[\theta_1]$ from $\mathbf{M}_1$ to $\mathbf{M}_2$ is an epimorphism because $\mathscr{A}$ is invariant. Finally we see that for each $t\in\mathbf{T}$ the cocycle $\Phi_1(\theta_1,t)$ depends only on the equivalence class $[\theta_1]$. So we conclude the proof by defining Φ_2 by $\Phi_2([\theta_1],t):=\Phi_1(\theta_1,t)$. Q.E.D.

The following lemma appears in Ellis and Johnson (1982), but we include a proof for completeness of exposition.

LEMMA 3.2. *Let Φ be a cocycle over a compact Hausdorff space $\mathbf{M}$ with $\mathbf{T}=\mathbf{R}$. Then Φ is cohomologous to a smooth cocycle Ψ over $\mathbf{M}$, i.e. $\Psi(\theta,t)$ is a fundamental matrix solution to $x'=A(\theta\cdot\tau)x$ where A is given by* (3.1) *below.*

Proof. Let $V \subseteq \mathcal{GL}(m)$ be a compact convex neighborhood of the identity I and choose $r>0$ so that $\Phi(\theta, t) \in V$ for all $\theta \in \mathbf{M}$ and $0 \leqq t \leqq r$. Define $F(\theta) :=$ $(1/r)\int_0^r \Phi(\theta, s)\, ds$. Then $F(\theta)$ is invertible, and it is easily verified that the cocycle

$$\Psi(\theta, t) := F(\theta \cdot t)\Phi(\theta, t)F(\theta)^{-1} = \frac{1}{r}\int_t^{t+r} \Phi(\theta, s)\, ds\, F(\theta)^{-1},$$

which is cohomologous to Φ, is the fundamental matrix solution to $x' = A(\theta \cdot t)x$ where

$$A(\theta) := \frac{1}{r}[\Phi(\theta, r) - I]F(\theta)^{-1}. \tag{3.1}$$

Q.E.D.

The next lemma will allow us to conclude that the Multiplicative Ergodic theorem 2.2 is valid for linear skew product flows on a vector bundle $\mathscr{E}$ over a compact Hausdorff space $\mathbf{M}$. The same lemma shows that Theorems 2.3 and 2.4 extend to vector bundles as well. Before stating this we need to derive the following general fact concerning smooth approximations to continuous mappings on a compact invariant set.

LEMMA 3.3. *Let* $\mathbf{M}$ *be a compact Hausdorff space with a flow* $\theta \cdot t$ *and let* $f : \mathbf{M} \to \mathbf{N}$ *be a continuous mapping where* $\mathbf{N}$ *is a smooth compact Riemannian manifold. Then for every* $\delta > 0$ *there is a continuous function* $g : \mathbf{M} \to \mathbf{N}$ *with the following properties*:

(i) $\sup\{\operatorname{dist}(f(\theta), g(\theta)) \colon \theta \in \mathbf{M}\} \leqq \delta$,

(ii) *for every* $\theta \in \mathbf{M}$, *the mapping* $\theta \to (d/dt)g(\theta \cdot t)|_{t=0}$ *of* $\mathbf{M}$ *into the tangent bundle* $\mathbf{TN}$ *is a continuous mapping in* θ.

Proof. Let $\delta > 0$ be given. The Tubular Neighborhood theorem, see Guillemin and Pollack (1974), assures us that for a sufficiently large $m \geqq 1$ there is a smooth imbedding $h : \mathbf{N} \to \mathbf{R}^m$, an open set $W \supseteq h(\mathbf{N})$ and a smooth retract $R : W \to h(\mathbf{N})$. Now choose $\eta > 0$ so that if $\phi_1, \phi_2 \in h(\mathbf{N})$ and $|\phi_1 - \phi_2| \leqq \eta$ then $\operatorname{dist}(h^{-1}(\phi_1), h^{-1}(\phi_2)) \leqq \delta$. Next choose $\tau > 0$ so that

$$V(\theta) := \operatorname{Co}\{h(f(\theta \cdot t)) : 0 \leqq t \leqq \tau\} \subseteq W$$

for every $\theta \in \mathbf{M}$, where Co refers to the closed convex hull, and $|h(f(\theta)) - R(y)| \leqq \eta$ for every $\theta \in \mathbf{M}$ and $y \in V(\theta)$. We now define $g : \mathbf{M} \to \mathbf{N}$ by

$$g := h^{-1} \circ R\left(\frac{1}{\tau}\int_0^\tau h \circ f(\theta \cdot s)\, ds\right).$$

Since $1/\tau \int_0^\tau hf(\theta \cdot s)\, ds \in V(\theta)$ we see that $|f(\theta) - g(\theta)| \leqq \delta$ for all $\theta \in \mathbf{M}$. Furthermore it is easy to conclude that g is C^1 along trajectories and the mapping $\theta \to (d/dt)g(\theta \cdot t)|_{t=0}$ is continuous. Q.E.D.

LEMMA 3.4. *Let* $\mathscr{E}$ *be a finite dimensional vector bundle over a compact Hausdorff base space* $\mathbf{M}$ *and let* $\pi(x, \theta, t) = (\Phi(\theta, t)x, \theta \cdot t)$ *be a linear skew product flow on* $\mathscr{E}$. *Then for any* $\lambda \in R$ *there exists an integer* $m \geqq 1$, *a monomorphism* $H : \mathscr{E} \to \mathbf{R}^m \times \mathbf{M}$, *a smooth cocycle* $\Psi : \mathbf{M} \to \mathcal{GL}(m)$ *and an orthogonal invariant resolution of the identity* $Q = (Q_1, Q_2)$ *such that* $H(\mathscr{E}) = \text{Range } Q_1$ *and*

$$Q_1(\theta \cdot t)\Psi(\theta, t) = \Psi(\theta, t)Q_1(\theta), \qquad H(\theta \cdot t)\Phi(\theta, t) = \Psi(\theta, t)H(\theta),$$

$$Q_2(\theta \cdot t)\Psi(\theta, t) = \Psi(\theta, t)Q_2(\theta) = e^{\lambda t}Q_2(\theta).$$

Proof. Since $\dim \mathscr{E}(\theta)$ is constant on the components of $\mathbf{M}$, there is no loss in generality in assuming that $\dim \mathscr{E}(\theta) = k$ for all $\theta \in \mathbf{M}$. The first step is to apply a standard result in the theory of vector bundles, Atiyah (1967, p. 25), which states that there is an integer $m > 0$ and a projector $P_1 : \mathbf{R}^m \times \mathbf{M} \to \mathbf{R}^m \times \mathbf{M}$ such that the vector bundle Range P_1 is isomorphic to $\mathscr{E}$. Let $\hat{H} : \mathscr{E} \to \text{Range } P_1 \subseteq \mathbf{R}^m \times \mathbf{M}$ be the isomorphism. Without any loss of generality we can assume that $P_1(\theta)$ is an orthogonal projection on $\mathbf{R}^m$ for all $\theta \in \mathbf{M}$. Let $P_2 := I - P_1$.

The mapping $\hat{W}_1: \theta \to \text{Range } P_1(\theta)$ defines a continuous mapping of $\mathbf{M}$ into the smooth manifold $\mathscr{G}(m, k)$ of k-planes in $\mathbf{R}^m$. By Lemma 3.3, there is a smooth mapping $W_1: \mathbf{M} \to \mathscr{G}(m, k)$ that is close to $\hat{W}_1$. Define $Q_1(\theta)$ to be the orthogonal projection with $W_1(\theta) = \text{Range } Q_1(\theta)$. Since W_1 is smooth this means that $\theta \to (d/dt)Q_1(\theta \cdot t)|_{t=0}$ is continuous. Also $Q_2(\theta) = I - Q_1(\theta)$ is smooth. Since Q_1 is close to P_1 it follows that $H = Q_1\hat{H}$ is an isomorphism of $\mathscr{E}$ onto Range Q_1.

Next we define a flow on $\mathbf{R}^m \times \mathbf{M}$ under which Q_1 and Q_2 are invariant. Let $S(\theta) = Q_1'(\theta)Q_1(\theta) + Q_2'(\theta)Q_2(\theta)$ and let $\Psi_1(\theta, t)$ be the fundamental matrix solution of $x' = S(\theta \cdot t)x$ satisfying $\Psi_1(\theta, 0) = I$. Then as shown by Daletskii and Krein (1974) one has

$$Q_i(\theta \cdot t)\Psi_1(\theta, t) = \Psi_1(\theta, t)Q_i(\theta)$$

for all $\theta \in \mathbf{M}$, $t \in \mathbf{R}$, and $i = 1, 2$. Define a cocycle Ψ on $\mathbf{M}$ by

$$\Psi(\theta, t)Q_1(\theta) = H(\theta \cdot t)\Phi(\theta, t)H^{-1}(\theta),$$

$$\Psi(\theta, t)Q_2(\theta) = e^{\lambda t}\Psi_1(\theta, t)Q_2(\theta).$$

It is now straightforward to check the remaining details. Lemma 3.2 assures us that Ψ can be chosen to be smooth. Q.E.D.

4. Triangularization of cocycles. We turn next to the theory of the Gram-Schmidt factorization of isomorphisms on $\mathbf{R}^m$, where $\mathbf{R}^m$ has the Euclidean inner product $\langle\ ,\ \rangle$. Let $\mathscr{GL}(m)$ denote the group of all isomorphisms of $\mathbf{R}^m$. Each element $L \in \mathscr{GL}(m)$ is identified with the $(m \times m)$ matrix whose column vectors satisfy $\text{col}_i\, L = Le_i$, $1 \leqq i \leqq m$, where $\{e_1, \cdots, e_m\}$ is a fixed orthonormal basis in $\mathbf{R}^m$. Let $\mathscr{O} = \mathscr{O}(m)$ denote the subgroup of $\mathscr{GL}(m)$ consisting of all orthogonal linear transformations, and let $\mathscr{T}^+(m)$ denote the subcollection of all upper triangular matrices $L \in \mathscr{GL}(m)$ with positive entries on the main diagonal. Then $\mathscr{T}^+(m)$ is also a subgroup of $\mathscr{GL}(m)$ and one has

$$\mathscr{O}(m) \cap \mathscr{T}^+(m) = \{I\}. \tag{4.1}$$

The Gram-Schmidt orthogonalization process assures us that for every $A \in \mathscr{GL}(m)$ there are unique matrices $G(A) \in \mathscr{O}(m)$ and $T(A) \in \mathscr{T}^+(m)$ such that

$$G(A) = AT(A). \tag{4.2}$$

Since the entries in $T(A)$ are algebraic functions of $\langle \text{col}_i\, A, \text{col}_j\, A\rangle$ we see that both $T(A)$ and $G(A)$ are smooth functions of A.

Next we note that one has

$$G(AB) = G(AG(B)), \qquad T(AB) = T(B)T(ABT(B)). \tag{4.3}$$

In order to prove (4.3), we define $U, V \in \mathscr{O}(m)$ by

$$U := G(AB) = ABT(AB), \qquad V := G(AG(B)) = ABT(B)T(ABT(B)),$$

where (4.2) is used above. One then has

$$U^{-1}V = T(AB)^{-1}T(B)T(ABT(B)) \in \mathscr{O}(m) \cap \mathscr{T}^+(m). \tag{4.4}$$

Hence by (4.1) $U^{-1}V = I$, which proves (4.3).

Let $\Phi: \mathbf{M} \to \mathscr{GL}(m)$ be a cocycle on $\mathbf{M}$. Then (4.2) admits the factorization

$$G(\Phi(\theta, t)U) = \Phi(\theta, t)UT(\Phi(\theta, t)U) \tag{4.5}$$

for every $U \in \mathcal{O}(m)$. This permits us to define a new flow on $\mathbf{H} := \mathbf{M} \times \mathcal{O}(m)$ as follows: Let $\phi := (\theta, U) \in \mathbf{M} \times \mathcal{O}(m)$ and define

$$\phi \cdot t := (\theta \cdot t, G(\Phi(\theta, t)U)). \tag{4.6}$$

LEMMA 4.1. *Equation* (4.6) *defines a flow on* $\mathbf{H} = \mathbf{M} \times \mathcal{O}(m)$.

Proof. It suffices to verify the group property

$$G[\Phi(\theta \cdot t, s)G(\Phi(\theta, t)U)] = G(\Phi(\theta, t+s)U).$$

However, this is an immediate consequence of (2.1) and (4.3). Q.E.D.

We noted in § 2 that the cocycle $\Phi(\theta, t)$ on $\mathbf{M}$ defines a linear skew product flow on $\mathbf{R}^m \times \mathbf{M}$ by $\pi(x, \theta, t) = (\Phi(\theta, t)x, \theta \cdot t)$. By using (4.6) we see that π can be lifted to a new flow $\hat{\pi}$ on $\mathbf{R}^m \times \mathbf{H}$ by

$$\hat{\pi}(x, \phi, t) := (\Phi(\theta, t)x, \phi \cdot t)$$

where $\phi = (\theta, U)$. Let $q: \mathbf{M} \times \mathcal{GL}(m) \to \mathcal{GL}(m)$ and $r: \mathbf{M} \times \mathcal{GL}(m) \to \mathbf{M}$, (or $r: \mathbf{H} \to \mathbf{M}$) be the natural projections. Define $\Psi(\phi, t)$ by

$$\Psi(\phi, t) := q(\phi \cdot t)^{-1}\Phi(\theta, t)q(\phi) = G(\Phi(\theta, t)U)^{-1}\Phi(\theta, t)U \tag{4.7}$$

where the (-1) denotes the matrix inverse and $\phi = (\theta, U)$. Since $\phi \cdot t$ is a flow on $\mathbf{H}$, it follows that Ψ is a cocycle on $\mathbf{H}$, and

$$\tilde{\pi}(x, \phi, t) := (\Psi(\phi, t)x, \phi \cdot t) \tag{4.8}$$

is a linear skew product flow on $\mathbf{R}^m \times \mathbf{H}$ which is cohomologous to $\hat{\pi}$. The following lemma is now an immediate consequence of (4.2) and (4.7).

LEMMA 4.2. *Let* Φ *be a cocycle on* $\mathbf{M}$ *and define* $\phi \cdot t$ *and* $\Psi(\phi, t)$ *by* (4.6) *and* (4.7). *Then one has*

$$\Psi(\phi, t) = T(\Phi(\theta, t)U)^{-1} \in \mathcal{T}^+(m) \tag{4.9}$$

for all $\phi = (\theta, U) \in \mathbf{H}$ *and* $t \in \mathbf{T}$.

Remark 4.1. The triangularization method described above is directly related to the familiar technique developed by Lyapunov (1892), Perron (1930) and Diliberto (1950). Let $\mathbf{T} = \mathbf{R}$ and let $\Phi(\theta, t)$ be a smooth cocycle and (therefore) the fundamental solution matrix of a differential equation

$$x' = A(\theta \cdot t)x, \qquad x \in \mathbf{R}^m, \quad \theta \in \mathbf{M}, \tag{4.10}$$

where A is a continuous $(m \times m)$ matrix valued function defined on $\mathbf{M}$. Then $\Psi(\phi, t)$ is the fundamental solution matrix of

$$y' = B(\phi \cdot t)y, \qquad y \in \mathbf{R}^m, \quad \phi \in \mathbf{H}, \tag{4.11}$$

where $\phi = (\theta, U)$, $B = G^{-1}(AG - G')$, $G = G(\Phi(\theta, t)U)$ and $G' = (d/dt)G$. The change of variables which maps solutions of (4.11) onto those of (4.10) is

$$x = P(t)y = G(\Phi(\theta, t)U)y.$$

Also since the fundamental matrix solution of (4.11) is Ψ, an upper triangular matrix, we see that B is also upper triangular.

Remark 4.2. For $\mathbf{T} = \mathbf{Z}$ this is basically the triangularization method described in Oseledec (1968).

5. Invariant measures. In this section we record for reference a number of known results concerning invariant measures associated with the flows on $\mathbf{M}$ and $\mathbf{H}$. Let r be the natural projection of $\mathbf{H}$ onto $\mathbf{M}$. By (4.6) we see that r is a flow epimorphism, i.e. $r(\phi) \cdot t = r(\phi \cdot t)$.

Because of Lemma 3.1 we see that there is no loss in generality in assuming **M** (and therefore **H**) to be compact metric spaces. The Riesz Representation theorem says that for any compact metric space **M** there is an isomorphism between bounded positive linear functionals l on $\mathscr{C}(\mathbf{M}, \mathbf{R})$ satisfying $l(1) = 1$ with the (regular, positive, Borel, probability) measures μ on **M**, and this isomorphism is given by the formula

$$l(f) = \int_M f(\theta)\mu(d\theta).$$

Hereafter we will interchange freely such functionals and the associated measures and write $\mu(f)$ in place of $l(f)$.

The measure μ is invariant for the flow $\theta \cdot t$ if and only if $\mu(f_\tau) = \mu(f)$ for all $f \in \mathscr{C}(\mathbf{M}, \mathbf{R})$ and all $\tau \in \mathbf{T}$ where $f_\tau(\theta) = f(\theta \cdot \tau)$. Also μ is ergodic if and only if for $f \in \mathscr{L}^1(\mathbf{M}, \mathbf{R})$ one has

$$\mu(f_\tau) = \mu(f) \quad \text{for all } \tau \in \mathbf{T} \Leftrightarrow f \equiv \text{constant}.$$

The Krylov-Bogoliubov method, cf. Nemytskii and Stepanov (1960), is a method for constructing invariant measures. Let us review this for the case $\mathbf{T} = \mathbf{R}$. Let μ be a given measure on **M** and define

$$\mu_T(f) := \frac{1}{T} \int_0^T \mu(f_\tau)\, d\tau \tag{5.1}$$

for $T > 0$. Let $T_n \to +\infty$, and suppose (by choosing a subsequence if necessary) that μ_{T_n} converges weakly to a measure $\hat{\mu}$. Then $\hat{\mu}$ is easily seen to be invariant.

If the original measure μ is a δ-measure, i.e. $\mu(f) = \delta_\theta(f) = f(\theta)$, then (5.1) becomes

$$\mu_T(f) := \frac{1}{T} \int_0^T f(\theta \cdot \tau)\, d\tau. \tag{5.2}$$

Notice that if the original measure μ has support in a closed invariant set $\mathbf{M}_0$, then the induced invariant measure $\hat{\mu}$ has support in $\mathbf{M}_0$ as well.

Let μ be a given invariant measure on **M**. Let $I(\mu)$ denote the collection of all invariant measures ν on **H** that cover μ, i.e. $\nu \in I(\mu)$ if it is invariant and $r(\nu) = \mu$. If μ is an ergodic measure on **M** we let $E(\mu)$ denote the ergodic measures $\nu \in I(\mu)$. By using the Krylov-Bogoliubov method we see that $I(\mu)$ is nonempty. Indeed if l is any measure on $\mathcal{O}(m)$, then $\mu \times l$ is a measure on **H**. Now form

$$(\mu \times l)_T(g) = \frac{1}{T} \int_0^T (\mu \times l)(g_\tau)\, d\tau,$$

and let ν be a resulting invariant measure. In order to show that ν covers μ we need to show that $\nu(f) = \mu(f)$ whenever $f = f(\theta)$ depends only on the coordinate $\theta \in \mathbf{M}$. However in this case one has

$$(\mu \times l)(f_\tau) = \mu(f_\tau) = \mu(f) = (\mu \times l)_T(f)$$

since μ is invariant. Hence the limit ν satisfies $\nu(f) = \mu(f)$. Since $I(\mu)$ is nonempty, compact and convex it has extreme points. The extreme points in $I(\mu)$ are ergodic measures ν when μ is ergodic.

6. Proof of the Multiplicative Ergodic theorem. Throughout this section we will adopt without any loss of generality the following Standing Hypotheses which will lead to a proof of Theorems 2.1 and 2.2: Let $\pi(x, \theta, t) = (\Phi(\theta, t)x, \theta \cdot t)$ be a given linear skew-product flow on the trivial (Lemma 3.4) vector bundle $\mathbf{R}^m \times \mathbf{M}$, where $\mathbf{M}$ is a compact metric space (Lemma 3.1). If $\mathbf{T}=\mathbf{R}$ we assume that $\Phi(\theta, t)$ is smooth and is the fundamental solution matrix (Lemma 3.2) of

$$x' = A(\theta \cdot t)x, \qquad x \in \mathbf{R}^m, \quad \theta \in \mathbf{M}. \tag{6.1}$$

Let $\hat{\pi}(x, \phi, t) = (\Psi(\phi, t)x, \phi \cdot t)$ be the cohomologous triangular flow induced on $\mathbf{R}^m \times \mathbf{H}$ (§ 4). If $\mathbf{T}=\mathbf{R}$ then $\Psi(\phi, t)$ is also smooth and is the fundamental solution matrix of

$$y' = B(\phi \cdot t)y, \qquad y \in \mathbf{R}^m, \quad \phi \in \mathbf{H} \tag{6.2}$$

where B is a continuous upper triangular matrix. Let μ be a given invariant measure on $\mathbf{M}$ and let $\nu \in I(\mu)$ be any invariant measure on $\mathbf{H}$ that covers μ. If μ is ergodic we assume that ν is ergodic (§ 5). Also $r: \mathbf{H} \to \mathbf{M}$ is the natural projection.

We shall say that a point $\theta \in \mathbf{M}$ (or $\phi \in \mathbf{H}$) is a *Lyapunov point for* π (or $\hat{\pi}$) if there are real numbers $\gamma_1, \cdots, \gamma_m$ and a basis $e_1, \cdots, e_m$ of $\mathbf{R}^m$ such that

$$\lambda(e_i, \theta) := \lim_{|t| \to \infty} \frac{1}{t} \log |\Phi(\theta, t)e_i| = \gamma_i, \tag{6.3}$$

$$(\text{or } \lambda(e_i, \phi) := \lim_{|t| \to \infty} \frac{1}{t} \log |\Psi(\phi, t)e_i| = \gamma_i) \tag{6.4}$$

for $1 \leqq i \leqq m$.

Roughly speaking, the Multiplicative Ergodic theorem asserts that there are *many* Lyapunov points (i.e. $\mu(\mathbf{M}_\mu) = 1$) and that they fit together in a measurable manner. As we now show this follows from the triangularization technique described in § 4.

LEMMA 6.1. *Let $\phi = (\theta, U) \in \mathbf{H}$ be a Lyapunov point for $\hat{\pi}$. Then $\theta \in \mathbf{M}$ is a Lyapunov point for π.*

Proof. Choose $\gamma_1, \cdots, \gamma_m$ in $\mathbf{R}$ and a basis $e_1, \cdots, e_m$ in $\mathbf{R}^m$ so that (6.4) is satisfied. Define $f_1, \cdots, f_m$ by $f_i = Ue_i$, $1 \leqq i \leqq m$. Equation (4.7) yields

$$G(\Phi(\theta, t)U)\Psi(\phi, t) = \Phi(\theta, t)U. \tag{6.5}$$

Since $G(\Phi(\theta, t)U)$ is an orthogonal matrix one has

$$|\Phi(\theta, t)f_i| = |\Psi(\phi, t)e_i|, \qquad 1 \leqq i \leqq m.$$

It follows that (6.3) is satisfied with the same γ_i when e_i is replaced by f_i, $1 \leqq i \leqq m$. Q.E.D.

The next lemma is the key step in our proof.

LEMMA 6.2. *Let $\phi = (\theta, U) \in \mathbf{H}$ be fixed. Assume that the diagonal entries $\Psi(\phi, t)$ satisfy*

$$\lim_{|t| \to \infty} \frac{1}{t} \log |\psi_{ii}(\phi, t)| = \gamma_i \tag{6.6}$$

for some constants γ_i, $1 \leqq i \leqq m$. Then ϕ is a Lyapunov point for $\hat{\pi}$ where the growth rates $\gamma_1, \cdots, \gamma_m$ are given by (6.6), *and the associated matrix V of basis vectors $\{e_i, \cdots, e_m\}$ is an upper triangular matrix given by* (6.7) *with $v_{ii} = 1$, $1 \leqq i \leqq m$.*

If $\mathbf{T}=\mathbf{R}$, then $\psi_{ii}(\phi, t) = \exp\left(\int_0^t b_{ii}(\phi \cdot s)\, ds\right)$ where b_{ii}, $1 \leqq i \leqq m$, are the diagonal entries of the triangular matrix B in (6.2). In this case (6.6) becomes

$$\lim_{|t| \to \infty} \frac{1}{t} \log |\psi_{ii}(\phi, t)| = \lim_{|t| \to \infty} \frac{1}{t} \int_0^t b_{ii}(\phi \cdot s)\, ds = \gamma_i, \qquad 1 \leqq i \leqq m.$$

Also if $\mathbf{T}=\mathbf{Z}$, then the diagonal elements of ψ satisfy

$$\psi_{ii}(\phi, t)=\prod_{s=0}^{t-1} \psi_{ii}(\phi \cdot s, 1), \qquad t>0$$

with a similar expression valid for $t<0$. For $t>0$ one has

$$\frac{1}{t}\log|\psi_{ii}(\phi, t)|=\frac{1}{t}\sum_{s=0}^{t-1} \log|\psi_{ii}(\phi \cdot s, 1)|.$$

We see then that for both $\mathbf{T}=\mathbf{R}$ and $\mathbf{T}=\mathbf{Z}$, the limits in (6.6) are time-averages of continuous real-valued functions defined on $\mathbf{H}$. This fact will be used later when we apply the Birkhoff Ergodic theorem.

Proof. The argument we now give applies to any triangular cocycle Ψ over any compact metric space $\mathbf{H}$. We will not use the special form of the flow on $\mathbf{H}$.

Let i satisfy $1\leqq i\leqq m$. For any upper triangular $(m\times m)$ matrix T we let T_i denote the lower-right $(k\times k)$-dimensional block where $k=(m-i+1)$. Thus $T_1=T$ and $T_m=(t_{mm})$. For the matrix B given by (6.2) we let β_i denote the $(m-i)$-dimensional row vector that satisfies

$$B_i=\begin{pmatrix} b_{ii} & \beta_i \\ 0 & B_{i+1}\end{pmatrix}$$

for $1\leqq i\leqq m-1$.

The upper triangular matrix V of basis vectors is obtained by constructing the V_i inductively starting with $V_m=(1)$. Suppose $1\leqq i\leqq m-1$ and that V_{i+1} has been constructed with the properties that its diagonal elements are 1 and

$$\lambda(\operatorname{col}_j(V_{i+1}), \phi)=\gamma_j$$

for $i+1\leqq j\leqq m$. To construct V_i with the corresponding properties we first define $v_{ii}=1$. For $1\leqq i\leqq m-1$ and $i+1\leqq j\leqq m$ we define[1]

$$(6.7)\qquad v_{ij}=v_{ij}(\phi):=\int_\tau^0 \psi_{ii}^{-1}(\phi, s)\beta_i(\phi\cdot s)\Psi_{i+1}(\phi, s)\operatorname{col}_j(V_{i+1}(\phi))\,ds$$

$$\operatorname{col}_j(V_i):=\begin{pmatrix} v_{ij} \\ \operatorname{col}_j(V_{i+1})\end{pmatrix}$$

where

$$\tau=\tau_{ij}:=\begin{cases}\infty & \text{if } \gamma_i>\gamma_j,\\ 0 & \text{if } \gamma_i=\gamma_j,\\ -\infty & \text{if } \gamma_i<\gamma_j.\end{cases}$$

Equations (6.4) and (6.6) and the induction hypothesis imply that for every $\varepsilon>0$ there are constants K_1 and K_2 such that for $t\geqq 0$ one has

$$|\psi_{ii}(\theta, t)|\leqq K_2\exp[(\gamma_i+\varepsilon)t], \qquad |\psi_{ii}^{-1}(\theta, t)|\leqq K_2\exp[(-\gamma_i+\varepsilon)t],$$

$$K_1\exp[(\gamma_j-\varepsilon)t]\leqq|\Psi_{i+1}(\phi, t)\operatorname{col}_j(V_{i+1})|\leqq K_2\exp[(\gamma_j+\varepsilon)t]$$

for $i+1\leqq j\leqq m$. Since $|\beta_i|$ is uniformly bounded on $\mathbf{H}$, it follows that the infinite integral in (6.7) is well defined.

[1] We assume for the moment that $\mathbf{T}=\mathbf{R}$. The modification of our argument needed for the case $\mathbf{T}=\mathbf{Z}$ is described in the last paragraph of the proof.

The variation of constants formula for the block-triangular system $u' = B_i(\phi \cdot t)u$ yields

$$(6.8) \qquad \Psi_i(\phi, t)\operatorname{col}_j(V_i) = \begin{pmatrix} \psi_{ii}(\phi, t)\int_\tau^t \psi_{ii}^{-1}(\phi, s)\beta_i(\phi \cdot s)\Psi_{i+1}(\phi, s)\operatorname{col}_j(V_{i+1})\,ds \\ \Psi_{i+1}(\phi, t)\operatorname{col}_j(V_{i+1}) \end{pmatrix}$$

for $i+1 \leqq j \leqq m$ and $\Psi_i(\phi, t)\operatorname{col}_i(V_i) = \operatorname{col}(\psi_{ii}(\theta, t), 0, \cdots, 0)$. Let $v_{ij}(t)$ denote the first entry in (6.8). One then has $\lambda(\operatorname{col}_i(V_i), \phi) = \gamma_i$. While it is known that

$$(6.9) \qquad \lambda(\operatorname{col}_j(V_i), \phi) = \lambda(\operatorname{col}_j(V_{i+1}), \phi) = \gamma_j$$

for $i+1 \leqq j \leqq m$, cf. Millionscikov (1968), we shall include a proof for completeness. Indeed it follows from (6.8) and the inequalities after (6.7) that there is a constant K such that

$$|v_{ij}(t)| \leqq K \exp[(\gamma_j + 3\varepsilon)t]$$

for $i+1 \leqq j \leqq m$ and $t \geqq 0$. Since ε is arbitrary one has

$$(6.10) \qquad \limsup_{t \to +\infty} \frac{1}{|t|} \log |v_{ij}(t)| \leqq \gamma_j, \qquad i+1 \leqq j \leqq m,$$

and therefore by (6.8) we have

$$\gamma_j = \lim_{t \to +\infty} \frac{1}{t} \log |\Psi_{i+1}(\phi, t)\operatorname{col}_j(V_{i+1})| \leqq \liminf_{t \to +\infty} \frac{1}{t} \log |\Psi_i(\phi, t)\operatorname{col}_j(V_i)|$$

$$\leqq \limsup_{t \to +\infty} \frac{1}{t} \log |\Psi_i(\phi, t)\operatorname{col}_j(V_i)| \leqq \limsup_{t \to +\infty} \frac{1}{t} \log |v_{ij}(t)| \leqq \gamma_j.$$

A similar argument applies as $t \to -\infty$. Also (6.10) is valid as $t \to -\infty$.

This completes the argument for $\mathbf{T} = \mathbf{R}$. If $\mathbf{T} = \mathbf{Z}$ the integrals in (6.7)-(6.8) are replaced by sums. For example by the variation of constants formula in Sacker and Sell (1976b), $v_{ij}(t)$ takes the form

$$v_{ij}(t) = \psi_{ii}(\phi, t) \sum_{s=\tau}^{t-1} \psi_{ii}^{-1}(\phi, s+1)\beta_i(\phi \cdot s)\Psi_{i+1}(\phi, s)\operatorname{col}_j(V_{i+1})$$

where $\beta_i = \beta_i(\phi)$ is the $(m-i)$-dimensional row vector that satisfies

$$\Psi_i(\phi, 1) = \begin{pmatrix} \psi_{ii}(\phi, 1) & \beta_i(\phi) \\ 0 & \Psi_{i+1}(\phi, 1) \end{pmatrix}$$

for all $\phi \in \mathbf{H}$. We will omit the details, which are easily verified. Q.E.D.

LEMMA 6.3. *Let $\phi = (\theta, U)$ satisfy the hypotheses of Lemma 6.2 and let V be the matrix of basis vectors constructed above. Then there are upper triangular $(m \times m)$ matrices $S(\phi, t)$ and $D(\phi, t)$ that satisfy*

(i) $\Psi(\phi, t)V = S(\phi, t)D(\phi, t)$,

(ii) $D(\phi, t) = \operatorname{diag}(\psi_{11}(\phi, t), \cdots, \psi_{mm}(\phi, t))$,

(iii) $\limsup_{|t| \to \infty} (1/|t|) \log |S(\phi, t)| \leqq 0$,

(iv) $\limsup_{|t| \to \infty} (1/|t|) \log |S^{-1}(\phi, t)| \leqq 0$.

Proof. S and D are uniquely determined by (i) and (ii) and

$$S(\phi, t) = \begin{pmatrix} 1 & \psi_{22}^{-1}(\phi, t)v_{12}(t) & \cdots & \psi_{mm}^{-1}(\phi, t)v_{1m}(t) \\ 0 & 1 & \cdots & \psi_{mm}^{-1}(\phi, t)v_{2m}(t) \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

It follows from (6.10) that

$$\limsup_{|t|\to+\infty} \frac{1}{|t|} \log |\psi_{jj}^{-1}(\phi, t) v_{ij}(t)| \leqq 0$$

for $1 \leqq i \leqq m-1$ and $i+1 \leqq j \leqq m$. Parts (iii) and (iv) then follow from the last inequality and the fact that the entries in S^{-1} are polynomials in the entries of S. Q.E.D.

Next we apply the Birkhoff Ergodic theorem which assures us that there are a good supply of Lyapunov points in $\mathbf{H}$.

LEMMA 6.4. *There is a Borel measurable invariant set* $\mathbf{H}_\nu \subseteq \mathbf{H}$ *with* $\nu(\mathbf{H}_\nu)=1$ *and such that every point* $\phi \in \mathbf{H}_\nu$ *is a Lyapunov point.*

Proof. As noted above the limits

$$\lim_{|t|\to\infty} \frac{1}{t} \log |\psi_{ii}(\phi, t)|, \qquad 1 \leqq i \leqq m,$$

are time-averages of continuous functions defined on $\mathbf{H}$. The Birkhoff Ergodic theorem, see Nemytskii and Stepanov (1960), asserts that there is a Borel set $\mathbf{H}_\nu \subseteq \mathbf{H}$ with $\nu(\mathbf{H}_\nu)=1$, and there exist bounded Borel measurable invariant functions $\rho_1, \cdots, \rho_m : \mathbf{H}_\nu \to \mathbf{R}$ with the property that

$$\text{(6.11)} \qquad \lim_{|t|\to\infty} \frac{1}{t} \log |\psi_{ii}(\phi, t)| = \rho_i(\phi), \qquad 1 \leqq i \leqq m,$$

for all $\phi \in \mathbf{H}_\nu$. It then follows from Lemma 6.2 that each $\phi \in \mathbf{H}_\nu$ is a Lyapunov point. Q.E.D.

Let $\rho_1, \cdots, \rho_m$ satisfy (6.11). For each integer k, $1 \leqq k \leqq m$, let $N(k)$ denote the collection of vectors $\vec{m} = (m_1, \cdots, m_k)$ with integer entries that satisfy $1 \leqq m_j$, $1 \leqq j \leqq k$, and $m_1 + \cdots + m_k = m$. We will now construct a measurable decomposition of $\mathbf{H}_\nu$. Fix $\phi \in \mathbf{H}_\nu$. We then note that there is an integer k, $1 \leqq k \leqq m$, and an $\vec{m} \in N(k)$ such that the following two properties hold:

(i) There are exactly k distinct values in the collection $\{\rho_1(\phi), \cdots, \rho_m(\phi)\}$, which we rewrite as $\{\lambda_1(\phi), \cdots, \lambda_k(\phi)\}$ where

$$\text{(6.12)} \qquad \lambda_1(\phi) < \lambda_2(\phi) < \cdots < \lambda_k(\phi).$$

(ii) The cardinality of the set $\{i : 1 \leqq i \leqq m \text{ and } \rho_i(\phi) = \lambda_j(\phi)\}$ is m_j for each j, $1 \leqq j \leqq k$.

We denote the ordered pair $(k, \vec{m})$ briefly by p and define $\mathbf{H}_\nu(p)$ to be the set of all $\phi \in \mathbf{H}_\nu$ to which k and $\vec{m}$ correspond as above. The set $\mathbf{H}_\nu(p)$ is Borel measurable since it is the pullback of the closed set

$$\{(x_1, \cdots, x_m) : x_1 = x_2 = \cdots = x_{m_1} < x_{m_1+1} = \cdots = x_{m_2} < \cdots < x_{m_{k-1}+1} = \cdots = x_{m_k}\}$$

by the Borel measurable function that is the composition of the Borel measurable function $\phi \to (\rho_1(\phi), \cdots, \rho_m(\phi))$ with the continuous function σ that maps $(x_1, x_2, \cdots, x_m)$ onto its permutation $(x_1', x_2', \cdots, x_m')$ where $x_1' \leqq x_2' \leqq \cdots \leqq x_m'$.

It is easy to see that $\mathbf{H}_\nu = \bigcup \mathbf{H}_\nu(p)$, where the union is taken over all such points $p = (k, \vec{m})$, is a measurable decomposition of $\mathbf{H}_\nu$. Since each ρ_i is invariant we see that the sets $\mathbf{H}_\nu(p)$ are invariant. If ν is an ergodic measure, then all but one $\mathbf{H}_\nu(p)$ has ν-measure 0. The following result is a consequence of the measurability and invariance of $\rho_1, \cdots, \rho_m$.

LEMMA 6.5. *The functions* $\lambda_1, \cdots, \lambda_k : \mathbf{H}_\nu(p) \to \mathbf{R}$ *are Borel measurable and invariant.*

Let $\phi \in \mathbf{H}_\nu(p)$ and let $\rho_1, \cdots, \rho_m$ and $\lambda_1, \cdots, \lambda_k$ be given as above. Let $e_1, \cdots, e_m$ be the basis in $\mathbf{R}^m$ constructed in Lemma 6.2. Thus one has $\lambda(e_i, \phi) = \rho_i(\phi)$, $1 \leqq i \leqq m$. For $1 \leqq j \leqq k$ define

$$\mathcal{W}_j(\phi) := \text{Span}\,\{e_i \colon \lambda(e_i, \phi) = \lambda_j(\phi)\}.$$

One then has $\dim \mathcal{W}_j(\phi) = m_j$ and $R^m = \mathcal{W}_1(\phi) + \cdots + \mathcal{W}_k(\phi)$.

The next lemma shows that every $y \in \mathcal{W}_j(\phi)$, $y \neq 0$, has $\lambda_j(\phi)$ as a strong Lyapunov exponent.

LEMMA 6.6. *Let $\phi \in \mathbf{H}$ satisfy the hypothesis of Lemma 6.2. Then for all $y \in W_j(\phi)$, $y \neq 0$, one has*

$$\lim_{|t| \to +\infty} \frac{1}{t} \log |\Psi(\phi, t) y| = \lambda_j(\phi), \qquad 1 \leqq j \leqq k. \tag{6.13}$$

Moreover the limit in (6.13) *is uniform for* $|y| = 1$.

Proof. We will use Lemma 6.3. Fix j and let $y \in \mathcal{W}_j(\phi)$, $x \neq 0$. Then

$$y \in \text{Span}\,\{\text{col}_i\,(V) \colon \lambda_i(\phi) = \lambda_j(\phi)\}.$$

If $\{e_1, \cdots, e_m\}$ is the natural basis in $\mathbf{R}^m$ then $z = V^{-1} y$ satisfies

$$z \in \text{Span}\left\{e_i \colon \lim_{|t| \to \infty} \frac{1}{t} \log |\psi_{ii}(\phi, t)| = \lambda_j(\phi)\right\}.$$

Consequently one has

$$\lim_{|t| \to +\infty} \frac{1}{t} \log |D(\phi, t) z| = \lambda_j(\phi),$$

and the last limit is uniform for $|y| = 1$. Since one has

$$|S^{-1}(\phi, t)|^{-1} |D(\phi, t) z| \leqq |\Psi(\phi, t) V z| = |\Psi(\phi, t) y| \leqq |S(\phi, t)|\,|D(\phi, t) z|$$

and

$$\liminf_{t \to +\infty} \frac{1}{t} \log |S^{-1}(\phi, t)|^{-1} = -\limsup_{t \to +\infty} \frac{1}{t} \log |S^{-1}(\phi, t)| \geqq 0$$

(by Lemma 6.3), we get

$$\begin{aligned} \lambda_j(\phi) &\leqq \liminf_{t \to +\infty} \frac{1}{t} \log |S^{-1}(\phi, t)|^{-1} + \liminf_{t \to +\infty} \frac{1}{t} \log |D(\phi, t) z| \\ &\leqq \liminf_{t \to +\infty} \frac{1}{t} \log |\Psi(\phi, t) y| \leqq \limsup_{t \to +\infty} \frac{1}{t} \log |\Psi(\phi, t) y| \\ &\leqq \limsup_{t \to +\infty} \frac{1}{t} \log |S(\phi, t)| + \limsup_{t \to +\infty} \frac{1}{t} \log |D(\phi, t) y| \\ &\leqq \lambda_j(\phi). \end{aligned}$$

A similar argument applies as $t \to -\infty$. Q.E.D.

LEMMA 6.7. *Let $\phi \in \mathbf{H}_\nu(p)$ and let $y \in \mathbf{R}^m$, $y \neq 0$. Assume that $\lambda(y, \phi) = \gamma$. Then there is a j, $1 \leqq j \leqq k$ such that $\lambda(y, \phi) = \lambda_j(\phi)$ and $y \in \mathcal{W}_j(\phi)$.*

Proof. Since $e_1, \cdots, e_m$ is a basis one has

$$y = \alpha_1 e_1 + \cdots + \alpha_m e_m \tag{6.14}$$

where $\alpha_1, \cdots, \alpha_m$ are scalars. It is an easy exercise to see that one has

$$\lim_{t\to\infty} \frac{1}{t} \log |\Psi(\phi, t)y| = \max \{\rho_i(\phi): 1 \leqq i \leqq m \text{ and } \alpha_i \neq 0\},$$

$$\lim_{t\to-\infty} \frac{1}{t} \log |\Psi(\phi, t)y| = \min \{\rho_i(\phi): 1 \leqq i \leqq m \text{ and } \alpha_i \neq 0\}.$$

Therefore if the two-sided limit $\lambda(y, \phi) = \gamma$ exists, the only nonzero α's in (6.14) are coefficients of basis vectors used to define a single $\mathscr{W}_j(\phi)$. Hence $\lambda(y, \phi) = \lambda_j(\phi)$ and $y \in \mathscr{W}_j(\phi)$. Q.E.D.

Let $\phi_1 = (\theta, U_1)$, $\phi_2 = (\theta, U_2)$ be two points in $\mathbf{H}$ with the same θ-coordinate. From (6.5) one has

$$(6.15) \qquad G(\Phi(\theta, t)U_1)\Psi(\phi_1, t)U_1^{-1} = \Phi(\theta, t) = G(\Phi(\theta, t)U_2)\Psi(\phi_2, t)U_2^{-1}.$$

Therefore if $y \in \mathbf{R}^m$ then $|\Psi(\phi_1, t)y| = |\Psi(\phi_2, t)Vy|$ where $V = U_2^{-1}U_1$. We see that

$$(6.16) \qquad \lambda(y, \phi_1) = \gamma \Leftrightarrow \lambda(Vy, \phi_2) = \gamma.$$

Now assume further that $\phi_1 \in \mathbf{H}_\nu$. Then ϕ_1 is a Lyapunov point by Lemma 6.2 and (6.4). Let $\gamma_1, \cdots, \gamma_m$ be the strong Lyapunov exponents and let $e_1, \cdots, e_m$ be a basis with $\lambda(e_i, \phi_1) = \gamma_i$, $1 \leqq i \leqq m$. It follows from (6.16) that $Ve_1, \cdots, Ve_m$ is a basis for which $\lambda(Ve_i, \phi_2) = \gamma_i$, $1 \leqq i \leqq m$. We have just proved the following result:

LEMMA 6.8. *Let $\phi = (\theta, U) \in \mathbf{H}_\nu$, and let $\gamma_1, \cdots, \gamma_m$ be the set of strong Lyapunov exponents given by Lemma 6.2 and (6.4). Then every point $\hat{\phi} = (\theta, \hat{U})$ in the fiber over θ is a Lyapunov point with precisely the same set of strong Lyapunov exponents.*

By combining Lemmas 6.7 and 6.8 and (6.15) we immediately have the following:

LEMMA 6.9. *Let $\phi_1 = (\theta, U_1)$ and $\phi_2 = (\theta, U_2)$ be two points in $\mathbf{H}_\nu$ with the same θ-coordinate. Then ϕ_1 and ϕ_2 lie in the same set $\mathbf{H}_\nu(p)$, and for $1 \leqq j \leqq k$ one has*

$$(6.17) \qquad \lambda_j(\phi_1) = \lambda_j(\phi_2), \qquad U_1\mathscr{W}_j(\phi_1) = U_2\mathscr{W}_j(\phi_2).$$

Hence $\lambda_j(\phi_1)$ and $U_1\mathscr{W}_j(\phi_1)$ depend only on the θ-coordinate.

By using (6.16) together with Lemma 6.6 we see that if $\phi_1 = (\theta, U_1) \in \mathbf{H}_\nu$ then for any $\phi_2 = (\theta, U_2)$, with the same θ-coordinate, we have

$$\lambda(y, \phi_1) = \lambda(Vy, \phi_2) = \lambda_j(\phi_1)$$

for all $y \in \mathscr{W}_j(\phi_1)$, $y \neq 0$, where $V = U_2^{-1}U_1$.

We now use $r: \mathbf{H} \to \mathbf{M}$ to project $\mathbf{H}_\nu$ and $\mathbf{H}_\nu(p)$ to $\mathbf{M}$. Define $\mathbf{M}_\mu$ and $\mathbf{M}_\mu(p)$ by

$$r(\mathbf{H}_\nu) := \mathbf{M}_\mu, \qquad r(\mathbf{H}_\nu(p)) := \mathbf{M}_\mu(p).$$

Note that since $\mathbf{M}$ and $\mathbf{H}$ are compact metric spaces, and $\mathbf{H}_\nu$ and $\mathbf{H}_\nu(p)$ are Borel measurable sets in $\mathbf{H}$, the images $\mathbf{M}_\mu$ and $\mathbf{M}_\mu(p)$ are μ-measurable sets in $\mathbf{M}$, see Federer (1969, Chap. 2). Furthermore one has $\mu(\mathbf{M}_\mu) = 1$. (Strictly speaking, $\mathbf{M}_\mu$ and $\mathbf{M}_\mu(p)$ depend on the choice of $\nu \in I(\mu)$. Since $\mu(r(\mathbf{H}_\nu)) = \nu(\mathbf{H}_\nu) = 1$ we see that any two such sets $\mathbf{M}_\mu$ agree except on a set of μ-measure 0.)

Let $\phi = (\theta, U) \in \mathbf{H}_\nu(p)$. Then $\theta \in \mathbf{M}_\nu(p)$. Next define

$$\lambda_j(\theta) := \lambda_j(\phi), \quad \mathscr{W}_j(\theta) := U\mathscr{W}_j(\phi), \quad 1 \leqq j \leqq k.$$

From Lemma 6.9 we see that $\lambda_j(\theta)$ and $U(\mathscr{W}_j(\phi))$ depend only on the θ-coordinate. Also from Lemmas 6.5 and 6.9 we see that $\lambda_1, \cdots, \lambda_k: \mathbf{M}_\mu(p) \to \mathbf{R}$ are μ-measurable and invariant. For $\theta \in \mathbf{M}_\mu(p)$ we see that the spaces $\mathscr{W}_1(\theta), \cdots, \mathscr{W}_k(\theta)$ satisfy conclusions (v)–(vii) of Theorem 2.1.

The only point that remains to be proven is that the mappings $\mathscr{W}_i:\mathbf{M}_\mu(p)\to\mathscr{G}(m, m_i)$, $1\leqq i\leqq k$, are μ-measurable. Because of Lemma 6.9 it suffices to show that each $\mathscr{W}_i$ is Borel measurable on $\mathbf{H}_\nu(p)$, $1\leqq i\leqq k$. We will do this by noting that the basis matrix $V=\{e_1,\cdots,e_m\}$ constructed in (6.7) is Borel measurable in ϕ since the coefficients in the integral depend continuously in ϕ, and therefore the integral is measurable[2] in ϕ, cf. Federer (1969). This completes the proof of Theorem 2.1.

Theorem 2.2, the Multiplicative Ergodic theorem on a vector bundle $\mathscr{E}$, follows directly from Lemma 3.4 and Theorem 2.1. In Lemma 3.4 one can choose the $\lambda\in\mathbf{R}$ arbitrarily. A good choice for λ is $\lambda\notin\operatorname{dyn}\Sigma(\mathscr{E})$, where $\operatorname{dyn}\Sigma(\mathscr{E})$ is the dynamical spectrum of the linear skew product flow on $\mathscr{E}$. With this choice one knows that the measurable subbundle associated with λ is Range (Q_2) and is disjoint from Range (Q_1). (See Theorem 8.1 below.)

Remark 6.1. The uniformity described in Lemma 6.6 can be strengthened. Let $\phi=(\theta, U)\in\mathbf{H}_\nu(p)$, let $\lambda_1,\cdots,\lambda_k:\mathbf{H}_\nu(p)\to\mathbf{R}$ be the growth rates with $\lambda_1(\phi)<\cdots<\lambda_k(\phi)$, and let

$$\mathbf{R}^m=\mathscr{W}_1(\phi)+\cdots+\mathscr{W}_k(\phi)$$

be the decomposition of $\mathbf{R}^m$ into the measurable bundles. Then every $y\in\mathbf{R}^m$ can be written uniquely as $y=y_1+\cdots+y_k$ where $y_i\in\mathscr{W}_i(\phi)$, $1\leqq i\leqq k$. Furthermore for $y\neq 0$ one has

$$\lim_{t\to+\infty}\frac{1}{t}\log|\Psi(\phi, t)y|=\lambda_b(\phi), \tag{6.18}$$

$$\lim_{t\to-\infty}\frac{1}{t}\log|\Psi(\phi, t)y|=\lambda_a(\phi), \tag{6.19}$$

where $a=\min\{i: y_i\neq 0\}$ and $b=\max\{i: y_i\neq 0\}$. (See Lemma 6.7.) By using the argument of Lemma 6.6, it is easily seen that the limits in (6.18) and (6.19) are uniform on compact sets of the form

$$\{y\in\mathbf{R}^m: 0<\alpha\leqq|y_b|, |y|\leqq\beta\},\qquad \{y\in\mathbf{R}^m: 0<\alpha\leqq|y_a|, |y|\leqq\beta\}.$$

These considerations extend immediately to the cocycle $\Phi(\theta, t)$ over $\mathbf{M}$, where $\mathscr{W}_i(\phi)$ is replaced by $U\mathscr{W}_i(\phi)$, $1\leqq i\leqq k$. (See Lemma 6.9.)

Remark 6.2. As noted by Oseledec (1968) the uniformity condition in Lemma 6.6 implies that the limits

$$\lim_{|t|\to+\infty}\frac{1}{|t|}\log\beta_i(\theta, t),\qquad 1\leqq i\leqq m, \tag{6.20}$$

exist almost everywhere, where $\beta_1\geqq\beta_2\geqq\cdots\geqq\beta_m$ are the eigenvalues of the positive self-adjoint matrix $\Phi^*(\theta, t)\Phi(\theta, t)$.

Remark 6.3. The basis $e_1,\cdots,e_m$, which we construct in Lemma 6.2, is very closely related to Lyapunov's concept of "regularity" or "biregularity", see Lyapunov (1892) and Bylov et al. (1966). Note that if $\theta\in\mathbf{M}_\mu$ then there are real numbers $\gamma_1<\gamma_2<\cdots<\gamma_k$ and a splitting

$$\mathbf{R}^m=W_1+\cdots+W_k$$

such that if $x\in W_i$, $x\neq 0$, then $\lambda(x, \theta)=\gamma_i$, $1\leqq i\leqq k$, and

$$\sum_{i=1}^{k} m_i\gamma_i=\lim_{|t|\to+\infty}\frac{1}{t}\log|\det(\Phi(\theta, t))|, \tag{6.21}$$

[2] Discontinuities in ϕ can arise from the definition of τ in Lemma 6.2.

where $m_i = \dim W_i$, $1 \leqq i \leqq k$. If Φ is a smooth cocycle, i.e. if Φ is a fundamental solution matrix of (6.1), then (6.21) becomes

$$\sum_{i=1}^{k} m_i \gamma_i = \lim_{|t| \to +\infty} \frac{1}{t} \int_0^t \operatorname{tr} A(\theta \cdot s)\, ds.$$

Also see Vinograd (1956).

Remark 6.4. *Other proofs of the Multiplicative Ergodic theorem.* The proof of Oseledec (1968) uses many features of our argument, including the triangularization method described in § 4 and the theory of regularity described above. Some complication in Oseledec's argument seems to be due to the fact that he used neither (6.7) nor the factorization technique described in Lemmas 6.3 and 6.6. Also, Oseledec did not assume the base space **M** to be a compact metric space, and consequently his proof of the measurability (with respect to θ) of the bundles $\mathcal{W}_i(\theta)$ leaves some unanswered questions.

A portion of the Multiplicative Ergodic theorem was derived by Millionscikov (1968) for the case where μ is an ergodic measure. He constructed the measurable spectrum meas $\Sigma(\mu)$ and showed that it was constant almost everywhere. Equation (6.7) was used by Millionscikov; however, he did not derive Lemma 6.6, nor did he address the question of the measurability of the bundles $\mathcal{W}_i(\theta)$.

Raghunathan (1979), Ruelle (1979), Crauel (1981), and Kifer (1985) give alternative proofs of the Multiplicative Ergodic theorem. Their approach is based on either a theorem of Furstenberg and Kesten (1960) (see § 10) or the Subadditive Ergodic theorem, which was proved by Kingman (1968). Ruelle, for example, first shows that the limits in (6.20) exist almost everywhere. By using the eigenspaces of the associated self-adjoint operator $\Phi^*(\theta, t)\Phi(\theta, t)$, he constructs the measurable subbundles $\mathcal{W}_i(\theta)$.

The proof by Ruelle is more general than ours in that it applies to certain linearized semiflows generated by evolutionary equations on an infinite dimensional Hilbert space. Ruelle does not assume the base space **M** to be compact; instead he uses a logarithmic-boundedness condition on the cocycle Φ. This boundedness condition is automatically satisfied when the base space is compact. As we shall see in § 10, the assumption that **M** be compact is not a serious restriction, since this can be satisfied in practically every application.

7. Flow on the projective bundle. In this section we shall study the ergodic properties of the induced flow on the projective bundle, see Johnson (1978), (1980b) and Crauel (1981).

As in § 6, we let $\phi = (\theta, U) \in \mathbf{H}_\nu(p)$ and let $\lambda_1(\phi) < \cdots < \lambda_k(\phi)$ be the growth rates with multiplicity $\vec{m} = (m_1, \cdots, m_k)$, where $m_1 + \cdots + m_k = m$. By Lemma 6.9 we recall that $\lambda_i(\phi)$ depends only on θ, $1 \leqq i \leqq k$. Next define

$$U_i^{\pm}(\phi) = \operatorname{Span}\left\{ y \in \mathbf{R}^m : y \neq 0 \text{ and } \limsup_{t \to \pm\infty} \frac{1}{|t|} \log |\Psi(\phi, t) y| \leqq \lambda_i(\phi) \right\}.$$

Then $\dim U_i^+(\phi) = m_1 + \cdots + m_i$ and $\dim U_i^-(\phi) = m_i + \cdots + m_k$. Also one has $\mathcal{W}_i(\phi) = U_i^+(\phi) \cap U_i^-(\phi)$, $\dim \mathcal{W}_i(\phi) = m_i$ and $\mathbf{R}^m = \mathcal{W}_1(\phi) + \cdots + \mathcal{W}_k(\phi)$. By Lemma 6.9 we see that $\mathcal{W}_i(\theta) := U\mathcal{W}_i(\phi)$ depends only on θ.

Let $\mathbf{P}^{m-1}(\mathbf{R})$ denote the projective space of lines in $\mathbf{R}^m$ containing the origin, with the usual topology. We define the induced flow $\tilde{\pi}$ on the projective bundle $\mathbf{N} = \mathbf{P}^{m-1}(\mathbf{R}) \times \mathbf{M}$ by $(l, \theta) \cdot t = (\Phi(\theta, t)l, \theta \cdot t)$. (Since Φ is linear it maps lines onto lines.)

If $\mathbf{T} = \mathbf{R}$, we define $f : \mathbf{N} \to \mathbf{R}$ by $f(l, \theta) = \langle A(\theta)x, x \rangle$, where A is the matrix function (6.1), $\langle\ ,\ \rangle$ is the Euclidean inner product on $\mathbf{R}^m$, and $x \in l$ satisfies $|x| = 1$. Then for

$|x|=1$ one has

$$\log|\Phi(\theta,t)x|=\int_0^t f((l,\theta)\cdot s)\,ds. \tag{7.1}$$

If $\mathbf{T}=\mathbf{Z}$, we define $f:\mathbf{N}\to\mathbf{R}$ by $f(\theta,l)=\frac{1}{2}\log\langle A^*(\theta)A(\theta)x,x\rangle$, where $A(\theta)=\Phi(\theta,1)$ and $|x|=1$. Then one has

$$\begin{aligned}\log|\Phi(\theta,t)x|&=\sum_{s=0}^{t-1}f((l,\theta)\cdot s), && t\geqq 1,\\ \log|\Phi(\theta,t)x|&=\sum_{s=1}^{|t|}f((l,\theta)\cdot -s), && t<0.\end{aligned} \tag{7.2}$$

Next define the time-averages

$$f^+(l,\theta)=\limsup_{t\to+\infty}\frac{1}{t}\int_0^t f((l,\theta)\cdot s)\,ds,$$

and

$$f^-(l,\theta)=\liminf_{t\to-\infty}\frac{1}{|t|}\int_0^t f((l,\theta)\cdot s)\,ds$$

when $\mathbf{T}=\mathbf{R}$. (For $\mathbf{T}=\mathbf{Z}$, $f^{\pm}$ are defined similarly by using (7.2).) Then by (7.1) and (7.2) $f^+(l,\theta)$ and $f^-(l,\theta)$ are the Lyapunov exponents $\lambda_s^+(x,\theta)$ and $\lambda_i^-(x,\theta)$, respectively, where $x\in l$, $x\neq 0$. Also the functions $f^{\pm}:\mathbf{N}\to\mathbf{R}$ are Borel measurable.

For $\theta\in\mathbf{M}_\mu(p)=r(\mathbf{H}_\nu(p))$ and $1\leqq i\leqq k$ we define

$$u_i^+(\theta)=\{l\in\mathbf{P}^{m-1}(\mathbf{R})|f^+(l,\theta)\leqq\lambda_i(\theta)\},$$

$$u_i^-(\theta)=\{l\in\mathbf{P}^{m-1}(\mathbf{R})|f^-(l,\theta)\leqq-\lambda_i(\theta)\}.$$

For θ fixed, $u_i^{\pm}(\theta)$ are closed subsets of $\mathbf{P}^{m-1}(\mathbf{R})$, and in fact are the "traces" in $\mathbf{P}^{m-1}(\mathbf{R})$ of the vector subspaces $U_i^{\pm}(\theta)$ of $\mathbf{R}^m$ defined above. This leads us to introduce the space $\mathcal{K}$ of closed subsets of $\mathbf{P}^{m-1}(\mathbf{R})$, with the Hausdorff topology. Thus $F_n\to F$ in $\mathcal{K}\Leftrightarrow$ to each $x\in F$, there corresponds a sequence $x_n\in F_n$ so that $x_n\to x$ in $\mathbf{P}^{m-1}(\mathbf{R})$. Observe that the "trace" of $\mathcal{W}_i(\theta)$ in $\mathbf{P}^{m-1}(\mathbf{R})$ is $u_i^+(\theta)\cap u_i^-(\theta)$.

Fix the pair $p=(k,\vec{m})$ and restrict attention to $\mathbf{M}_\mu(p)$. The following proposition is a direct consequence of the measurability of the exponents $\lambda_1,\cdots,\lambda_k$.

LEMMA 7.1. *For every $r>0$, there is a compact set $Z\subseteq\mathbf{M}_\mu(p)$ such that $\mu(\mathbf{M}_\mu(p)\backslash Z)<r$ and the restriction $\lambda_i|_Z$ is continuous, $1\leqq i\leqq k$.*

We will now show that the functions $u_i^{\pm}$ are μ-measurable. (The measurability of $\mathcal{W}_i$ is also a consequence of this fact.) Consider u_i^+ and $\mathbf{T}=\mathbf{R}$. (The arguments for u_i^- and $\mathbf{T}=Z$ are similar and we will omit them.) Define $g_t(l,\theta):=(1/t)\int_0^t f((l,\theta)\cdot s)\,ds$. Then g_t is continuous on $\mathbf{N}$, and $\limsup_{t\to+\infty}g_t(l,\theta_0)=f^+(l,\theta)$ for all $(l,\theta)\in\mathbf{P}^{m-1}(\mathbf{R})\times\mathbf{M}_\mu(p)$. Let $r>0$ be given, and let $Z\subseteq\mathbf{M}_\mu(p)$ be a compact set with $\mu(\mathbf{M}_\mu(p)\backslash Z)<r$, where λ_i is continuous on Z, $1\leqq i\leqq k$. Choose δ so that

$$0<3\delta<|\lambda_i(\theta)-\lambda_j(\theta)| \tag{7.3}$$

for all $i\neq j$ and $\theta\in Z$. Finally define

$$v_t^+(\theta):=\{l\in\mathbf{P}^{m-1}(\mathbf{R})\,|\,g_t(l,\theta)\in(-\infty,\lambda_i(\theta)+\delta]\}.$$

Then $v_t^+(\theta)\in\mathcal{K}$ for $\theta\in Z$. It is not difficult to verify that $v_t^+:Z\to\mathcal{K}$ is a Borel measurable function. (In general it is not continuous.)

We claim that $v_t^+(\theta) \to u_i^+(\theta)$ in $\mathscr{K}$ as $t \to +\infty$, for each $\theta \in Z$. Assume on the contrary that there is a monotone subsequence $t_k \to +\infty$ and an element $Q \in \mathscr{K}$ such that $Q \neq u_i^+(\theta)$ and $v_{t_k}^+(\theta) \to Q$. Then $u_i^+(\theta) \subseteq Q$, since $l \in u_i^+(\theta) \Rightarrow l \in v_t^+(\theta)$ for large t. On the other hand, let $l \in Q \backslash u_i^+(\theta)$. Since $v_{t_k}^+(\theta) \to Q$ in the Hausdorff topology, there is a sequence $l_k \in v_{t_k}^+(\theta)$ with $l_k \to l$. Therefore l_k is eventually in every neighborhood of l in $\mathbf{P}^{m-1}(\mathbf{R})$. By Lemma 7.2 (below) we conclude that $g_{t_k}(l_k, \theta) \geqq \lambda_i(\theta) + 2\delta$, which contradicts the definition of $v_{t_k}^+(\theta)$.

LEMMA 7.2. *Let $l \in \mathbf{P}^{m-1}(\mathbf{R})$ with $l \notin u_i^+(\theta)$ where $\theta \in Z$. Then given any $\delta > 0$ there is a neighborhood $N(l)$ of l and a $\tau \in \mathbf{T}$ such that for all $t \geqq \tau$ one has $g_t(\tilde{l}, \theta) \geqq \lambda_i(\theta) + 2\delta$ for all $l \in N(\tilde{l})$.*

Proof. We will use the notation of Remark 6.1. Let $x \in l$, $x \neq 0$. Since $l \notin u_i^+(\theta)$ one has $i < b$ where

$$\lim_{t \to +\infty} \frac{1}{t} \log |\Phi(\theta, t) x_b| = \lambda_b(\theta)$$

and $|x_b| = 2\alpha > 0$. Let $N(l)$ be those lines $\tilde{l}$ in $\mathbf{P}^{m-1}(\mathbf{R})$ with the property that $\tilde{x} \in \tilde{l}$, $|\tilde{x}| = 1$, satisfies $|\tilde{x}_b| > \alpha$. The uniformity assertion in Remark 6.1 implies that for every $\beta > 0$ there is a $\tau \in \mathbf{T}$ such that

$$g_t(\tilde{l}, \theta) = \frac{1}{t} \log |\Phi(\theta, t) \tilde{x}| \geqq \lambda_b(\theta) - \beta$$

for all $t \geqq \tau$ and all $\tilde{x} \in \tilde{l} \in N(l)$ with $|\tilde{x}| = 1$. Now set $\beta = \delta$, then the lemma follows from (7.3). Q.E.D.

We see then that u_i^+ is the point-wise limit of a sequence of Borel measurable functions on Z. By the Lusin theorem, it follows that u_i^+ is measurable on $\mathbf{M}_\mu(p)$.

8. Comparison with the continuous spectrum. Let Φ be a cocycle on a compact, connected Hausdorff space $\mathbf{M}$. Let $\operatorname{dyn} \Sigma = \bigcup_{i=1}^k [a_i, b_i]$ be the dynamical spectrum with the corresponding Whitney decomposition of $\mathbf{R}^m \times \mathbf{M}$ into continuous spectral subbundles $\mathbf{R}^m \times \mathbf{M} = \mathscr{V}_1 + \cdots + \mathscr{V}_k$. Let $\mathscr{V}_i(\theta)$ denote the fiber of $\mathscr{V}_i$ in $\mathbf{R}^m$, $1 \leqq i \leqq k$. The following result is proved in Sacker and Sell (1978):

THEOREM 8.1. *The spectral subbundles $\mathscr{V}_i$ are characterized by*

$$V_i(\theta) = \operatorname{Span} \{x \in \mathbf{R}^m \colon x \neq 0 \text{ and } \lambda_s^\pm(x, \theta), \lambda_i^\pm(x, \theta) \in [a_i, b_i]\}$$

for $1 \leqq i \leqq k$, where

$$\lambda_s^\pm(x, \theta) = \limsup_{t \to \pm\infty} \frac{1}{t} \log |\Phi(\theta, t) x|, \qquad \lambda_i^\pm(x, \theta) = \liminf_{t \to \pm\infty} \frac{1}{t} \log |\Phi(\theta, t) x|.$$

We will next give a proof of Theorem 2.3. The essence of the argument is to verify (2.7), i.e.

$$\text{boundary dyn } \Sigma \subseteq \bigcup_\mu \text{meas } \Sigma(\mu) \subseteq \text{dyn } \Sigma,$$

and to show that the measurable subbundle decomposition implied by the Multiplication Ergodic theorem leads to a refinement of the continuous decomposition given by the Spectral theorem.

It follows immediately from Theorem 8.1 that for any invariant probability measure μ on $\mathbf{M}$ one has meas $\Sigma(\mu, \theta) \subseteq \operatorname{dyn} \Sigma$ for all $\theta \in \mathbf{M}_\mu$. In particular one has meas $\Sigma(\mu) \subseteq \operatorname{dyn} \Sigma$ for every ergodic measure μ on $\mathbf{M}$. Furthermore the measurable bundles $\mathscr{W}_j(\theta)$ are contained in the associated continuous spectral bundle $\mathscr{V}_i(\theta)$ when $\lambda_j(\theta) \in [a_i, b_i]$, and $\theta \in \mathbf{M}_\mu$. Since the sum of both the $\mathscr{W}_j(\theta)$'s and the $\mathscr{V}_i(\theta)$'s span $\mathbf{R}^m$ for $\theta \in \mathbf{M}_\mu$, it follows that for all $\theta \in \mathbf{M}_\mu$ one has $\mathscr{V}_i(\theta) = \sum \mathscr{W}_j(\theta)$, where the summation is over all j with $\lambda_j(\theta) \in [a_i, b_i]$, $1 \leqq i \leqq k$. (This also follows from applying Theorem 2.2 to the spectral subbundle $\mathscr{V}_i$.)

It remains to show that if $\beta \in$ boundary dyn Σ then $\beta \in$ meas $\Sigma(\mu)$ for some ergodic measure μ. Let β be an endpoint of one of the spectral intervals $[a_i, b_i]$ in dyn Σ. Let $\mathscr{V}_i$ be the continuous spectral subbundle associated with $[a_i, b_i]$. As noted in Lemma 3.1 there is no loss in generality in assuming $\mathbf{M}$ to be a compact metric space. Let X_i be the trace of $\mathscr{V}_i$ in the projective bundle $\mathbf{N} = \mathbf{P}^{m-1}(\mathbf{R}) \times \mathbf{M}$, i.e.

$$X_i = \{(l, \theta) \colon l \text{ is a line in } \mathscr{V}_i(\theta)\}.$$

Since $\mathscr{V}_i$ is invariant under the flow $\pi(x, \theta, t) = (\Phi(\theta, t)x, \theta \cdot t)$, X_i is invariant under the induced flow $\hat{\pi}$ on $\mathbf{N}$. Also X_i is compact. Let $f(l, \theta)$ be given as in § 7. Recall that the time-average of $f(l, \theta)$ along orbits in $\mathbf{N}$ determines the Lyapunov exponents of the solution $\Phi(\theta, t)x$ where x is on the line l, $x \neq 0$.

Let J be the set of all invariant measures η on X_i. We claim that

$$a_i \leqq \int_{X_i} f \, d\eta \leqq b_i \tag{8.1}$$

for all $\eta \in J$. If, on the contrary, (8.1) is false for some $\eta \in J$, then for η-almost all $(l, \theta) \in X_i$ one has

$$\lim_{|t| \to \infty} \frac{1}{t} \log |\Phi(\theta, t)x| = \hat{f}(l, \theta)$$

where f is an invariant function defined on $\mathbf{N}$ with $\int_{X_i} \hat{f} \, d\eta = \int_{X_i} f \, d\eta$, x is on l, $x \neq 0$, and $\int_{X_i} f \, d\eta \notin [a_i, b_i]$. This implies that $\hat{f}(l, \theta) \notin [a_i, b_i]$ on some invariant set of positive η-measure, which contradicts Theorem 8.1.

Next we claim that there is a measure $\eta \in J$ such that $\int_{X_i} f \, d\eta = \beta$. To see this, assume for definiteness that $\beta = b_i$ is the right endpoint of $[a_i, b_i]$. Recall that J is compact, and that the mapping $\eta \to \int_{X_j} f \, d\eta$ is continuous. Therefore if there is no $\eta \in J$ with $\int_{X_i} f \, d\eta = \beta$, then it follows from (8.1) that there is an $\varepsilon > 0$ such that $\int_{X_j} f \, d\eta \leqq \beta - \varepsilon$ for all $\eta \in J$. It follows from the Krylov–Bogoliubov method described in § 5 that for every $(x, \theta) \in \mathscr{V}_i$ with $x \neq 0$ one has

$$\limsup_{t \to \infty} \frac{1}{t} \log |\Phi(\theta, t)x| = \limsup_{t \to \infty} \frac{1}{t} \int_0^t f(\pi(l, \theta, s)) \, ds \leqq \beta - \varepsilon.$$

It then follows from Sacker and Sell (1978, Lemma 4) that $\beta \notin$ dyn Σ, a contradiction. A similar argument works for $\beta = a_i$.

We want to show next that η can be chosen to be an ergodic measure on X_i. Now fix $\eta \in J$ with $\int_{X_j} f \, d\eta = \beta$. Since X_j is a metric space, we can use the Choquet Representation theorem, Phelps (1966), to find a probability measure A_η on the set of *ergodic* measures E in J such that

$$\int_{X_j} g \, d\eta = \int_E \left(\int_{X_j} g \, d\sigma \right) dA_\eta(\sigma)$$

for each $g \in \mathscr{C}(X_i, \mathbf{R})$. In particular for $f = g$ one has

$$\beta = \int_{X_j} f \, d\eta = \int_E \left(\int_{X_j} f \, d\sigma \right) dA_\eta(\sigma).$$

From (8.1) we see that one cannot have $\int_{X_j} f\,d\sigma < \beta = b_i$ for all $\sigma \in E$. It follows then that there is an ergodic measure $\sigma \in E$ with $\int_{X_j} f\,d\sigma = \beta$.

Finally let μ be the projection of σ to $\mathbf{M}$. Then μ is an ergodic measure on $\mathbf{M}$, and from § 6 we see that $\beta \in \operatorname{meas} \Sigma(\mu)$. This completes the proof of Theorem 2.3. Q.E.D.

Remark 8.1. In general, one cannot find a single ergodic measure μ such that $a_i, b_i \in \operatorname{meas} \Sigma(\mu)$ for all endpoints a_i, b_i, even if $\mathbf{M}$ is minimal. Here is a simple example. According to Furstenberg (1961), there is a discrete flow on the 2-torus $\mathbf{M} = \mathbf{T}^2$ with more than one (in fact uncountably many) ergodic measures. Moreover, there is a continuous function g on $\mathbf{M}$ so that $\int_{\mathbf{M}} g\,d\mu_1 \neq \int_{\mathbf{M}} g\,d\mu_2$ for distinct ergodic measures μ_1, μ_2. Define $\Phi: \mathbf{M} \times \mathbf{Z} \to \mathbf{R}$ by $\Phi(y, 1) = \exp g(y)$. The dynamical spectrum of the cocycle Φ is $[a, b]$, where $a = \inf \int_{\mathbf{M}} g\,d\mu$, $b = \sup \int_{\mathbf{M}} g\,d\mu$, and the inf and sup are taken over all ergodic measures on $\mathbf{M}$. However, the measurable spectrum contains just the point $\{\int_{\mathbf{M}} g\,d\mu\}$ for each ergodic measure μ.

Remark 8.2. If there is only one ergodic measure μ on $\mathbf{M}$, for example if the flow $\theta \cdot t$ on $\mathbf{M}$ is almost periodic, then $\operatorname{meas} \Sigma(\mu)$ is a subset of the $\operatorname{dyn} \Sigma$ and *all* endpoints a_i, b_i of $\operatorname{dyn} \Sigma$ are in $\operatorname{meas} \Sigma(\mu)$. For $m = 2$ we conclude that $\operatorname{meas} \Sigma(\mu) = $ boundary $\operatorname{dyn} \Sigma$. An example in Johnson (1986) shows that for $m = 3$, even if $\mathbf{M}$ is almost periodic, the measurable spectrum need not consist entirely of endpoints a_i, b_i.

9. Computation of the measurable spectrum. Wedge product flows. Let $\theta \in \mathbf{M}$ be a Lyapunov point and let $\gamma_1(\theta) \leqq \cdots \leqq \gamma_m(\theta)$ denote the growth rates with associated basis $e_1, \cdots, e_m$. Thus one has $\lambda(e_i, \theta) = \gamma_i(\theta)$, $1 \leqq i \leqq m$. By a standard argument, see Naylor and Sell (1982, p. 268) for example, there is a constant K such that for any vector $x \in \mathbf{R}^m$ one has $x = \alpha_1 e_1 + \cdots + \alpha_m e_m$ and

$$|\Phi(\theta, t)x| \leqq K \max\{|\Phi(\theta, t)e_i| : \alpha_i \neq 0\}|x\} \tag{9.1}$$

for all $t \in \mathbf{T}$. It follows from (9.1) that

$$\limsup_{t \to +\infty} \frac{1}{t} \log |\Phi(\theta, t)| \leqq \gamma_m(\theta).$$

On the other hand one has

$$|\Phi(\theta, t)e_m| \leqq |\Phi(\theta, t)|\,|e_m|,$$

which implies that

$$\gamma_m(\theta) = \lim_{t \to +\infty} \frac{1}{t} \log |\Phi(\theta, t)e_m| \leqq \liminf_{t \to \infty} \frac{1}{t} \log |\Phi(\theta, t)|,$$

and hence one has

$$\lim_{t \to \infty} \frac{1}{t} \log |\Phi(\theta, t)| = \gamma_m(\theta). \tag{9.2}$$

A similar argument yields

$$\lim_{t \to -\infty} \frac{1}{t} \log |\Phi(\theta, t)| = \gamma_1(\theta). \tag{9.3}$$

An early version of (9.2) for stationary stochastic process of $(m \times m)$ matrices appears in Furstenberg and Kesten (1960).

The same considerations extend to the induced wedge product cocycles $\Lambda^k\Phi(\theta, t)$ on $\Lambda^k\mathbf{R}^m$, where $2 \leqq k \leqq m$. If $\Psi(\phi, t)$ satisfies (4.7), then one has

$$\Lambda^k\Psi(\phi, t) = (\Lambda^k q(\phi \cdot t))^{-1}(\Lambda^k\Phi(\theta, t))(\Lambda^k q(\phi)).$$

Hence the cocycles $\Lambda^k\Psi$ and $\Lambda^k\Phi$ are cohomologous. Therefore $\Lambda^k\Psi$ and $\Lambda^k\Phi$ have the same Lyapunov exponents.

For $1 \leqq k \leqq m$ let Ord (k, m) denote the collection of all strictly monotone mappings $\sigma: \{1, \cdots, k\} \to \{1, \cdots, m\}$. We will use the lexicographic ordering on Ord (k, m); thus $\sigma < \tau$, where $\sigma, \tau \in \text{Ord}\,(k, m)$, provided there is a j, $1 \leqq j \leqq k$ such that $\sigma(i) = \tau(i)$ for $1 \leqq i \leqq j-1$ and $\sigma(j) < \tau(j)$. If $\{e_1, \cdots, e_m\}$ is any basis for $\mathbf{R}^m$, then $\{e_\sigma: \sigma \in \text{Ord}\,(k, m)\}$ is a basis for $\Lambda^k\mathbf{R}^m$ where

$$e_\sigma = e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(k)}. \tag{9.4}$$

Furthermore if T is an upper-triangular $(m \times m)$ matrix (with respect to the basis $\{e_1, \cdots, e_m\}$), then $\Lambda^k T$ is an upper-triangular matrix with respect to the basis e_σ, $\sigma \in \text{Ord}\,(k, m)$. Also the diagonal entry $t_{\sigma\sigma}$ (in the σth position on the diagonal) is given by the product

$$t_{\sigma\sigma} = t_{\sigma(1)\sigma(1)} \cdot \cdots \cdot t_{\sigma(k)\sigma(k)}.$$

If, in addition, one has $t_{ii} > 0$ for $1 \geqq i \geqq m$, then $t_{\sigma\sigma} > 0$ for all $\sigma \in \text{Ord}\,(k, m)$.

Let us return to the triangular cocycle $\Psi(\phi, t)$. It follows from the last paragraph that if $\phi = (\theta, U)$ is fixed and if the diagonal entries of $\Psi(\phi, t)$ satisfy (6.6), then the diagonal entry $\psi_{\sigma\sigma}$ of $\Lambda^k\Psi$ satisfies

$$\lim_{|t|\to\infty} \frac{1}{t} \log |\psi_{\sigma\sigma}(\phi, t)| = \gamma_{\sigma(1)} + \cdots + \gamma_{\sigma(k)}, \tag{9.5}$$

where $\sigma \in \text{Ord}\,(k, m)$ and $2 \leqq k \leqq m$. The collection of numbers given by (9.5), where σ varies over Ord (k, m), represents the Lyapunov exponents of $\Lambda^k\Psi$. This analysis applies for every $\phi \in \mathbf{H}_\nu$, where $\mathbf{H}_\nu$ is given by Lemma 6.4. For $\phi \in \mathbf{H}_\nu(p)$ we shall rewrite the growth rates in the form (2.8) where

$$\gamma_1(\phi) \leqq \gamma_2(\phi) \leqq \cdots \leqq \gamma_m(\phi). \tag{9.6}$$

It then follows from (9.5) that the largest growth rate for $\Lambda^k\Psi$ is

$$\gamma_{m+1-k}(\phi) + \cdots + \gamma_m(\phi)$$

and the smallest is

$$\gamma_1(\phi) + \cdots + \gamma_k(\phi).$$

The argument in the first paragraph in this section now applies to $\Lambda^k\Psi$, which completes the proof of Theorem 2.4.

Remark 9.1. One can give a precise description of the measurable bundles $W_\tau^{(k)}(\phi)$ corresponding to $\Lambda^k\Psi$. Fix $\phi \in \mathbf{H}_\nu(p)$, where the growth rates of Ψ satisfy (9.6), and let $e_1(\phi), \cdots, e_m(\phi)$ be a basis in $\mathbf{R}^m$ that satisfies $\lambda(e_i(\phi), \phi) = \gamma_i(\phi)$, $1 \leqq i \leqq m$. For $\tau \in \text{Ord}\,(k, m)$ we define

$$\gamma_\tau(\phi) = \gamma_{\tau(1)}(\phi) + \cdots + \gamma_{\tau(k)}(\phi).$$

Then one has

$$W_\tau^{(k)}(\phi) = \text{span}\,\{e_\sigma: \lambda(e_\sigma, \phi) = \gamma_\tau(\phi)\}, \tag{9.7}$$

where e_σ is defined by (9.4) for $\sigma \in \text{Ord}\,(k, m)$. We will omit the proof of (9.7), which is a simple application of the techniques developed in § 6.

10. Applications and illustrations. In this section we collect several illustrative examples of the theory described above. Included here is a discussion of spiral systems, products of "random" matrices, conservative second-order Schrödinger equations with almost periodic potentials, and linear stochastic differential equations with bounded measurable coefficients.

(A) *Spiral systems.* The theory above applies to every compact invariant set **N** in **M**. In this case the dynamical spectrum dyn $\Sigma(\mathbf{N})$ depends on **N**. Next we want to study the case where **N** is a single orbit together with its ω-limit set, i.e. a *spiral system.* More precisely let **M** be a compact Hausdorff space with a flow $\theta \cdot t$. Let θ_0 be a given point in **M** and define

$$\mathbf{N} = H^+(\theta_0) = \text{closure}\,\{\theta_0 \cdot t\colon t \geqq 0\}.$$

Then **N** is positively invariant and the ω-limit set $\Omega = \bigcap_{\tau \geqq 0} H^+(\theta_0 \cdot \tau)$ is a compact invariant set. We are interested in the case where $\theta_0 \notin \Omega$. Thus the positive trajectory $\theta_0 \cdot t$ forms a spiral. See Fig. 1.

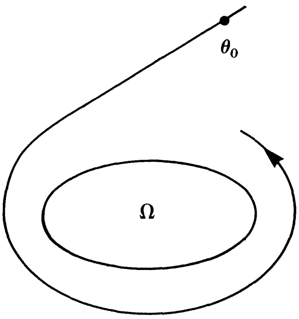


FIG. 1. N: *A spiral system.*

Let $\Phi(\theta, t)$ be a cocycle defined on **M**. Then the theory described above applied directly to the restriction of Φ to the ω-limit set Ω. The problem we wish to study here is the limiting behavior (as $s, \tau \to +\infty$) of the cocycles $\Lambda^k\Phi(\theta_0 \cdot s, t)$, $1 \leqq k \leqq m$, along the spiral trajectory θ_0. In particular we want to show that this limiting behavior can be used to evaluate the measurable spectrum of the ω-limit set Ω.

Before doing the analysis it should be noted that this study addresses a basic question which arises naturally when one is doing a numerical evaluation of the measurable (or dynamical) spectrum. The initial point θ_0 in **M** is determined by the code or program. If by good fortune it happens to lie in the set $\mathbf{M}_\mu$ (Theorem 2.1), then Theorem 2.4 explains how to compute the spectrum. With our present understanding, one does not know whether or not we have had good fortune. However, what is always true is that θ_0 does determine a spiral system.

For $k = 1, \cdots, m$ we define

$$b^k := \limsup_{s,\tau \to +\infty} \frac{1}{\tau} \log |\Lambda^k\Phi(\theta_0 \cdot s, \tau)|.$$

Also let Σ^k denote the dynamical spectrum of the linear skew-product flow $(\Lambda^k\Phi(\theta, t), \theta \cdot t)$ over Ω, and define

$$a^k := \max \Sigma^k.$$

We will now prove the following result:

THEOREM 10.1. *The following statements are valid*:

(A) *For* $k=1,\cdots,m$ *one has* $a^k=b^k$.

(B) *If the flow on* Ω *is uniquely ergodic* (*e.g. almost periodic*) *then* meas $\Sigma=\{\gamma_1,\cdots,\gamma_m\}$ *where* $\gamma_m=b^1$ *and*

$$\gamma_{m-k}=b^{k+1}-(\gamma_m+\cdots+\gamma_{m-k+1}),\qquad 1\leqq k\leqq m-1.$$

Proof. We will give the proof of (A) for $k=1$. The argument for $k\geqq 2$ is the same. Statement (B) is an immediate corollary of part (A) and Theorems 2.1, 2.3 and 2.4.

Let a and b denote a^1 and b^1. For $S\geqq 0$ and $T\geqq 0$ we define

$$\beta(S,T)=\sup_{\substack{S\leqq s\\ T\leqq \tau}}\frac{1}{T}\log|\Phi(\theta_0\cdot s,\tau)|.$$

Then $b=\lim_{S,T\to+\infty}\beta(S,T)$. Fix $\varepsilon>0$ and choose $S\geqq 0$, $T\geqq 0$ so that $\beta(S,T)\leqq b+\varepsilon$. For $t=s+\tau$ one then has

$$|\Phi(\theta_0\cdot s,\tau)|=|\Phi(\theta_0,t)\Phi^{-1}(\theta_0,s)|\leqq K\,e^{(b+\varepsilon)(t-s)},\qquad 0\leqq s\leqq t$$

where $K=\max\{|\Phi(\theta_0\cdot s,\tau)|: 0\leqq s\leqq S,\ 0\leqq\tau\leqq T\}$. It then follows directly, see Sacker and Sell (1974, Thms. 2 and 5) and (1976a, Lemma 4) for example, that $a\leqq b+\varepsilon$ for every $\varepsilon>0$. Hence $a\leqq b$.

If one has $a<b$, then we can replace $\Phi(\theta,t)$ by the shifted flow $e^{\lambda t}\Phi(\theta,t)$ where $3\lambda=2a+b$. This has the effect of shifting a and b to $a-\lambda$ and $b-\lambda$, respectively. Without any loss of generality, therefore, we can assume then that $a<0<b$ and set $b=3\alpha$. Since $a<0$, the linear skew-product flow $(\Phi(\theta,t)x,\theta\cdot t)$ has an exponential dichotomy over Ω with $\mathcal{S}=\mathbf{R}^m\times\Omega$. This means that there is a constant K such that

$$|\Phi(\theta,t)\Phi^{-1}(\theta,s)|\leqq K\,e^{-\sigma(t-s)}$$

for all $\theta\in\Omega$, and $s\leqq t$, $s,t\in\mathbf{T}$. In particular it follows from Sacker and Sell (1974, p. 452) that $\mathcal{B}=\{0\}\times\Omega$, $\mathcal{S}=\mathbf{R}^m\times\Omega$ and $\mathcal{U}=\{0\}\times\Omega$, where $\mathcal{B}$, $\mathcal{S}$ and $\mathcal{U}$ are defined to be those $(x,\theta)\in\mathbf{R}^m\times\Omega$ that satisfy

$$\sup_{t\in\mathbf{T}}|\Phi(\theta,t)x|<\infty,\quad \lim_{t\to+\infty}|\Phi(\theta,t)x|=0\quad\text{and}\quad \lim_{t\to-\infty}|\Phi(\theta,t)x|=0,$$

respectively.

Since $b>0$ there are sequences $s_n\to+\infty$, $\tau_n\to+\infty$ such that

$$|\Phi(\theta_0\cdot s_n,\tau_n)|\geqq e^{\alpha\tau_n}.$$

Let e_n be a vector with $|e_n|=1$ and

$$|\Phi(\theta_0\cdot s_n,\tau_n)e_n|\geqq e^{\alpha\tau_n}.$$

Fix σ_n so that $0\leqq\sigma_n\leqq\tau_n$ and

$$|\Phi(\theta_0\cdot s_n,t)e_n|\leqq|\Phi(\theta_0\cdot s_n,\sigma_n)e_n|,\qquad 0\leqq t\leqq\tau_n.$$

Set $\xi_n=\Phi(\theta_0\cdot s_n,\sigma_n)e_n$. Then $|\xi_n|\geqq e^{\alpha\tau_n}$, and

$$|\xi_n|^{-1}|\Phi(\theta_0\cdot s_n,t)e_n|\leqq 1,\qquad 0\leqq t\leqq\sigma_n.$$

Since $|\xi_n|\to+\infty$ one has $\sigma_n\to+\infty$. Also one has $e_n=\Phi(\theta_n,-\sigma_n)\xi_n$ where $\theta_n=\theta_0\cdot(s_n+\sigma_n)$. By choosing subsequences (if necessary) we can assume that $\theta_n\to\theta\in\Omega$ and $|\xi_n|^{-1}\xi_n\to e$ where $|e|=1$. It follows from Sacker and Sell (1976a, Lemma 4) that $(e,\theta)\in\mathcal{U}$, which contradicts the fact that $\mathcal{U}=\{0\}\times\Omega$. We conclude that $a=b$.

Remark 10.1. The conclusions of Theorem 10.1 can be reformulated in another manner. The numbers a^k represent the "largest possible" Lyapunov exponents for the cocycle $\Lambda^k\Phi(\theta, t)$, where $\theta \in \Omega$ and $1 \leqq k \leqq n$. By the inequality (2.7) and Theorem 2.4 we see that there are ergodic measures μ_k on Ω with the property that a^k is the largest value in the measurable spectrum generated by $\Lambda^k\Phi$, $1 \leqq k \leqq n$. Consequently if μ is any invariant measure on Ω and $\gamma_1(\theta) \leqq \cdots \leqq \gamma_m(\theta)$ are the growth rates satisfying (9.6) for $\theta \in \Omega_\mu$, then one has

$$\gamma_m(\theta) + \cdots + \gamma_{m-k+1}(\theta) \leqq a^k$$

for all $\theta \in \Omega_\mu$.

Remark 10.2. The ergodic measures μ_k referred to in the last paragraph need not be the same. Let Ω be a minimal set in the flow $\theta \cdot t$, and assume that this flow is not uniquely ergodic. Then there are ergodic measures μ_1 and μ_2 on Ω and a continuous real-valued function g that satisfies $\int g\, d\mu_1 \neq \int g\, d\mu_2$. Since g can be replaced by $g + c$ and/or αg, where c and α are constants with $\alpha \neq 0$, we can assume that μ_1, μ_2 and g are chosen to satisfy

$$-1 = \int g\, d\mu_2 \leqq \int g\, d\mu \leqq \int g\, d\mu_1 = 3$$

for every ergodic measure μ. Now set $h = -2g$ and consider the linear skew-product flow on $\mathbf{R}^2 \times \Omega$ generated by

$$x' = \operatorname{diag}(g(\theta \cdot t), h(\theta \cdot t))x.$$

In the notation introduced above one then has $a^1 = 3$, $a^2 = 2$, and $a^i = \max \operatorname{meas} \Sigma(\mu_i)$, $i = 1, 2$.

Remark 10.3. It may happen that all the limits

$$\lim_{t \to +\infty} \frac{1}{t} \log |\Lambda^k\Phi(\theta_0, t)| = c^k, \qquad 1 \leqq k \leqq m,$$

exist. If so, then one has $c^k \leqq a^k$, $1 \leqq k \leqq m$. By using the associated triangular flow and the Krylov-Bogoliubov method described in § 5, one can show that there is an invariant measure μ on Ω with the property that $\operatorname{meas} \Sigma(\mu) = \{\gamma_1, \cdots, \gamma_m\}$ where $\gamma_1 \leqq \cdots \leqq \gamma_m$ and

$$\gamma_{m-k+1} + \cdots + \gamma_m = c^k, \qquad 1 \leqq k \leqq m.$$

Remark 10.4. The conclusions of Theorem 10.1 and the above remarks are related to results of Pelikan (1983) who has analyzed the structure of Bowen-Ruelle invariant measures on certain attractors.

(B) *Products of random matrices.* In this example we show that the theory of products of matrices considered in Furstenberg and Kesten (1960) is often imbeddable in our theory. Let $\mathbf{K}$ denote a fixed compact subset of $\mathscr{GL}(m)$ and let $\{X_1, X_2, \cdots\}$ be a given sequence of $(m \times m)$ matrices with values in $\mathbf{K}$. For $n = 1, 2, \cdots$ form the product

$$Y_n = X_n X_{n-1} \cdots X_1.$$

We wish to study the limiting behavior of

$$\frac{1}{n} \log |Y_n|$$

as $n \to \infty$. This will be done by imbedding this problem into a spiral system, where $Y_n = \Phi(\theta_0, n)$ for an appropriate cocycle Φ, and using Theorem 10.1.

Let $\mathbf{M} := \mathbf{K}^{\mathbf{Z}}$ denote the collection of all two-sided sequences

$$\theta = (\cdots, A_{-2}, A_{-1}; A_0; A_1, A_2, \cdots) \tag{10.1}$$

with entries $A_i \in \mathbf{K}$, $i \in \mathbf{Z}$. (We will use semi-colons to designate the zeroth position of θ.) Then $\mathbf{M}$ is a compact metric space with the shift flow $\theta \cdot n$ where

$$\theta \cdot n = (\cdots, A_{n-2}, A_{n-1}; A_n; A_{n+1}, A_{n+2}, \cdots)$$

for $n \in \mathbf{Z}$. Define $F: \mathbf{M} \to \mathbf{K}$ by $F(\theta) := A_0$ where θ is given by (10.1). Then one constructs a cocycle Φ over $\mathbf{M}$ by defining $\Phi(\theta, 0) = I$

$$\Phi(\theta, n) = F(\theta \cdot (n-1)) \cdots F(\theta), \qquad n \geqq 1,$$

$$\Phi(\theta, n) = [F(\theta \cdot (-1)) \cdots F(\theta \cdot n)]^{-1}, \qquad n \leqq -1.$$

It is not difficult to see that the cocycle identity (2.1) is valid for $t, s \in \mathbf{Z}$.

The distinguished sequence $\{X_1, X_2, \cdots\}$ is imbedded into this flow as a spiral, i.e. let A be a fixed element in $\mathbf{K}$ and set

$$\theta_0 = (\cdots, A, A; X_1; X_2, X_3, \cdots),$$

where every negative entry in θ_0 is A. Then $\theta_0 \in \mathbf{M}$ and $Y_n = \Phi(\theta_0, n)$ for $n \geqq 1$. By Theorem 10.1 we see that

$$\limsup_{m,n \to +\infty} \frac{1}{m} \log |\Phi(\theta \cdot n, m)| = \limsup_{m,n \to +\infty} \frac{1}{m} \log |Y_{m+n} Y_n^{-1}| \tag{10.2}$$

exists and this is the maximum value of the dynamical spectrum over Ω, the ω-limit set of θ_0.

When the distinguished sequence $\{X_1, X_2, \cdots\}$ is a stationary stochastic process, then the expectation satisfies $E(\log^+ |X_1|) < \infty$ since X_1 assumes values in the compact set $\mathbf{K}$. If, in addition, this stochastic process is metrically transitive (i.e. ergodic) then Theorem 2.4 is applicable, and one concludes that

$$\lim_{n \to \infty} \frac{1}{n} \log |Y_n| \tag{10.3}$$

exists with probability 1. Also the limits in (10.2) and (10.3) agree. We refer the reader to Furstenberg and Kesten (1960) for more details.

Remark 10.5. Some interesting applications of products of random matrices to problems in demographics can be found in Cohen (1979).

(C) *Schrödinger equation.* In the study of the Schrödinger equation

$$\mathrm{L}y = \left(-\frac{d^2}{dt^2} + q(t)\right) y = \lambda y,$$

where $q(t)$ is real and Bohr almost periodic, it is of interest to compute the "Lyapunov number" $\beta(\lambda)$ as a function of $\lambda \in R$. $\beta(\lambda)$ is defined as follows: First introduce the *hull* $\mathbf{M}$ of q by

$$\mathbf{M} = \text{closure}\,\{q_\tau \mid \tau \in \mathbf{R}\},$$

where $q_\tau(t) = q(t+\tau)$, and the closure is in the uniform topology. Then $\mathbf{M}$ is a compact metric space with translation flow $\theta \cdot \tau = \theta_\tau$. In fact $\mathbf{M}$ is a compact topological group, and the normalized Haar measure μ is the unique invariant measure on $\mathbf{M}$. Define

$Q(\theta)=\theta(0)$, and consider the operators $L_\theta=-(d^2/dt^2)+Q(\theta\cdot t)$ and the associated equations

$$x'=\begin{pmatrix}0 & 1\\ -\lambda+Q(\theta\cdot t) & 0\end{pmatrix}x,\qquad x=\begin{pmatrix}y\\ y'\end{pmatrix}. \tag{10.4}$$

Since the trace of the coefficient matrix is zero, one obtains from Liouville's formula that

$$\operatorname{meas}\Sigma(\mu)=\{-\beta(\lambda),\beta(\lambda)\}$$

where $\beta(\lambda)\geqq 0$. This defines $\beta(\lambda)$. Also, as noted in Remark 8.2, one has boundary dyn $\Sigma=\operatorname{meas}\Sigma(\mu)$.

Spectral properties of the self-adjoint linear operators L_θ on $L^2(-\infty,\infty)$ are reflected in the dynamics of (10.4). For example, λ is in the resolvent set for L_θ if and only if (10.4) admits an exponential dichotomy, Johnson (1982). Also if $\beta(\lambda)>0$ for all λ in an interval I, then for μ-almost all θ, the (functional analytic) spectrum of L_θ has no absolutely continuous component in I, Pastur (1980), Ishii (1973). Moreover if $\beta(\lambda)=0$ for all λ in I, then I is in the purely absolutely continuous spectrum of L_θ for μ-almost all θ, Kotani (1982).

The numerical computation of $\beta(\lambda)$ when $q(t)=\cos t+\cos\pi t$, for example, is a challenging problem. An investigation of this problem is described in Perry (1986). The basic idea is to use Theorem 2.4 to estimate $\beta(\lambda)$. Also special properties of second order linear equations, as described in Johnson (1980a) and Johnson and Moser (1982), can be exploited to help determine whether or not (10.4) admits an exponential dichotomy for $\lambda=0$. This, in turn, leads to a resolution of the question of whether or not one has dyn $\Sigma=\operatorname{meas}\Sigma(\mu)$.

Another method for computing $\beta(\lambda)$, which was suggested by R. Helleman, is to use the theory of Johnson and Moser (1982). In this setting one extends λ to the complex plane so that for $\operatorname{Im}\lambda>0$, $\beta(\lambda)$ is the real part of a holomorphic function $w(\lambda)$, called the Floquet exponent of (10.4). When $\operatorname{Im}\lambda>0$, (10.4) has an exponential dichotomy. One can compute $\beta(\lambda)$ for real λ by a limiting formula:

$$\beta(\lambda)=\lim\nolimits_{\eta\to 0^+}\beta(\lambda+i\eta).$$

(D) *Linear stochastic differential equations.* An interesting variation of the Schrödinger equation occurs when the potential $q(t)$ is a stochastic variable. More generally consider the m-dimensional case $x'=A(t)x$, $x\in\mathbf{R}^m$, where the entries $a_{ij}(t)$ are stochastic variables in t. We will show how this can be imbedded in a linear skew-product flow on $\mathbf{R}^m\times\mathbf{M}$, where $\mathbf{M}$ is a compact space, under the assumption that the coefficients $a_{ij}(t)$ are bounded and measurable in t, i.e. $a_{ij}\in L^\infty(\mathbf{R})$. (See Kurzweil (1957), Miller and Sell (1970) and Sacker and Sell (1974) for more information.)

The set $\mathbf{M}$ is the hull of A and is defined by

$$\mathbf{M}=\text{closure}\,\{A_\tau\colon \tau\in\mathbf{R}\},$$

where $A_\tau(t)=A(\tau+t)$ and the closure is taken in the "weak L^1-local" topology. That is, a generalized sequence $\{A_n\}$ converges to a limit B if for every $\tau\in\mathbf{R}$ and every $\phi\in L^1[\tau,\tau+1]$ one has

$$\int_\tau^{\tau+1} A_n(t)\phi(t)\,dt\to\int_\tau^{\tau+1} B(t)\phi(t)\,dt.$$

Since $A(t)$ is bounded and measurable, the hull of A is a compact Hausdorff space. If $B\in\mathbf{M}$ we let $\Phi(B, t)$ be the fundamental solution matrix for $x'=B(t)x$. Then $\Phi:\mathbf{M}\times\mathbf{R}\to\mathcal{GL}(m)$ is continuous and the associated linear skew-product flow is

$$\pi(B, x, \tau)=(\Phi(B, \tau)x, B_\tau).$$

When the coefficients a_{ij} are stationary stochastic variables, it is not difficult to show (by using the ideas of § 5) that a given underlying invariant probability measure μ lifts to an invariant measure ν on $\mathbf{M}$. If the coefficients are metrically transitive, i.e. if μ is ergodic, then the lifted measure ν can be chosen to be ergodic.

Appendix. Further geometric properties of cocycles. A *projector* is a continuous mapping $P(x, \theta)=(P(\theta)x, \theta)$ on $\mathbf{R}^m\times\mathbf{M}$ where $\mathbf{M}$ is a compact Hausdorff space and $P(\theta)$ is a linear projection on $\mathbf{R}^m$. A *resolution of the identity* on $\mathbf{R}^m\times\mathbf{M}$ is a k-tuple $P=(P_1,\cdots, P_k)$, $k\geqq 1$, satisfying: (i) each P_i is a projector, (ii) $P_iP_j=\vec{0}$ when $i\neq j$ and (iii) $I=P_1+\cdots+P_k$. (Here we define $\vec{0}(x, \theta):=(0, \theta)$.) Let $P=(P_1,\cdots, P_k)$ be a k-tuple of projectors on $\mathbf{R}^m\times\mathbf{M}$ and define

$$\mathcal{R}_i:=\text{Range}\,(P_i):=\{(x, \theta)\in\mathbf{R}^m\times\mathbf{M}\colon P_i(\theta)x=x\},\qquad 1\leqq i\leqq k.$$

Then P is a resolution of the identity if and only if $\mathbf{R}^m\times\mathbf{M}=\mathcal{R}_1+\cdots+\mathcal{R}_k$ (as a Whitney sum). A resolution of the identity P is said to be *orthogonal* if $\mathcal{R}_i\perp\mathcal{R}_j$ whenever $i\neq j$, i.e. the Euclidean inner product $\langle\cdot,\cdot\rangle$ satisfies $\langle x, y\rangle=0$ for all $(x, \theta)\in\mathcal{R}_i$, $(y, \theta)\in\mathcal{R}_j$ when $i\neq j$. The latter is equivalent to saying that each $P_i(\theta)$ is an orthogonal projection on $\mathbf{R}^m$.

Now let Φ be a cocycle on $\mathbf{R}^m\times\mathbf{M}$ and assume $\mathbf{R}^m\times\mathbf{M}=\mathcal{V}_1+\cdots+\mathcal{V}_k$ as a Whitney sum. Let $P=(P_1,\cdots, P_k)$ be the induced resolution of the identity where Range $(P_i)=\mathcal{V}_i$, $1\leqq i\leqq k$. Then the subbundles $\mathcal{V}_i$ are invariant under the linear skew product flow induced by Φ if and only if one has

$$P_i(\theta\cdot t)\Phi(\theta, t)=\Phi(\theta, t)P_i(\theta),\qquad 1\leqq i\leqq k \tag{A.1}$$

for all $\theta\in\mathbf{M}$ and $t\in\mathbf{T}$. We shall say that a resolution of the identity $P=(P_1,\cdots, P_k)$ is *invariant* if (A.1) is satisfied. It is not always the case that an invariant resolution of the identity is orthogonal; however, the next lemma shows that one can replace Φ with a cohomologous flow in which the new invariant resolution of the identity is orthogonal.

LEMMA A. *Let Φ be a cocycle over a compact Hausdorff space $\mathbf{M}$ and let $P=(P_1,\cdots, P_k)$ be an invariant resolution of the identity. Then there is a continuous self-adjoint mapping $R:\mathbf{M}\to\mathcal{GL}(m)$ with the property that $Q=(Q_1,\cdots, Q_k)$ is an orthogonal resolution of the identity, where*

$$Q_i(\theta)=R(\theta)P_i(\theta)R^{-1}(\theta),\qquad 1\leqq i\leqq k. \tag{A.2}$$

Furthermore Q is invariant under the cocycle

$$\Psi(\theta, t)=R(\theta\cdot t)\Phi(\theta, t)R^{-1}(\theta). \tag{A.3}$$

Proof. Define $S(\theta)$ by

$$S(\theta):=\sum_{i=1}^{k}P_i^*(\theta)P_i(\theta)$$

where P^* denotes the adjoint operation. Then $S(\theta)$ is positive definite and self-adjoint, so it has a unique positive definite, self-adjoint square root $R(\theta)$, i.e. $R^2(\theta)=S(\theta)$. If $Q_i(\theta)$ is defined by (A.2) and Ψ is given by (A.3) it is easy to verify that $Q_i^*(\theta)=Q_i(\theta)$ and $\Psi(\theta, t)Q_i(\theta)=Q_i(\theta\cdot t)\Psi(\theta, t)$, $1\leqq i\leqq k$. Q.E.D.

Lemma 3.4 also gives information about the case where one has a linear skew-product flow on a vector bundle $\mathscr{E}$ over a compact Hausdorff space $\mathbf{M}$ where $\mathscr{E} = \mathscr{V}_1 + \cdots + \mathscr{V}_k$ is a Whitney sum of invariant subbundles. Each of these subbundles $\mathscr{V}_i$ can be separately imbedded in a trivial bundle $\mathbf{R}^{m_i} \times \mathbf{M}$ where $m = m_1 + \cdots + m_k$. The construction of Lemma 3.4 shows that one can construct a cocycle on $\mathbf{R}^m \times \mathbf{M}$ so that the given flow on $\mathscr{E}$ is cohomologous to a flow on a subbundle of $\mathbf{R}^m \times \mathbf{M}$.

If π is a discrete flow on a vector bundle $\mathscr{E}$, i.e. if $\mathbf{T} = \mathbf{Z}$, then Lemma 3.4 can be extended to this case by first suspending the discrete flow to get an equivalent continuous-time flow on a new vector bundle. See Ellis and Johnson (1982) for the suspension construction. One should note that even if the original bundle is trivial, the suspended bundle may be nontrivial.

The triangularization technique can be used to put some cocycles into a block-diagonal, upper-triangular form. Let Φ be a cocycle on $\mathbf{M}$ and let $P = (P_1, \cdots, P_k)$ be an invariant partition of unity of $\mathbf{R}^m \times \mathbf{M}$. Because of Lemma A we can assume P to be orthogonal. For any point $\phi = (\theta, U) \in \mathbf{H}$ we define the P-partition of U to be the partitioning of U into block matrices $U = (U_1, \cdots, U_k)_P$ where the number m_i of column vectors in U_i is $\dim \operatorname{Range} P_i(\theta)$, $1 \leqq i \leqq k$. Let $\mathbf{H}_P$ denote the set of all $\phi = (\theta, U) \in \mathbf{H}$ with the property that the P-partition $U = (U_1, \cdots, U_k)_P$ satisfies

$$P_i(\theta) U_j = \delta_{ij} U_i, \qquad 1 \leqq i, \quad j \leqq k. \tag{A.4}$$

By using Lemma A one can easily verify the following:

LEMMA B. *$\mathbf{H}_P$ is a compact invariant set in $\mathbf{H}$ in the flow $\phi \cdot t$. Furthermore if $\phi = (\theta, U) \in \mathbf{H}_P$ then the column vectors in $\Phi(\theta, t) U_i$ are orthogonal to those in $\Phi(\theta, t) U_j$ when $i \neq j$.*

For $\phi = (\theta, U) \in \mathbf{H}_P$ let $T(\Phi(\theta, t) U)$ be given by (4.5). The P-partition of U prescribes an induced block partition of $T(\Phi(\theta, t) U)$ where the diagonal blocks are square matrices of size $(m_i \times m_i)$, $1 \leqq i \leqq k$. Since the off-diagonal blocks of $T(\Phi(\theta, t) U)$ depend on the inner products of column vectors of $\Phi(\theta, t) U_i$ and $\Phi(\theta, t) U_j$ for $i \neq j$, it follows from Lemma B that these off-diagonal blocks are zero. Hence $\Psi(\phi, t) = T(\Phi(\theta, t) U)^{-1}$ is a block-diagonal, upper-triangular matrix.

The block-diagonalization of Ψ involves an "untwisting" of the spectral subbundles of Φ. A similar untwisting with additional useful structures appears in Ellis and Johnson (1982). Also compare with Coppel (1967), Palmer (1980) and Vinograd et al. (1977).

REFERENCES

[1] M. Atiyah (1967), *K-Theory*, W. A. Benjamin, New York.

[2] B. V. Bylov et al. (1966), *The Theory of Lyapunov Exponents and Its Applications to Stability Problems*, Nauka, Moscow. (In Russian.)

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

[4] J. E. Cohen (1979), *Ergodic theorems in demography*, Bull. Amer. Math. Soc. (N.S.), 1, pp. 275-295.

[5] P. Constantin and C. Foias (1983), *Global Lyapunov exponents, Kaplan-Yorke formulas and the dimension of the attractors for* 2D *Navier-Stokes equations*, IMA Preprint #21, Comm. Pure Appl. Math, to appear.

[6] W. A. Coppel (1978), *Dichotomies in Stability Theory*, Lecture Notes in Mathematics 629, Springer-Verlag, New York-Heidelberg-Berlin.

[7] H. Crauel (1981), *Ergodentheorie linearer stochastischer systeme*, Universität Bremen, Report No. 59.

[8] J. Daletskii and M. Krein (1974), *Stability of Solutions of Differential Equations in Banach Space*, AMS Translations, Amer. Math. Soc., Providence, RI.

[9] S. P. Diliberto (1950), *On systems of ordinary differential equations*, Contributions Thy. Nonlin. Oscil., Ann. of Math. Studies, 20, pp. 1-38.

[10] R. Ellis (1969), *Lectures on Topological Dynamics*, W. A. Benjamin, New York.

[11] R. Ellis and R. Johnson (1982), *Topological dynamics and linear differential systems*, J. Differential Equations, 44, pp. 21-39.

[12] H. Federer (1969), *Geometric Measure Theory*, Springer-Verlag, New York.

[13] N. Fenichel (1971), *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21, pp. 193-226.

[14] C. Foias, G. R. Sell and R. Temam (1985), *Variété inertielles des équations différentielles dissipatives*, C.R. Acad. Sci. Paris, Serie I, 301, pp. 139-141.

[15] H. Furstenberg (1961), *Strict ergodicity and transformations of the torus*, Amer. J. Math., 83, pp. 573-601.

[16] H. Furstenberg and H. Kesten (1960), *Products of random matrices*, Ann. Math. Stat., 31, pp. 457-469.

[17] V. Guillemin and A. Pollack (1974), *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ.

[18] E. Hewitt and K. Ross (1963), *Abstract Harmonic Analysis* I, Springer-Verlag, New York.

[19] W. M. Hirsch, C. C. Pugh and M. Shub (1977), *Invariant Manifolds*, Lecture Notes in Mathematics 583, Springer-Verlag, New York-Heidelberg-Berlin.

[20] K. Ishii (1973), *Localization of eigenstates and transport phenomena in one-dimensional disordered systems*, Supp. Theor. Phys., 53, pp. 77-138.

[21] R. A. Johnson (1978), *Ergodic theory and linear differential equations*, J. Differential Equations, 28, pp. 23-34.

[22] ——— (1980a), *On a Floquet theory for almost-periodic, two-dimensional linear systems*, J. Differential Equations, 37, pp. 184-205.

[23] ——— (1980b), *Analyticity of spectral subbundles*, J. Differential Equations, 35, pp. 366-387.

[24] ——— (1982), *On the recurrent Hill equation*, J. Differential Equations, 46, pp. 165-194.

[25] ——— (1986), *The Oseledec and Sacker-Sell Spectra: An example*, Proc. Amer. Math. Soc., to appear.

[26] R. A. Johnson and J. Moser (1982), *The rotation number for almost periodic potentials*, Comm. Math. Phys., 84, pp. 403-438.

[27] A. Katok (1980), *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publ. I.H.E.S., 51, pp. 137-174.

[28] J. Kingman (1968), *The ergodic theory of subadditive stochastic processes*, J. Royal Statist. Soc. Ser. B, 30, pp. 499-510.

[29] S. Kotani (1982), *Lyapunov indices determine absolutely continuous spectra of stationary random one-dimensional Schrödinger operators*, Proc. Kyoto Stochastic Conference.

[30] J. Kurzweil (1957), *Generalized ordinary differential equations and continuous dependence on a parameter*, Czechoslovak Math. J., 7 (82), pp. 415-448.

[31] S. T. Liao (1966), *Applications to phase-space structure of ergodic properties of the one-parameter transformation group induced on the tangent bundle by a differential system on a manifold* I, Acta Sci. Natur. Univ. Pekinensis, 12, pp. 1-43. (In Chinese.)

[32] ——— (1973), *An ergodic property theorem for a differential system*, Sci. Sinica, 16, pp. 1-24.

[33] M. A. Lyapunov (1892), *Problème général de la stabilité du mouvement*, Ann. Math. Studies, 17, 1947.

[34] G. A. Margulis (1975), *Discrete groups of motions of manifolds of nonpositive curvature*, Proc. Internat. Cong. Math., Vancouver, 1974, pp. 21-34.

[35] Y. Matshushima (1972), *Differentiable Manifolds*, Marcel Dekker, New York.

[36] R. K. Miller and G. R. Sell (1970), *Volterra integral equations and topological dynamics*, Memoir Amer. Math. Soc., No. 102.

[37] V. M. Millionscikov (1968), *Metric theory of linear systems of differential equations*, Math. USSR-Sbornik, 6, pp. 149-158.

[38] ——— (1978), *Typicality of almost reducible systems with almost periodic coefficients*, Differential Equations, 14, pp. 448-450.

[39] A. W. Naylor and G. R. Sell (1982), *Linear Operator Theory in Engineering and Science*, 2nd ed., Appl. Math. Sci., 40, Springer-Verlag, New York.

[40] V. Nemytskii and V. Stepanov (1960), *Qualitative Theory of Differential Equations*, Princeton Univ. Press, Princeton, NJ.

[41] V. L. Novikov (1975), *On almost reducible systems with almost periodic coefficients*, Math. Notes, 16, pp. 1065-1071.

[42] V. Oseledec (1968), *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc., 19, pp. 197-231.

[43] K. J. Palmer (1980), *On the reducibility of almost periodic systems of linear differential equations*, J. Differential Equations, 35, pp. 374-390.

[44] ——— (1984), *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55, pp. 225-256.

[45] L. PASTUR (1980), *Spectral properties of disordered systems in the one-body approximation*, Comm. Math. Phys., 75, pp. 179-196.

[46] S. PELIKAN (1983), *The dimension of attractors are surfaces*, Ph.D. dissertation, Boston University, Boston, MA.

[47] O. PERRON (1930), *Über eine Matrixtransformation*, Math. Z., 32, pp. 465-473.

[48] D. PERRY (1986), *A numerical study of the relationship between the measurable spectrum and the continuous spectrum*, Ph.D. dissertation, University of Minnesota, Minneapolis, MN, to appear.

[49] YA. B. PESIN (1977), *Lyapunov characteristic exponents and smooth ergodic theory*, Russian Math. Surveys, 32, No. 4, pp. 55-114.

[50] R. PHELPS (1966), *Lectures on Choquet Theory*, American Book Co., New York.

[51] M. RAGHUNATHAN (1979), *A proof of Oseledec's multiplicative ergodic theorem*, Israel J. Math., 32, pp. 356-362.

[52] D. RUELLE (1979), *Ergodic theory of differentiable dynamical systems*, Publ. I.H.E.S., 50, pp. 275-320.

[53] R. J. SACKER (1969), *A perturbation theorem for invariant manifolds and Hölder continuity*, J. Math. Mech., 18, pp. 705-762.

[54] R. J. SACKER AND G. R. SELL (1974), *Existence of dichotomies and invariant splittings for linear differential systems* I, J. Differential Equations, 15, pp. 429-458.

[55] ——— (1975), *A spectral theory for linear almost periodic equations. Preliminary report*, in International Conference on Differential Equations, Academic Press, New York, pp. 698-709.

[56] ——— (1976a), *Existence of dichotomies and invariant splittings for linear differential systems* II, J. Differential Equations, 22, pp. 478-496.

[57] ——— (1976b), *Existence of dichotomies and invariant splittings for linear differential systems* III, J. Differential Equations, 22, pp. 497-522.

[58] ——— (1978), *A spectral theory for linear differential systems*, J. Differential Equations, 27, pp. 320-358.

[59] ——— (1980), *The spectrum of an invariant submanifold*, J. Differential Equations, 37, pp. 135-160.

[60] J. SELGRADE (1975), *Isolated invariant sets for flows on vector bundles*, Trans. Amer. Math. Soc., 203, pp. 359-390.

[61] G. R. SELL (1979), *Bifurcation of higher dimensional tori*, Arch. Rational Mech. Anal., 69, pp. 199-230.

[62] ——— (1984), *Linearization and global dynamics*, Proc. Internat. Cong. Math., Warsaw, 1982, Vol. 2, pp. 1283-1296.

[63] S. SMALE (1967), *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73, pp. 747-817.

[64] R. E. VINOGRAD (1956), *Necessary and sufficient criteria for the behavior of solutions of regular systems*, Math. Sbornik, 38 (80), pp. 23-50.

[65] R. E. VINOGRAD ET AL. (1977), *On the topological causes of the anomalous behavior of certain almost periodic systems*, in Problems of the Asymptotic Theory of Nonlinear Oscillations, Naukova Dumka, Kiev, pp. 54-61. (In Russian.)

[66] Y. KIFER (1985), *A multiplicative ergodic theorem for random transformations*, J. Anal. Math., 45, pp. 207-233.

[67] W. A. COPPEL (1967), *Dichotomies and reducibility*, J. Differential Equations, 3, pp. 500-521.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

002

SIMPLE CRITERIA FOR STABLE BIFURCATING PERIODIC SOLUTIONS OF O.D.E.'s*

G. CICOGNA†

Abstract. We propose a very simple criterion (generalizing classical Hopf theory) ensuring the existence of bifurcating periodic solutions for systems of O.D.E. of order larger than one. We show also that these solutions can be evaluated by means of a practical recursive procedure, and give a direct rule for finding the critical Floquet exponent. Some further generalizations are also considered, mainly based on symmetry properties and stability theory.

Key words. nonlinear systems of O.D.E.'s, Hopf bifurcation, recursive methods, Floquet exponent, equivariant bifurcation problems

AMS(MOS) subject classifications. 34A34, 58E07, 34A45, 34D20

1. Introduction. We will deal in this paper with systems of nonlinear differential equations, for a n-component real variable $x = x(t)$, of order higher than one: our aim is mainly to provide simple and readily applicable criteria concerning the existence of bifurcating periodic solutions (§ 2), and their stability (§ 6). It can be observed, actually, that any system of differential equations of order higher than one could be transformed—in principle—into a system containing only first-order derivatives; however, the price that must be paid is the introduction of a very high number of equations and variables, which are in general not easily handled. In addition, it may happen that the initial system cannot be explicitly transformed into a first-order system, so that the classical Hopf bifurcation theory cannot be applied. Our criteria do not require these transformations; rather they involve only linear algebra in the vector space C^n.

The interest in this type of problem is increasing: we can refer to the books and reviews [2], [4], [8], [9], [11], [12], and to the Proceedings of recent courses [1], [10], [17].

Another relevant point is that our scheme enables us to apply a general recursion procedure (again in an n-dimensional space) for explicitly calculating the periodic solution (cf. [11]): this will be illustrated also by means of an example (§ 4). Another example will be examined in some detail (§ 7), in order to discuss the stability of the bifurcating periodic solution, in terms of Floquet exponents [11]. In § 3, we will compare our results with the classical Hopf bifurcation theory.

In § 5 we discuss some symmetry properties of the problem. It is known that symmetries play an interesting role in bifurcation theory; what concerns us especially is the case of bifurcating periodic solutions (see the paper by Golubitsky and Stewart [7]) where one can find a detailed analysis of this point of view. Symmetries are also treated in § 8, where some generalizations of the method are briefly considered: precisely, the case of multiple critical eigenvectors (Theorem 3), and finally the case when the classical conditions of the Hopf method are not verified, and one has to resort to different arguments based on stability theory (Theorem 4).

2. Existence of bifurcating periodic solutions. Let $x = x(t)$ be a real n-dimensional vector, depending on time t; we shall consider a family, depending on a real parameter

* Received by the editors July 3, 1985, and in revised form September 20, 1985.

† Dipartimento di Fisica, Università di Pisa, I-56100 Pisa, Italy.

λ, of systems of nonlinear ordinary differential equations of the following type

$$\text{(1a)} \qquad F(\lambda, x) \equiv L(\lambda)x + H(\lambda, x), \qquad \lambda \in R, \quad x \in R^n$$

where L is the linear differential operator

$$\text{(1b)} \qquad L(\lambda) = \sum_{r=0}^{k} A_r(\lambda) \frac{d^r}{dt^r}$$

and $A_r = A_r(\lambda)$ $(r = 0, \cdots, k)$ are given $n \times n$ real matrices, $H: R \times R^n \to R^n$ is the remaining nonlinear higher order part, with $H(\lambda, 0) = 0$; we will assume also for simplicity that $A_r = A_r(\lambda)$ and $H = H(\lambda, x)$ are analytical functions.

Theorem 1 below provides a condition for the existence of a bifurcating nonzero periodic solution of (1): this will be readily obtained introducing first the $n \times n$ auxiliary matrix T defined by

$$\text{(2)} \qquad T = T(\lambda, \omega) = \sum_{r=0}^{k} (i\omega)^r A_r(\lambda)$$

where ω is a new real parameter.

THEOREM 1. *Suppose that when* $\lambda = \lambda_0$ *and* $\omega = \omega_0 > 0$ *the matrix* $T_0 = T(\lambda_0, \omega_0)$ *defined in* (2) *is singular, and that its kernel in* C^n *is one-dimensional. Assume also that the usual "no-resonance" condition is fulfilled, i.e. that, in the case there is some other* $\omega' \neq \omega_0$ *such that* $\det T(\lambda_0, \omega') = 0$, *then* ω'/ω_0 *is not an integer number. Finally, denoting by* ζ *and* ζ' *unit vectors in* C^n *such that*[1]

$$\text{(3)} \qquad T_0\zeta = 0 \quad \textit{and} \quad T_0^+\zeta' = 0$$

assume that the two complex numbers

$$\text{(4)} \qquad \langle T_\lambda \rangle = \left(\frac{\partial T}{\partial \lambda}\zeta, \zeta'\right) \quad \textit{and} \quad \langle T_\omega \rangle = \left(\frac{\partial T}{\partial \omega}\zeta, \zeta'\right)$$

(*where* (,) *is the scalar product in* C^n *and derivatives are evaluated at* $\lambda = \lambda_0$ *and* $\omega = \omega_0$) *are not aligned with the origin of the complex plane. Then, the problem* (1) *possesses a periodic nonzero solution, branching at* $\lambda = \lambda_0$, *with period* $2\pi/\omega$, *which can be parameterized in this form*

$$\text{(5)} \qquad \begin{aligned} x &= x(s, t) = s \operatorname{Re}(\zeta e^{i\omega t}) + w(s, t), \\ \lambda &= \lambda(s), \\ \omega &= \omega(s), \end{aligned}$$

where s is a real parameter, defined in a neighbourhood of zero, and such that

$$\lim_{s \to 0} \lambda(s) = \lambda_0, \quad \lim_{s \to 0} \omega(s) = \omega_0, \quad \lim_{s \to 0} s^{-1} w(s, t) = 0.$$

Proof. Having rescaled the time variable t

$$\text{(6)} \qquad t \to \tau = \omega t$$

(so one has to look for 2π-periodic solutions in τ), let us put

$$\text{(7)} \qquad L(\lambda, \omega) = \sum_{r=0}^{k} \omega^r A_r(\lambda) \frac{d^r}{d\tau^r}$$

[1] The choice of phase factors of ζ and ζ' is irrelevant here and in the following.

and introduce the $L^2((0,2\pi), R^n)$ scalar product

$$(x(\tau), y(\tau))_{L^2} = \frac{1}{\pi}\int_0^{2\pi} (x(\tau), y(\tau))_{R^n}\, d\tau \tag{8}$$

(easily extended, whenever necessary, to $L^2((0,2\pi), C^n)$). Our assumptions imply that the operator $L_0 = L(\lambda_0, \omega_0)$ is a Fredholm operator of index zero, with two-dimensional (in the real sense) kernel V, spanned by

$$\mathrm{Re}\,(\zeta\, e^{i\tau}) \quad \text{and} \quad \mathrm{Im}\,(\zeta\, e^{i\tau}). \tag{9a}$$

Let V' be the kernel, spanned by

$$\mathrm{Re}\,(\zeta'\, e^{i\tau}) \quad \text{and} \quad \mathrm{Im}\,(\zeta'\, e^{i\tau}), \tag{9b}$$

of the formal adjoint

$$L_0^+ = \sum_{r=0}^{k} (-\omega_0)^r A_r(\lambda_0) \frac{d^r}{d\tau^r}$$

and W, W' the orthogonal complementary subspaces in $L^2((0,2\pi), R^n)$ of V, V' respectively. According to the usual Lyapunov-Schmidt technique (see e.g. [4], [8]), introduce the projectors P, P' on V, V', and Q, Q' on W, W'; then, writing $w = Qx$ and $v = Px$, the projection of (1) on W' uniquely fixes w as a function of λ, ω and v; next, the projection on V' gives the bifurcation equation:

$$P'F(\lambda, \omega, v + w(\lambda, \omega, v)) \equiv \Phi(\lambda, \omega, v) \equiv \Psi(\lambda, \omega, v)v = 0 \tag{10}$$

where $\Phi: R^2 \times V \to V'$, and $\Psi(\lambda, \omega, v)$ is a 2×2 real matrix with the property

$$\Psi(\lambda_0, \omega_0, 0) = 0.$$

Let now $\hat{v}$ be any unit vector in V (the choice of $\hat{v}$ is arbitrary, as we shall see more clearly in § 5), and s a real parameter; then (10), written in the one-dimensional subspace $\{s\hat{v}\}$, becomes

$$\Psi(\lambda, \omega, s)\hat{v} = 0$$

($s = 0$ corresponds to the trivial solution $v = w = x = 0$). An application of the implicit function theorem gives that this equation has nonzero solution if

$$\frac{\partial\Psi}{\partial\lambda}(\lambda_0, \omega_0, 0)\hat{v} \quad \text{and} \quad \frac{\partial\Psi}{\partial\omega}(\lambda_0, \omega_0, 0)\hat{v} \tag{11}$$

are two linearly independent vectors (in the real sense: see [4] and also [6, § 6]). If this condition is satisfied, the solution can be put precisely in the explicit form (5). On the other hand, one has

$$\frac{\partial\Psi}{\partial\lambda}(\lambda_0, \omega_0, 0)\hat{v} = \frac{\partial}{\partial\lambda} P'F_x(\lambda_0, \omega_0, 0)\hat{v} = P'\frac{\partial L}{\partial\lambda}(\lambda_0, \omega_0)\hat{v}$$

and similarly for $(\partial\Psi/\partial\omega)\hat{v}$, having taken into account that $w_v(\lambda_0, \omega_0, 0) = 0$, as known from usual bifurcation theory. The projection P' on V' is obtained by evaluating the two scalar products with vectors (9b); using now definition (8), one can observe that, for any real function $y(\tau) \in L^2$,

$$\begin{aligned}(y(\tau), \mathrm{Re}\,(\zeta'\, e^{i\tau}))_{L^2} &= \mathrm{Re}\,(\eta_1, \zeta')_{C^n},\\ (y(\tau), \mathrm{Im}\,(\zeta'\, e^{i\tau}))_{L^2} &= \mathrm{Im}\,(\eta_1, \zeta')_{C^n}\end{aligned} \tag{12}$$

where $\eta_1 \in C^n$ is the vector of the first components of the Fourier series

$$y(\tau) = \sum_{p=0}^{\infty} \frac{1}{2}(\eta_p e^{ip\tau} + \bar{\eta}_p e^{-ip\tau}).$$

Choosing now $\hat{v} = \operatorname{Re}(\zeta e^{i\tau})$, it is immediately seen that the condition given above on the two vectors (11) is equivalent to the assumption on the scalar products (4) as given in the theorem. The other assertions follow from standard results of the Lyapunov-Schmidt procedure.

3. A particular case: the Hopf bifurcation problem. In the particular case of first-order differential systems ($k = 1$), the two quantities (4) become now

$$\langle T_\lambda \rangle = ((i\omega_0 A_{1\lambda} + A_{0\lambda})\zeta, \zeta') \quad \text{and} \quad \langle T_\omega \rangle = i(A_1\zeta, \zeta') \tag{13}$$

where $A_{1\lambda} = \partial A_1/\partial\lambda$ evaluated at $\lambda = \lambda_0$, etc. If, in addition, the matrix $A_1(\lambda)$ is assumed to be invertible, at least near λ_0 (Theorem 1 actually works also without this assumption), the problem is the classical Hopf bifurcation problem:

$$\frac{d}{dt}x = B(\lambda)x + C(\lambda, x) \quad \text{with } B(\lambda) = -A_1^{-1}(\lambda)A_0(\lambda).$$

Observing that

$$B(\lambda_0)\zeta = i\omega_0\zeta,$$
$$B^+(\lambda_0)\zeta'' = -i\omega_0\zeta'' \quad \text{if } \zeta'' = A_1^+(\lambda_0)\zeta',$$

one obtains in this case from (13) (all quantities are evaluated at $\lambda = \lambda_0$)

$$\langle T_\lambda \rangle = -\sigma_\lambda(\zeta, \zeta'') \quad \text{and} \quad \langle T_\omega \rangle = i(\zeta, \zeta'')$$

where $\sigma = \sigma(\lambda)$ is the critical branch of eigenvalues of $B(\lambda)$, with $\sigma(\lambda_0) = i\omega_0$. Our condition on (4) becomes then

$$(\zeta, \zeta'') \neq 0 \quad \text{and} \quad \operatorname{Re}\sigma_\lambda(\lambda_0) \neq 0$$

which is precisely the usual "transversality" assumption [11], [4], together with the condition that the critical eigenvalue $i\omega_0$ of B is semisimple (i.e. algebraically simple). For what concerns this last assertion, note in fact that $(\zeta, \zeta'') = 0$ would imply, using also (8) and (12), that $\operatorname{Re}(\zeta e^{i\tau})$ and $\operatorname{Im}(\zeta e^{i\tau})$ are orthogonal to the kernel of the formal adjoint

$$L_0^+ = -\omega_0\frac{d}{dt} - B^+(\lambda_0),$$

which in turn implies that there exists some vector $\xi \in C^n$ such that $(B(\lambda_0) - i\omega_0)\xi = \zeta$, thus showing the nonsemisimplicity of the eigenvalue $i\omega_0$, and vice versa.

4. The recursive method. First example. We want to show here a very general recursive method (cf. [11]) which, if the hypotheses of Theorem 1 are verified, can be applied for explicitly evaluating solution (5), and which appears to be more convenient—for practical uses—than the Lyapunov-Schmidt procedure.

Let us insert into (1) the following expansion of (5) in powers of the parameter s

$$\begin{aligned} x(s, \tau) &= sx_{(1)}(\tau) + s^2x_{(2)}(\tau) + s^3x_{(3)}(\tau) + \cdots, \\ \lambda(s) &= \lambda_0 + s\lambda_1 + s^2\lambda_2 + \cdots, \\ \omega(s) &= \omega_0 + s\omega_1 + s^2\omega_2 + \cdots. \end{aligned} \tag{14}$$

At the first order one gets

$$L(\lambda_0, \omega_0)x_{(1)}(\tau) = 0$$

which is solved just by all vectors in V. Choose now (the normalization of ζ fixes the scale of s)

$$x_{(1)} = \text{Re}\,(\zeta\, e^{i\tau}).$$

At the order s^2, one has the linear differential problem

$$L_0 x_{(2)}(\tau) = -\left(\lambda_1 \frac{\partial L}{\partial \lambda} + \omega_1 \frac{\partial L}{\partial \omega}\right) x_{(1)}(\tau) + H_2(\tau) \tag{15}$$

where $H_2(\tau)$ is the second order term in the expansion of $H(\lambda, x)$ in powers of s. Classical alternative theorems say that this equation can be solved (in $x_2(\tau)$) if its r.h.s. is orthogonal to V'; imposing this condition by means of (8) and (12) gives an equation for λ_1, ω_1. This equation, in turn, can be satisfied just if assumptions of Theorem 1 (in particular the one concerning quantities (4)) are verified.

The method can be illustrated by means of the following example. This is one of the simplest examples which could have some interest, because of the occurrence of the highest derivatives in just one of the equations. To my knowledge, such cases which are not (or cannot be) transformed into a first order system rarely appear to be considered in the literature. However, there are several problems (e.g. in nonlinear field theories, nonlinear quantum-mechanical systems, etc.: see [1], [2], [9]) which are of this form: our method may then be useful, even if this possibility cannot be considered in this paper. Consider then the system, with $k = 2$, $n = 2$, writing here the two-component vector x in the form $x \equiv (x(t), y(t))$

$$\ddot{x} + x - \ddot{y} - x^2 = 0, \qquad \dot{y} + \lambda x - x^3 = 0 \qquad (\dot{x} = dx/dt, \text{etc.}).$$

It is easily seen that the auxiliary matrix

$$T(\lambda, \omega) = \begin{pmatrix} -\omega^2 + 1 & \omega^2 \\ \lambda & i\omega \end{pmatrix}$$

is singular for $\lambda_0 = 0$ and $\omega_0 = 1$, and that all conditions of Theorem 1 are satisfied with $\langle T_\lambda \rangle / \langle T_\omega \rangle = -i$. The spaces V and V' are spanned by

$$\begin{pmatrix} \cos\tau \\ 0 \end{pmatrix}, \begin{pmatrix} \sin\tau \\ 0 \end{pmatrix} \quad \text{and} \quad \frac{1}{\sqrt{2}}\begin{pmatrix} \cos\tau \\ \sin\tau \end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix} -\sin\tau \\ \cos\tau \end{pmatrix}$$

respectively. An alternative theorem from (15) requires

$$\lambda_1 \langle T_\lambda \rangle + \omega_1 \langle T_\omega \rangle = 0$$

which gives

$$\lambda_1 = \omega_1 = 0$$

and then (neglecting for simplicity an independent solution with $y = \text{const.}$)

$$x_{(2)}(\tau) = \frac{1}{6}\begin{pmatrix} 3 - \cos 2\tau \\ 0 \end{pmatrix}.$$

At the third step, the solvability condition can be satisfied, as before, by fixing λ_2 and ω_2; we thus obtain, summarizing:

$$x(s,\tau)=\begin{pmatrix} s\cos\tau+\dfrac{s^2}{6}(3-\cos 2\tau)-\dfrac{s^3}{96}(2\cos 3\tau+9\sin 3\tau)+\cdots \\ -\dfrac{s^3}{12}\sin 3\tau+\cdots \end{pmatrix},$$

$$\lambda=-\tfrac{3}{4}s^2+\cdots,\qquad \omega=1+\tfrac{5}{12}s^2+\cdots,\qquad \tau=\omega t.$$

It can be easily seen that the equation for $x_n(\tau)$ obtained at the nth step is not very different from (15). Then one can be easily convinced that our assumptions guarantee that each step can be completed, and that this method works for any other example as well.

5. Symmetry properties. All systems as (1), being "autonomous," exhibit in a natural way a "covariance" with respect to the group SO_2 (isomorphic to the circle group S^1) of the "time translations" $\tau\to\tau+\tau'(\bmod 2\pi)$ [7], [20], [22]. Let us recall that a map $F:R\times X\to X'$, where X, X' are topological vector spaces, is said to be covariant with respect to a topological group G [20]-[22], if there exist two continuous representations D and D' of G, acting on X and X' respectively, such that, for any $\lambda\in R$, $x\in X$ and $g\in G$,

$$F(\lambda, D(g)x)=D'(g)F(\lambda,x). \tag{16}$$

It can be shown [20]-[22], [5] that covariance is inherited by the bifurcation equation (10), and that the kernel V of the linearized map $F_x(\lambda,0)$ is an invariant subspace under the group action. In our cases, this implies that the two-dimensional space V is the basis space of a real irreducible representation of SO_2; therefore, all unit vectors in V are equivalent, in the sense that they belong to the same orbit under SO_2 (changing the choice of $\hat{v}$ corresponds to choose another origin for the time τ).

Another consequence is that, exactly as in Hopf bifurcation [11], λ and ω are *even* functions of s:

$$\lambda(s)=\lambda(-s),\qquad \omega(s)=\omega(-s).$$

In fact, after a time $\tau=\pi$, one has $s\hat{v}\,e^{i\tau}\to -s\hat{v}\,e^{i\tau}$, and then, from covariance of (10)

$$\psi(\lambda,\omega,s)=\psi(\lambda,\omega,-s)$$

which proves the assertion. The same result is confirmed by the iterative method used in the previous section: it suffices to observe that the odd-order coefficients λ_{2m-1}, ω_{2m-1} in the expansion (15) are determined, via solvability condition, by the projection on V' of the terms of order $2m$ in the s-power expansion of the nonlinear part $H(\lambda,x)$, and take into account the computation rule (12).

As a final remark, let us consider the case in which the system (1) contains only even-order derivatives: in this case, both quantities $\langle T_\lambda\rangle$ and $\langle T_\omega\rangle$ are real, and the condition stated in Theorem 1 cannot be satisfied. However, in this case, the covariance of the problem is larger: it is described in fact by the whole group O_2 (i.e., it includes time inversion $\tau\to-\tau$). In this case, by restricting e.g. the problem to the subspace of even functions (with respect to time inversion), one can easily reduce the bifurcation equation, as a consequence of symmetry, to a standard one-dimensional case [4], [7], [20]-[22], [5], and then apply classical arguments [4].

6. Stability of the bifurcating periodic solution. We will give now a readily applicable criterion for finding the "critical" Floquet exponent [11], [19] governing the linearized stability of the periodic bifurcated solution (5). Assuming here that $A_k(\lambda)$ is invertible in a neighbourhood of λ_0, and introducing the vector variables

$$x_1=x,\quad x_2=\frac{dx_1}{dt},\quad \cdots,\quad x_k=\frac{dx_{k-1}}{dt}, \qquad u\equiv(x_1,\cdots,x_k)\in R^{kn}, \tag{17}$$

system (1) becomes equivalent to

$$\frac{du}{dt}=M(\lambda)u+K(\lambda,u) \tag{18}$$

with

$$M(\lambda)=\left(\begin{array}{c|cccc} & I & & & 0\\ 0 & & \ddots & & \\ & 0 & & & I\\ \hline B_0 & B_1 & & \cdots & B_{k-1}\end{array}\right)$$

where I is the unit $n\times n$ matrix, and

$$B_j=B_j(\lambda)=-A_k^{-1}(\lambda)A_j(\lambda)\qquad (j=0,\cdots,k-1),$$

$$K=K(\lambda,u)=\begin{pmatrix}\vdots\\ 0\\ \vdots\\ -A_k^{-1}(\lambda)H(\lambda,x_1)\end{pmatrix}.$$

For the problem (18), the Floquet exponent [11], at the lowest order of the parameter s, is given by

$$\gamma=-s\frac{d\lambda}{ds}\operatorname{Re}\frac{d\rho}{d\lambda}=-2s^2\lambda_2\operatorname{Re}\frac{d\rho}{d\lambda} \tag{19}$$

where $\rho=\rho(\lambda)$ is the critical branch of the eigenvalues of the matrix M, with $\rho(\lambda_0)=i\omega_0$, and derivatives are evaluated at $\lambda=\lambda_0$.

We can now state the following:

THEOREM 2. *The sign of the Floquet exponent governing stability of the bifurcating periodic solution* (5) *of the system* (1) *is equal to the sign of the quantity*

$$\delta=\operatorname{Im}(\langle H_3\rangle\overline{\langle T_\omega\rangle}), \tag{20}$$

$\langle T_\omega\rangle$ *being defined in* (4) *and*

$$\langle H_3\rangle=\frac{1}{\pi}\int_0^{2\pi}(H_{(3)}(\tau),\zeta')\,e^{-i\tau}\,d\tau,$$

where $H_3(\tau)$ is the third-order term in the s-power expansion of the nonlinear part $H(\lambda,x)$, and all other assumptions are as in Theorem 1. *Therefore, $\delta<0$ (>0) corresponds to linearized stability* (*instability*).

Proof. Let us denote by η, $\eta'\in C^{kn}$ the critical eigenvectors of the matrix $M(\lambda_0)$:

$$M(\lambda_0)\eta=i\omega_0\eta,\qquad M^+(\lambda_0)\eta'=-i\omega_0\eta'.$$

One can see that (apart from normalization factors) η and η' are given, in terms of vectors ζ, ζ' (3), by

$$\eta = \begin{pmatrix} \zeta \\ (i\omega_0)\zeta \\ (i\omega_0)^2\zeta \\ \vdots \\ (i\omega_0)^{k-1}\zeta \end{pmatrix}$$

and

$$\eta' = \begin{pmatrix} \dfrac{-1}{(-i\omega_0)} A_0^+\zeta' \\ \dfrac{-1}{(-i\omega_0)^2}[A_0^+ + (-i\omega_0)A_1^+]\zeta' \\ \vdots \\ \dfrac{-1}{(-i\omega_0)^{k-1}}[A_0^+ + (-i\omega_0)A_1^+ + \cdots + (-i\omega_0)^{k-2}A_{k-2}^+]\zeta' \\ A_k^+\zeta' \end{pmatrix}$$

and verify, after some calculations, that

$$(\eta, \eta') = -i\langle T_\omega\rangle, \qquad \left(\frac{\partial M}{\partial\lambda}\eta, \eta'\right) = -\langle T_\lambda\rangle, \qquad \left(\frac{\partial M}{\partial\lambda}\eta, \eta'\right) = \frac{d\rho}{d\lambda}(\eta, \eta');$$

then

$$\text{(21)} \qquad \operatorname{Re}\frac{d\rho}{d\lambda} = \operatorname{Im}\frac{\langle T_\lambda\rangle}{\langle T_\omega\rangle}.$$

Now using methods of § 4, from the solvability condition

$$\lambda_2\langle T_\lambda\rangle + \omega_2\langle T_\omega\rangle + \langle H_3\rangle = 0,$$

one gets

$$\text{(22)} \qquad \lambda_2 = \frac{\operatorname{Im}(\langle H_3\rangle\overline{\langle T_\omega\rangle})}{\operatorname{Im}(\overline{\langle T_\lambda\rangle}\langle T_\omega\rangle)},$$

and finally

$$\gamma = \frac{2s^2}{|\langle T_\omega\rangle|^2}\delta.$$

7. Second example. We will consider here in some detail this second example, writing $x \equiv (x(t), y(t))$:

$$\text{(23)} \qquad \begin{aligned} &\ddot{x} + \dot{y} + 2x + \lambda y - (x+y)(x^2+y^2) = 0, \\ &\ddot{y} - \dot{x} + 2y - \lambda x + (x-y)(x^2+y^2) = 0 \end{aligned}$$

(which, of course, has the special property of being covariant also with respect to an additional rotation group SO_2 (see § 5; cf. also [7], [14], [15]) acting, for each fixed t, on the vectors $(x(t), y(t))$ through its basic representation). One has now $\lambda_0 = 0$, $\omega_0 = 2$, and $\langle T_\omega\rangle/\langle T_\lambda\rangle = -3i$; the periodic solution is given (exactly) by

$$\text{(24)} \qquad x(t) = \frac{1}{\sqrt{2}}\begin{pmatrix}\cos\omega t \\ \sin\omega t\end{pmatrix}, \quad \lambda = \frac{s^2}{2}, \quad \omega = 2 + \frac{1}{2}(\sqrt{9-2s^2} - 3).$$

In view of Theorem 2, one gets

$$H_{(3)}(\tau)=\frac{1}{2\sqrt{2}}\begin{pmatrix}-\cos\tau-\sin\tau\\ \cos\tau-\sin\tau\end{pmatrix}=\frac{1}{2\sqrt{2}}\operatorname{Re}\begin{pmatrix}-1+i\\ 1+i\end{pmatrix}e^{i\tau}$$

and then, using (12) and Theorem 2,

$$\delta=-\frac{3}{2},\qquad \gamma=-\frac{s^2}{3}$$

corresponding to stability of this solution for small variations of s.

It can be interesting to check these conclusions by directly inspecting the equations governing the small "perturbations" of solution (24): precisely, by writing

$$x(t)=\frac{s}{\sqrt{2}}\begin{pmatrix}\cos\omega t\\ \sin\omega t\end{pmatrix}+\begin{pmatrix}\xi(t)\\ \eta(t)\end{pmatrix},\qquad \lambda=\frac{s^2}{2}$$

and introducing for convenience the two functions

$$f(t)=\xi(t)\cos\omega t+\eta(t)\sin\omega t,$$
$$g(t)=\eta(t)\cos\omega t-\xi(t)\sin\omega t,$$

one can see that f and g obey the equations

$$\dddot{f}+3(3-s^2)\dot{f}+s^2(2\omega-1)f=0,$$
$$\ddot{g}=(1-2\omega)\dot{f}-s^2f.$$

This is sufficient to show that $\xi(t)$ and $\eta(t)$ behave typically according to

$$\xi(t),\ \eta(t)\sim e^{\gamma_i t}$$

where γ_i can be directly evaluated for small s to be[2]

$$\gamma_1=-\frac{s^2}{3}\quad\text{and}\quad \gamma_{2,3}=\pm 3i+\frac{s^2}{6},$$

so, γ_1 confirms the above calculation, whereas the two others correspond to displacements toward nonclosed orbits in the R^2-plane "oscillating" around the periodic solution (24). Alternatively, the various aspects of this discussion can be viewed writing (23) in polar coordinates r, θ. In this way one can obtain e.g. that

$$\text{(25a)}\qquad r^2\dot{\theta}-\tfrac{1}{2}r^2=C=\text{const},$$

$$\text{(25b)}\qquad \ddot{r}=-\frac{9}{4}r+r^3+\frac{C^2}{r^3};$$

then the "equilibrium solution" $r=r_0$ of (25b) requires through (25a)

$$\dot{\theta}=\text{const}=\frac{1}{2}+\left(\frac{9}{4}-r_0^2\right)^{1/2}$$

which is just (24) with $\dot{\theta}=\omega$ and $r_0=s/\sqrt{2}$.

8. Some generalizations. One of the hypotheses of Theorem 1 was that the kernel of $T_0=T(\lambda_0,\omega_0)$ is one-dimensional (in C^n). If the system (1) is covariant under some "external" symmetry group G (i.e. a group acting on the vectors $x\in R^n$ for each fixed

[2] Apart from the value $\gamma_0=0$, always present in these cases, as is well known [11].

t, not only through the time translations considered in § 5), this hypothesis is usually not verified; but it can happen that some group-theoretical consideration allows us to restrict the problem to some subspace where the initial hypothesis is recovered. The situation is completely similar to the one considered in full detail in [7] for the case of Hopf bifurcation. Referring for a more complete description of the group-theoretical situation to [7], we briefly state the assumption in this form:

(G) Let system (1) be covariant under a symmetry group G, as in (16), with $D=D'$. Suppose that there is an isotropy subgroup $\tilde{G}$ in G and a two-dimensional (in the real sense) subspace $U \subset R^n$ such that

$$D(\tilde{g})u=u \quad \text{for all } \tilde{g} \in \tilde{G}, \quad u \in U$$

and that no other vector in R^n is left fixed by $D(\tilde{G})$.

The following result needs then no further comment:

THEOREM 3. *Let system* (1) *satisfy assumption* (G); *then the restriction* $T_0|_U$ *maps* U *in itself, and if all other assumptions in Theorem* 1 *are verified by this restriction, the same conclusions are true.*

Another crucial point is the classical Hopf "transversality condition": in particular, it was proved in § 3 that our hypothesis concerning the two scalar products (4) is equivalent—in the case of first-order systems—to this condition. However, even if transversality is *not* satisfied, it is known [1], [3], [13], [16], [18] that suitable assumptions concerning a well defined change in the stability property of the solution $x \equiv 0$ can ensure the existence of a nontrivial bifurcation. Combining these ideas with symmetry properties, we get the following result, which we state for first-order normal systems.

THEOREM 4. *Let system* (1) *be of order* 1, *with*

$$L(\lambda)=\frac{d}{dt}+A_0(\lambda),$$

and satisfy assumption (G). *Denoting by* $\tilde{F}=\tilde{F}(\lambda, u)$ *the restriction of* F *to* $R \times U$, *suppose now that a positively definite (continuously differentiable, for simplicity) Lyapunov function* $V: U \to R$ *can be defined in a neighbourhood of* $u=0$ *in such a way that its time derivative*

$$\frac{dV}{dt}=\partial_u V \cdot \tilde{F}$$

is negatively definite for $\lambda=\lambda_0$, *but when* $\lambda>\lambda_0$ *it becomes positively definite. Suppose finally that* $\tilde{F}(\lambda, u) \neq 0$ *for all* $u \neq 0$ *in a neighbourhood of* $u=0$ *and* $\lambda>\lambda_0$: *then for* $\lambda>\lambda_0$ *there exists a bifurcating invariant set. This set is in general an annular region (possibly reduced to a single periodic cycle) around the origin* $u=0$, *and stable under perturbations belonging to* U.

Proof. By assumption (G), we can restrict the problem to the subspace U. Assumptions on the Lyapunov function V imply that the trivial solution $u=0$ is asymptotically stable for $\lambda=\lambda_0$ and becomes completely unstable for $\lambda>\lambda_0$; then (see [1], [3], [13], [16], [18]) there exists a stable bifurcating set in U. For classical Bendixson theorems, being U a two-dimensional real space and having locally $\tilde{F}$ no critical points other than $u=0$, this bifurcating set has all the mentioned properties.

Just for illustrating the idea, consider the following example. Let $x \in R^m$, $y \in R^m$, with $m=9$, be 4×4 real symmetric traceless matrices; let $G=\mathrm{SO}_4$ act irreducibly on the space R^9 according to the rule (in matrix notation)

$$x \to gxg^t, \qquad g \in G \tag{26}$$

and on the space $R^9 \oplus R^9$ according to the "diagonal" direct sum of the transformation rule (26). Consider the system, covariant with respect to this representation of G

$$\dot{x} = \lambda^3 x - \lambda y + \partial_x(\det x),$$

$$\dot{y} = \lambda x + \lambda^3 x + \partial_y(\det y).$$

The two-dimensional subspace U is generated by the two vectors $\tilde{x}$, $\tilde{y}$ lying along the direction of the diagonal matrix (cf. [5])

$$\tilde{e} = \frac{1}{3\sqrt{2}} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & -3 \end{pmatrix}.$$

Even if restricted to this subspace, standard Hopf theory cannot be applied (nor Theorem 1, being now $T_0 \equiv 0$); in fact for $\lambda = \lambda_0 = 0$ transversality is not verified and, in addition, $\omega_0 = \operatorname{Im} \sigma(0) = 0$. However, denoting now by $\tilde{x}$, $\tilde{y}$ (instead of the vectors) their projections along the direction $\tilde{e}$, and choosing the Lyapunov function

$$V = \tfrac{1}{2}(\tilde{x}^2 + \tilde{y}^2),$$

which gives

$$\frac{dV}{dt} = \lambda^3(\tilde{x}^2 + \tilde{y}^2) - \frac{1}{48}(\tilde{x}^4 + \tilde{y}^4),$$

all hypotheses of Theorem 4 are verified, and in fact one can see that a periodic solution bifurcates from $\lambda_0 = 0$. Clearly, under the action of the group G, this solution generates orbits of equivalent solutions.

Note of course that, in this example, it is impossible to extend the function V to the whole space in such a way that the required properties of dV/dt are preserved; also in this case, therefore, symmetry plays an important role.

REFERENCES

[1] A. Ambrosetti, ed., *Nonlinear Oscillations for Conservative Systems*, Proceedings of the meeting held in Venice 1985, Pitagora, Bologna, 1985.

[2] C. Bardos and D. Bessis, eds., *Bifurcation Phenomena in Mathematical Physics and Related Topics*, NATO Adv. Stu. Inst. Ser., Dordrecht, 1980.

[3] S. R. Bernfeld, P. Negrini and L. Salvadori, *Quasi-invariant manifolds, stability and generalized Hopf bifurcation*, Ann. Mat. Pura Appl. (IV), 130 (1982), pp. 1070-1085.

[4] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[5] G. Cicogna, *Bifurcation and symmetries*, Boll. Un. Mat. Ital., 1-B (1982), pp. 878-796, and Nuovo Cim. Lettere, 31 (1981), pp. 600-602.

[6] G. Cicogna and M. Degiovanni, *Mathematical hints in nonlinear problems*, Nuovo Cim., B-82 (1984), pp. 54-70.

[7] M. Golubitsky and I. Stewart, *Hopf bifurcation in the presence of symmetry*, Arch. Rat. Mech. Anal., 87 (1985), pp. 107-165.

[8] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[9] R. H. G. Helleman, ed., *Nonlinear Dynamics*, Ann. of the New York Academy of Sciences, Vol. 357, 1980.

[10] G. Iooss, ed., *Proceedings of the CIMPA School*, Nice, 1983, to appear.

[11] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, 1980.

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

[13] P. NEGRINI AND L. SALVADORI, *Attractivity and Hopf bifurcation*, Nonlinear Anal. T.M.A., 3 (1979), pp. 87-99.

[14] D. RAND, *Dynamics and symmetry predictions for modulated waves in rotating fluids*, Arch. Rat. Mech. Anal., 79 (1982), pp. 1-37.

[15] M. RENARDY, *Bifurcation from rotating waves*, Arch. Rat. Mech. Anal., 79 (1982), pp. 49-84.

[16] L. SALVADORI, *An approach to bifurcation via stability theory*, Proc. Conference on Recent Advances, in Nonlinear Analysis and Differential Equations, Madras, 1981.

[17] ———, ed., *Bifurcation Theory and Applications*, lectures given at the CIME course, Montecatini 1983, Lecture Notes in Mathematics 1057, Springer-Verlag, Berlin, 1984.

[18] ———, *Exchange of stability and bifurcation for periodic differential Systems*, Proc. VI International Conference on Trends in Theory and Practice of Nonlinear Analysis, North-Holland Elsevier, Amsterdam, in press.

[19] D. H. SATTINGER, *Topics in Stability and Bifurcation Theory*, Springer-Verlag, Berlin, 1973.

[20] ———, *Group Theoretic Methods in Bifurcation Theory*, Springer-Verlag, New York, 1979.

[21] ———, *Branching in the Presence of Symmetry*, CBMS Regional Conference Series in Applied Mathematics, 40, Society for Industrial and Applied Mathematics, Philadelphia, 1983.

[22] A. L. VANDERBAUWHEDE, *Local Bifurcation and Symmetry*, Pitman, Boston, 1982.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

AN EXPLICIT SOLUTION OF THE INVERSE PERIODIC PROBLEM FOR HILL'S EQUATION*

ALLAN FINKEL†, ELI ISAACSON‡ AND EUGENE TRUBOWITZ§

Abstract. Let the periodic spectrum of the Hill's operator $-d^2/dx^2+p(x)$ have n nonzero gaps. We give explicit formulas for the isospectral manifold of operators $-d^2/dx^2+q(x)$ having the same spectrum. This allows us to realize the isospectral manifold explicitly as a torus. What makes this possible is an explicit solution of the flow

$$\frac{d}{dt}q=\frac{d}{dx}\frac{\partial}{\partial q(x)}\Delta(\lambda,q)\bigg|_{\lambda=\mu_n(q)}$$

introduced by McKean and Trubowitz, where Δ is the discriminant and $\mu_n(q)$ is the nth Dirichlet eigenvalue. The general case (in which there are an infinite number of nonzero gaps) is handled by a limiting process.

Key words. inverse eigenvalue problem, Hill's equation, isospectral manifold, flow on manifold

AMS(MOS) subject classifications. Primary 34B30, 34B25, 34K10

1. Introduction. Let $-d^2/dx^2+q(x)$ be the Hill's operator with[1] $q\in L^2_{\mathbf{R}}(S^1)$. We give a simple explicit formula for the isospectral manifold of all potentials having the same periodic spectrum as q. The formula, which involves only the Floquet solutions and Dirichlet eigenfunctions of q, represents the isospectral manifold explicitly as a torus.

Its and Matveev [4] and Dubrovin, Matveev, and Novikov [3] have given another formula which involves the directional derivative of the theta function for the Jacobi variety associated with q. Our formula involves no algebraic geometry but cannot be used to solve the KdV flow explicitly.

Before stating our results we fix our notation. Let $y_1(x,\lambda,q)$, $y_2(x,\lambda,q)$ be the solutions of

$$-y''+q(x)y=\lambda y \qquad (-\infty<x<\infty)$$

with[2]

$$y_1(0)=y_2'(0)=1,$$
$$y_1'(0)=y_2(0)=0.$$

* Received by the editors May 3, 1983, and in revised form February 15, 1985.

† IBM, Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

‡ University of Wyoming, Laramie, Wyoming 82071. The research of this author was supported in part by the National Science Foundation under grant PHY-80-90179 while he was at Rockefeller University, New York, New York 10021.

§ Forschungsinstitut für Mathematik, ETH, Zurich, Switzerland. The research of this author was supported in part by the Sloan Foundation, and by the National Science Foundation under grant MCS-80-02561.

[1] $L^2_{\mathbf{R}}(S^1)$ is the Hilbert space of real-valued square integrable functions of period 1 with the norm $\|f\|^2=\int_0^1|f(x)|^2\,dx$.

[2] x,λ, or q will often be suppressed.

The discriminant is given by

$$\Delta(\lambda,q)\equiv y_1(1,\lambda,q)+y_2'(1,\lambda,q).$$

The zeros

$$\lambda_0(q)<\lambda_1(q)\leqq\lambda_2(q)<\cdots,$$

of $\Delta^2(\lambda)-4$ are the eigenvalues of $-d^2/dx^2+q(x)$ with eigenfunctions of period 2; equality means that $\lambda_{2n-1}=\lambda_{2n}$ is equivalently a double zero or a double eigenvalue. Moreover, $\Delta(\lambda_0)=+2$ and the corresponding eigenfunction is periodic, while $\Delta(\lambda_{2n-1})=\Delta(\lambda_{2n})=2(-1)^n$ $(n\geqq 1)$ and the corresponding eigenfunctions are periodic when n is even and antiperiodic when n is odd.[3]

The Floquet multipliers and corresponding Floquet solutions are

$$m_\pm(\lambda,q)\equiv\frac{\Delta(\lambda)}{2}\pm\frac{1}{2}(\Delta^2(\lambda)-4)^{1/2}$$

and

$$\begin{aligned}f_\pm(x,\lambda,q)&\equiv y_1(x,\lambda,q)+\left[\frac{m_\pm-y_1(1,\lambda)}{y_2(1,\lambda)}\right]y_2(x,\lambda,q)\\&=y_1(x,\lambda,q)+\left[\frac{y_1'(1,\lambda)}{m_\pm-y_2'(1,\lambda)}\right]y_2(x,\lambda,q),\end{aligned}$$

respectively. We note that

$$f_\pm(x+1,\lambda)=m_\pm(\lambda)f_\pm(x,\lambda)$$

so that $f_\pm(x,\lambda_n)$ is an eigenfunction when it is well defined. Also,[4]

$$\frac{\partial\Delta(\lambda)}{\partial q(x)}=y_2(1,\lambda)f_+(x,\lambda)f_-(x,\lambda).$$

The zeros

$$\mu_1(q)<\mu_2(q)<\cdots$$

of $y_2(1,\lambda,q)$ are the Dirichlet eigenvalues of q; i.e., there is a nontrivial solution of

$$-y''+q(x)y=\mu_n y$$

with $y(0)=y(1)=0$. The normalized Dirichlet eigenfunction corresponding to $\mu_n(q)$ is[5]

$$g_n(x,q)\equiv\frac{y_2(x,\mu_n(q),q)}{(\dot{y}_2(1,\mu_n)y_2'(1,\mu_n))^{1/2}}\qquad(n\geqq 1).$$

[3] The function f is periodic if $f(x+1)=f(x)$ and antiperiodic if $f(x+1)=-f(x)$.

[4] The gradient $\partial\Delta(\lambda,q)/\partial q(x)$ is defined to be the kernel of the directional derivative of $\Delta(\lambda,q)$; i.e.,

$$\frac{d}{d\varepsilon}\Delta(\lambda,q+\varepsilon v)\bigg|_{\varepsilon=0}=\int_0^1\frac{\partial\Delta(\lambda,q)}{\partial q(x)}v(x)\,dx$$

for all $v\in L^2_{\mathbf{R}}(S^1)$.

[5] $\cdot\equiv\partial/\partial\lambda$.

The Dirichlet eigenvalues interlace the periodic eigenvalues: $\lambda_{2n-1} \leqq \mu_n \leqq \lambda_{2n}$ $(n \geqq 1)$. Details can be found in [1], [6].

Fix $p \in L^2_{\mathbf{R}}(S^1)$ and define the isospectral manifold

$$L(p) \equiv \{q \in L^2_{\mathbf{R}}(S^1)\colon \lambda_n(q) = \lambda_n(p),\ n \geqq 0\}.$$

Then $q \in L(p)$ if and only if $\Delta(\lambda, q) = \Delta(\lambda, p)$ for all λ. Furthermore, McKean and Trubowitz [7] have shown that $L(p)$ is a (generically infinite dimensional) torus.

The isospectral manifold is realized as a product of circles as follows (see [7]). The map $q \in L^2_{\mathbf{R}}[0,1] \to (\mu_n(q), y_1(1, \mu_n(q), q); n \geqq 1)$ is one-to-one by a theorem of Borg.[6] However, $y_1(1, \mu_n, q) = \Delta(\mu_n)/2 \pm \frac{1}{2}(\Delta^2(\mu_n) - 4)^{1/2}$ for an appropriate choice of the sign. Since $\Delta(\lambda)$ is the same for each $q \in L(p)$, it follows that the map

$$q \in L(p) \to (\mathbf{p}_n(q), n \geqq 1)$$

is one-to-one where

$$\mathbf{p}_n(q) \equiv (\mu_n(q), \sigma_n(q)) \qquad (n \geqq 1)$$

and $\sigma_n(q) \in \{\pm\}$ is the sign of the radical in $y_1(1, \mu_n(q), q)$. Thus, $L(p)$ is mapped into the product of circles in Fig. 1.

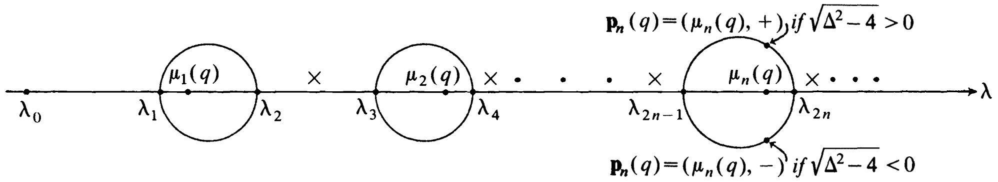


FIG. 1

McKean and Trubowitz show in addition that the map is onto and therefore a homeomorphism. Our main result is a concrete realization of this map. In fact in Theorems 1 and 2 we give a simple explicit construction of $L(p)$ in terms of p and $y_i(x, \lambda, p)$ $(i = 1, 2)$. Theorem 1 covers the case of finitely many gaps, and Theorem 2 covers the general case.

THEOREM 1. *Let* $p \in L^2_{\mathbf{R}}(S^1)$, $\tilde{\mathbf{p}}_i = (\omega_i, s_i) \in [\lambda_{2i-1}, \lambda_{2i}] \times \{\pm\}$ *be a point in the ith circle* $(1 \leqq i \leqq n)$, *and set*

$$\text{(1a)} \qquad q(x, \tilde{\mathbf{p}}_1, \cdots, \tilde{\mathbf{p}}_n) \equiv p(x) - 2\frac{d^2}{dx^2}\log W(f_1, g_1, f_2, g_2, \cdots, f_n, g_n)$$

where $f_i \equiv f_{s_i}(x, \omega_i, p)$, $g_i \equiv g_i(x, p)$, *and W denotes the Wronskian. Then q is the unique point in* $L(p)$ *with*

$$\text{(1b)} \qquad \mathbf{p}_i(q) = \begin{cases} \tilde{\mathbf{p}}_i, & 1 \leqq i \leqq n, \\ \mathbf{p}_i(p), & i > n. \end{cases}$$

In particular, $L(p)$ *is homeomorphic to an n-dimensional torus when p has n nonzero gaps.*

[6] Levinson [5] has shown that the map $q \in L^2_{\mathbf{R}}[0,1] \to (\mu_n(q), y_2'(1, \mu_n(q), q); n \geqq 1)$ is one-to-one, and the equality $y_1(1, \mu_n) = 1/y_2'(1, \mu_n)$ follows from the Wronskian identity.

Remark. The Wronskian is defined by

$$W(f_1,g_1,\cdots,f_n,g_n)\equiv$$

$$\det\begin{vmatrix} f_1 & g_1 & \cdots & f_n & g_n \\ f_1' & g_1' & \cdots & f_n' & g_n' \\ -\omega_1 f_1 & -\mu_1 g_1 & \cdots & -\omega_n f_n & -\mu_n g_n \\ -\omega_1 f_1' & -\mu_1 g_1' & \cdots & -\omega_n f_n' & -\mu_n g_n' \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-\omega_1)^{n-1} f_1 & (-\mu_1)^{n-1} g_1 & \cdots & (-\omega_n)^{n-1} f_n & (-\mu_n)^{n-1} g_n \\ (-\omega_1)^{n-1} f_1' & (-\mu_1)^{n-1} g_1' & \cdots & (-\omega_n)^{n-1} f_n' & (-\mu_n)^{n-1} g_n' \end{vmatrix}.$$

The standard definition is not used because p is only in $L^2_{\mathbf{R}}(S^1)$. The definition above can be obtained from the standard definition when p is sufficiently smooth by using the differential equations for f,g to eliminate the higher order derivatives.

The final conclusion of Theorem 1 follows from (1). In fact, $\mathbf{p}_i(q)=\mathbf{p}_i(p)$ whenever $q\in L(p)$ and $\lambda_{2i-1}=\lambda_{2i}$. Therefore, the map $q\in L(p)\to(\mathbf{p}_i(q),i\geqq 1)$, which is one-to-one by Borg's theorem, is onto by formula (1). Consequently, formula (1) gives an explicit construction of all points on $L(p)$.

The remainder of the proof of Theorem 1 is given in §3. The proof depends on the solution of certain flows on $L(p)$ described in §2. Theorem 2 will also be proven in §3.

THEOREM 2. *Let* $p\in L^2_{\mathbf{R}}(S^1)$, *and let* $\tilde{\mathbf{p}}_i=(\omega_i,s_i)\in[\lambda_{2i-1},\lambda_{2i}]\times\{\pm\}$ $(i\geqq 1)$ *be a point on the ith circle. Then the sequence*

$$q_n(x)\equiv p(x)-2\frac{d^2}{dx^2}\log W(f_1,g_1,\cdots,f_n,g_n) \tag{2a}$$

converges strongly in $L^2_{\mathbf{R}}(S^1)$ *to the unique point* $q\in L(p)$ *with*

$$\mathbf{p}_i(q)=\tilde{\mathbf{p}}_i \qquad (i\geqq 1). \tag{2b}$$

This result exhibits every point in $L(p)$ as the (strong) limit of an appropriate sequence constructed explicitly from $y_i(x,\lambda,p)(i=1,2)$ and the parameters on the torus. This gives another proof that the continuous, one-to-one map $q\in L(p)\to(\mathbf{p}_i(q),i\geqq 1)$ is also onto. That is, $L(p)$ is homeomorphic to a torus.

2. Some flows on $L(p)$. In order to obtain formula (1), we find an exact solution of the differential equation

$$\frac{d}{dt}q(x,t)=\frac{d}{dx}\left.\frac{\partial}{\partial q(x)}\Delta(\lambda,q(\cdot,t))\right|_{\lambda=\mu_n(q(\cdot,t))} \tag{3}$$

which was studied in [7]. Theorem 3 shows that this differential equation generates a flow on $L(p)$. In Theorem 4 we solve (3) explicitly, and in Theorem 5 we eliminate the time parameter to obtain a formula for certain points on $L(p)$.

THEOREM 3. *Let* $p\in L^2_{\mathbf{R}}(S^1)$. *Then*

$$Z_n(q)\equiv\frac{d}{dx}\left.\frac{\partial\Delta(\lambda,q)}{\partial q(x)}\right|_{\lambda=\mu_n(q)}$$

is a vector field on $L(p)$. That is, a solution of the differential equation

$$\frac{d}{dt}q(x,t)=\frac{d}{dx}\frac{\partial\Delta(\lambda)}{\partial q(x)}\bigg|_{\lambda=\mu_n(q(\cdot,t))}$$

with initial data in $L(p)$ stays in $L(p)$ for all time.

This flow was described first in [7] where the following geometric picture is given. Under the flow Z_n the point $\mathbf{p}_i(q(\cdot,t))$ $(i\neq n)$ remains fixed on its circle, while the point $\mathbf{p}_n(q(\cdot,t))$ moves clockwise around its circle without pausing. It moves in such a way that

$$\frac{d\mu_n}{dt}=\frac{1}{2}\left(\Delta^2(\mu_n)-4\right)^{1/2},$$

the radical having the sign of $\mathbf{p}_n(q(\cdot,t))$.

Proof of Theorem 3. Let $q(x,t)$ solve the differential equation

$$\frac{d}{dt}q(x,t)=\frac{d}{dx}\frac{\partial\Delta(\mu)}{\partial q(x)}\bigg|_{\mu=\mu_n(q(\cdot,t))}$$

with $q(x,0)\in L(p)$. We show that $\frac{d}{dt}\Delta(\lambda,q(\cdot,t))=0$ so that $\Delta(\lambda,q(\cdot,t))=\Delta(\lambda,q(\cdot,0))$ for all $\lambda\in C$. Indeed,

$$\begin{aligned}\frac{d}{dt}\Delta(\lambda,q(\cdot,t))&=\int_0^1\frac{\partial}{\partial q(x)}\Delta(\lambda,q(\cdot,t))\frac{dq}{dt}\,dx\\&=\int_0^1\frac{\partial\Delta(\lambda)}{\partial q(x)}\frac{d}{dx}\frac{\partial\Delta(\mu)}{\partial q(x)}\bigg|_{\mu=\mu_n(q(\cdot,t))}dx,\end{aligned}$$

and we need only show that

$$\int_0^1\frac{\partial\Delta(\lambda,q)}{\partial q(x)}\frac{d}{dx}\frac{\partial\Delta(\lambda',q)}{\partial q(x)}\,dx=0$$

for all $\lambda,\lambda'\in C$ and all $q\in L^2_{\mathbf{R}}(S^1)$. Now, for $\lambda,\lambda'\neq\mu_i(q)$ $(i\geq 1)$ and $\lambda\neq\lambda'$ we have

$$\begin{aligned}&2\int_0^1\frac{\partial\Delta(\lambda)}{\partial q(x)}\frac{d}{dx}\frac{\partial\Delta(\lambda')}{\partial q(x)}\,dx\\&=2\int_0^1 y_2(1,\lambda)f_+(x,\lambda)f_-(x,\lambda)\frac{d}{dx}\left(y_2(1,\lambda')f_+(x,\lambda')f_-(x,\lambda')\right)dx\\&=y_2(1,\lambda)y_2(1,\lambda')\int_0^1\bigg[f_+(x,\lambda)f_-(x,\lambda)\frac{d}{dx}\left(f_+(x,\lambda')f_-(x,\lambda')\right)\\&\qquad\qquad-f_+(x,\lambda')f_-(x,\lambda')\frac{d}{dx}\left(f_+(x,\lambda)f_-(x,\lambda)\right)\bigg]dx\\&=\frac{y_2(1,\lambda)y_2(1,\lambda')}{\lambda-\lambda'}\int_0^1\frac{d}{dx}\left(W\left(f_+(x,\lambda),f_+(x,\lambda')\right)W\left(f_-(x,\lambda),f_-(x,\lambda')\right)\right)dx\\&=0.\end{aligned}$$

The extension to λ, $\lambda'=\mu_i$ $(i\geq 1)$ and $\lambda=\lambda'$ follows from continuity.

THEOREM 4. *Let $q_0\in L(p)$, and let $\mu_n(t)$ denote the unique solution of*

$$\frac{d}{dt}\mu_n=\frac{1}{2}\left(\Delta^2(\mu_n)-4\right)^{1/2}$$

for which the point $\mathbf{p}_n(t) \equiv (\mu_n(t), \sigma_n(t))$ *starts at* $\mathbf{p}_n(q_0)$ *and moves clockwise around its circle without pausing. Then*

$$q(x,t) = q_0(x) - 2\frac{d^2}{dx^2}\log W\left[f_{\sigma_n(t)}(x,\mu_n(t),q_0), g_n(x,q_0)\right]$$

is the integral curve of Z_n *passing through* q_0.

The sign of the radical in the differential equation is taken to be $\sigma_n(t)$. The requirements on $\mathbf{p}_n(t)$ actually determine $\sigma_n(t)$ in addition to specifying a unique solution.

Proof of Theorem 4. We show first that $\chi(x,\mu) \equiv W(f_+(x,\mu), g_n(x))$ does not vanish for $\lambda_{2n-1} \leqq \mu \leqq \lambda_{2n}$. The proof for f_- is similar. It can be shown that $\chi(x,\mu_n)$, $\chi(0,\mu)$, $\chi(1,\mu)$ are all positive. Therefore, if $\chi(x,\mu)=0$, then there is a μ closest to μ_n for which χ vanishes, say $\chi(x^0,\mu^0)=0$. It follows that $(d/dx)\chi(x^0,\mu^0)$ also vanishes. These two conditions imply that $f_+(x^0,\mu^0)=0$ and $g_n(x^0)=0$. But this is impossible because both $f_+(x+x^0,\mu^0)$ and $g_n(x+x^0)$ would be Dirichlet eigenfunctions for the translated potential $q(x+x^0)$ with corresponding eigenvalues μ^0, μ_n lying in the same gap $[\lambda_{2n-1}, \lambda_{2n}]$.

To complete the proof we show that

$$\frac{dq}{dt}(x,t) = \frac{d}{dx}\frac{\partial}{\partial q(x)}\Delta(\mu, q(\cdot,t))\bigg|_{\mu=\mu_n(t)}.$$

It can be shown that

$$\frac{d}{dt}q(x,t) = -2\frac{d}{dx}\frac{g_n(x)}{W(f_{\sigma_n},g_n)}\left[f_{\sigma_n} + (\mu_n(t)-\mu_n)\frac{g_n(x)}{W(f_{\sigma_n},g_n)}W(f_{\sigma_n},\dot{f}_{\sigma_n})\right]\frac{d\mu_n(t)}{dt},$$

$$y_2(x,\mu_n(t),q(\cdot,t)) = \frac{g_n(x)}{W(f_{\sigma_n},g_n)},$$

and

$$f_{\sigma_n}(x,\mu_n(t),q(\cdot,t)) = f_{\sigma_n}(x,\mu_n(t)) + (\mu_n(t)-\mu_n)\frac{g_n(x)}{W(f_{\sigma_n},g_n)}\;W(f_{\sigma_n},\dot{f}_{\sigma_n})\big|_{q(\cdot,0)}.$$

Therefore

$$\begin{aligned}\frac{d}{dt}q(x,t) &= -2\frac{d}{dx}y_2(x,\mu_n(t))f_{\sigma_n}(x,\mu_n(t))\frac{d\mu_n(t)}{dt}\bigg|_{q(\cdot,t)}\\ &= \frac{d}{dx}\frac{\partial}{\partial q(x)}\Delta(\mu,q(\cdot,t))\bigg|_{\mu=\mu_n(t)}\end{aligned}$$

since

$$\frac{d\mu_n(t)}{dt} = \frac{1}{2}\left(y_1(1,\mu_n(t)) - y_2'(1,\mu_n(t))\right)\bigg|_{q(\cdot,t)},$$

so that in the limit $\mu \to \mu_n(t)$

$$-y_2(x,\mu_n(t))\frac{d\mu_n(t)}{dt} = y_2(1,\mu_n(t))f_{-\sigma_n}(x,\mu_n(t)).$$

THEOREM 5. *Let* $\tilde{\mathbf{p}}_n=(\omega_n,s_n)\in[\lambda_{2n-1},\lambda_{2n}]\times\{\pm\}$ *be a point in the nth circle. Then*

$$q(x,\tilde{\mathbf{p}}_n)\equiv p(x)-2\frac{d^2}{dx^2}\log W\left[f_{s_n}(x,\omega_n,p),g_n(x,p)\right]$$

is the unique point in $L(p)$ *with*

$$\mathbf{p}_i(q(\cdot,\tilde{\mathbf{p}}_n))=\begin{cases}\mathbf{p}_i(p), & i\neq n,\\ \tilde{\mathbf{p}}_n, & i=n.\end{cases}$$

This result is a straightforward corollary of Theorem 4. It removes the dependence on t allowing us to construct points in $L(p)$ using only the parameters in Borg's theorem and information obtained from p.

3. Proofs of Theorems 1 and 2.

Proof of Theorem 1. We use induction and Theorem 5. Suppose the result holds for some $n\geqq 1$, and let q^k denote the potential satisfying

$$\mathbf{p}_i(q^k)=\begin{cases}\tilde{\mathbf{p}}_i, & 1\leqq i\leqq k,\\ \mathbf{p}_i(p), & i>k.\end{cases}$$

Then by Theorem 5,

$$q^{n+1}=q^n-2\frac{d^2}{dx^2}\log W(f^n_{n+1},g^n_{n+1})$$

where f^n_{n+1},g^n_{n+1} are the Floquet solution and Dirichlet eigenfunction of q^n. Thus,

$$q^{n+1}=p-2\frac{d^2}{dx^2}\log W(f_1,g_1,\cdots,f_n,g_n)-2\frac{d^2}{dx^2}\log W(f^n_{n+1},g^n_{n+1}). \tag{4}$$

Direct calculation shows that

$$f^n_{n+1}=\frac{W(f_1,g_1,\cdots,f_n,g_n,f_{n+1})}{W(f_1,g_1,\cdots,f_n,g_n)\Pi^n_{j=1}(\omega_j-\omega_{n+1})}, \tag{5}$$

$$y^n_2(x,\mu_{n+1})=\frac{W(f_1,g_1,\cdots,f_n,g_n,y_2(x,\mu_{n+1}))}{W(f_1,g_1,\cdots,f_n,g_n)\Pi^n_{j=1}(\mu_j-\mu_{n+1})}, \tag{6}$$

$$W(f^n_{n+1},y^n_2)=-(y^n_2)^2\frac{d}{dx}\frac{f^n_{n+1}}{y^n_2}, \tag{7}$$

and (see Deift–Trubowitz [2])

(8)

$$W^2(h_1,\cdots,h_n)\frac{d}{dx}\frac{W(h_1,\cdots,h_{n-1},h_{n+1})}{W(h_1,\cdots,h_n)}=W(h_1,\cdots,h_n,h_{n+1})W(h_1,\cdots,h_{n-1}).$$

Consequently, substituting (5) and (6) into (7) and using (8) gives

$$W(f^n_{n+1},y^n_2)=\frac{W(f_1,g_1,\cdots,f_n,g_n,f_{n+1},g_{n+1})}{W(f_1,g_1,\cdots,f_n,g_n)\Pi^n_{j=1}(\omega_j-\omega_{n+1})(\mu_j-\mu_{n+1})}.$$

Combining this with (4) yields the result.

Proof of Theorem 2. First, $\int_0^1 q^2 dx$ is constant on $L(p)$ (see [7]), and $\Delta(\lambda,q)$ is weakly sequentially continuous. Therefore, every subsequence of $\{q_n\}$ has a weakly convergent subsequence which must converge to the point q with $\mathbf{p}_i(q)=\tilde{\mathbf{p}}_i$ $(i\geqq 1)$. Consequently, the original sequence converges weakly to q. The convergence is actually strong since all points in $L(p)$ have the same norm.

REFERENCES

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

[2] P. Deift and E. Trubowitz, *Inverse scattering on the line*, Comm. Pure. Appl. Math., 32 (1979), pp. 121–251.

[3] B. A. Dubrovin, V. Matveev and S. Novikov, *Nonlinear equations of Korteweg–deVries type, finite zone linear operators, and Abelian varieties*, Uspekhi. Mat. Nauk, 31 (1976), pp. 55–136; Russ. Math. Surveys, 31 (1976), pp. 59–146.

[4] A. K. Its and V. B. Matveev, *On Hill's operator with a finite number of lacunae*, Funct. Anal. Appl., 9 (1975), pp. 65–66.

[5] N. Levinson, *The inverse Sturm–Liouville problem*, Math. Tidskr., 4 (1949), pp. 25–30.

[6] W. Magnus and W. Winkler, *Hill's Equation*, Wiley-Interscience, New York, 1966.

[7] H. P. McKean and E. Trubowitz, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29 (1976), pp. 143–226.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

004

OSCILLATORY SECOND ORDER LINEAR DIFFERENCE EQUATIONS AND RICCATI EQUATIONS*

JOHN W. HOOKER†, MAN KAM KWONG‡ AND WILLIAM T. PATULA†

Abstract. Oscillation criteria are established for the equation $c_nx_{n+1}+c_{n-1}x_{n-1}=b_nx_n$, $c_n>0$, involving asymptotic behavior of the quantity $\alpha_{n,m}=4\Pi_{j=0}^{m}(4q_{n+j})^{-1}$, where $q_n=c_n^2/(b_nb_{n+1})$. We also show that the given equation is oscillatory if $y_{n+1}+y_{n-1}=(q_n^{-1}-1)y_n$ is oscillatory. This result is then employed to obtain several new oscillation criteria. Riccati difference equations are used to prove the basic results.

Key words. oscillation, nonoscillation, linear difference equation, Riccati transformation

AMS(MOS) subject classifications. Primary 39A10, 39A12

1. Introduction and main results. In this paper we continue our investigation begun in [4] and [5] of oscillation criteria for solutions of the second-order linear difference equation

$$c_nx_{n+1}+c_{n-1}x_{n-1}=b_nx_n,\qquad n=1,2,3,\cdots,\tag{1.1}$$

with $c_n>0$ and $b_n>0$ for all $n\geqq 0$, using Riccati transformation methods.

This equation models, for example, the amplitude of oscillation of the weights on a discretely weighted vibrating string [1, pp. 15–17]. It is equivalent to the self-adjoint equation

$$-\Delta(c_{n-1}\Delta x_{n-1})+a_nx_n=0,$$

where $a_n=b_n-c_n-c_{n-1}$ and Δ is the forward difference operator $\Delta u_n=u_{n+1}-u_n$. A nontrivial solution of (1.1) is called oscillatory if for every $N>0$ there exists an $n\geqq N$ such that $x_nx_{n+1}\leqq 0$.

Either all nontrivial solutions of (1.1) are oscillatory or none are (see [2, p. 153]), so (1.1) may be classified as oscillatory or nonoscillatory. The assumption $b_n>0$ is made because if $b_{n_k}\leqq 0$ for some subsequence $n_k\to\infty$, then (1.1) clearly must be oscillatory [6, Lemma 3].

Suppose that (1.1) is nonoscillatory, and let $\{x_n\}$, $n\geqq 0$, be a solution of (1.1) such that $x_n>0$, $n\geqq N$, for some N. The substitutions $r_n=x_{n+1}/x_n$, $z_n=c_nx_{n+1}/x_n$ and $s_n=b_{n+1}x_{n+1}/(c_nx_n)$, $n\geqq N$, lead, respectively, to the difference equations

$$c_nr_n+c_{n-1}/r_{n-1}=b_n,\qquad n>N,\tag{1.2}$$

$$z_n+c_{n-1}^2/z_{n-1}=b_n,\qquad n>N,\tag{1.3}$$

and

$$q_ns_n+1/s_{n-1}=1,\qquad n>N,\tag{1.4}$$

*Received by the editors April 22, 1985; accepted for publication (in revised form) September 3, 1985.

† Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901.

‡ Department of Mathematics, Northern Illinois University, DeKalb, Illinois 60115.

where

$$q_n = \frac{c_n^2}{b_n b_{n+1}}, \qquad n \geqq 1. \tag{1.5}$$

Equations (1.2), (1.3), and (1.4) we call equations of Riccati type, since the substitutions which lead to them are discrete analogues of the Riccati transformation for ordinary differential equations. The above remarks lead immediately to the following theorem.

THEOREM 1.1 [4, Thm. 2]. *The following conditions are equivalent.*

(i) *Equation* (1.1) *is nonoscillatory.*

(ii) *Equation* (1.2) *has a positive solution* $\{r_n\}$, $n \geqq N$, *for some* $N > 0$.

(iii) *Equation* (1.3) *has a positive solution* $\{z_n\}$, $n \geqq N$, *for some* $N > 0$.

(iv) *Equation* (1.4) *has a positive solution* $\{s_n\}$, $n \geqq N$, *for some* $N > 0$.

For convenient reference, and because the quantity $q_n = c_n^2/(b_n b_{n+1})$ defined by (1.5) will play a central role in the results below, we restate here in terms of q_n three more theorems proved in [4].

THEOREM 1.2 [4, Thm. 5]. *If* $q_n \geqq 1/(4-\varepsilon)$ *for some* $\varepsilon > 0$, *for all sufficiently large* n, *then* (1.1) *is oscillatory.*

THEOREM 1.3 [4, Thm. 6]. *If* $q_n \leqq \frac{1}{4}$ *for all sufficiently large* n, *then* (1.1) *is nonoscillatory.*

THEOREM 1.4 [4, Thm. 7]. *If* $q_{n_k} \geqq 1$ *for a sequence* $n_k \to \infty$, *then* (1.1) *is oscillatory.*

In [5], the following necessary condition for nonoscillation of (1.1) is presented.

THEOREM 1.5 [5, Thm. 2.3]. *Suppose* (1.1) *is nonoscillatory. Then there exists* $N > 0$ *such that for any* $n \geqq N$ *and any* $m \geqq 0$,

$$q_n q_{n+1} \cdots q_{n+m} < 4^{-m}. \tag{1.6}$$

We can rewrite condition (1.6) as

$$\alpha_{n,m} > 1, \tag{1.7}$$

where $\alpha_{n,m}$ is defined as

$$\alpha_{n,m} = 4 \prod_{j=0}^{m} (4q_{n+j})^{-1}, \qquad n \geqq 1, \quad m \geqq 0. \tag{1.8}$$

The contrapositive of Theorem 1.5 says that (1.1) is oscillatory if for every $K > 0$ there exists $n \geqq K$ such that $\alpha_{n,m} \leqq 1$ for some $m \geqq 0$. Since this statement involves two variables m and n, one can state corollaries in various forms. For example, we have:

COROLLARY 1.1. *If for some* $M \geqq 0$ *there exists a sequence* $n_k \to \infty$ *such that*

$$\alpha_{n_k,M} \leqq 1,$$

then (1.1) *is oscillatory.*

In particular, for $M = 0$, we have $\alpha_{n_k,0} = 1/q_{n_k}$. Thus Corollary 1.1 implies that if $1/q_{n_k} \leqq 1$ for some sequence $n_k \to \infty$, then (1.1) is oscillatory. This is precisely Theorem 1.4, so Corollary 1.1 generalizes that theorem.

The following theorem is also an immediate corollary of Theorem 1.5.

THEOREM 1.6. *If for every* $K > 0$ *there exists* $N \geqq K$ *such that*

$$\liminf_{m \to \infty} \alpha_{N,m} < 1, \tag{1.9}$$

then (1.1) *is oscillatory.*

On the other hand, if condition (1.9) is not satisfied for arbitrarily large N, then (1.1) may or may not be oscillatory, as the following examples show.

Example 1. Consider the equation

$$x_{n+1}+x_{n-1}=2x_n,$$

which has linearly independent solutions $u_n=1$, $v_n=n$, $n=0,1,2,\cdots$. This nonoscillatory example has $q_n=c_n^2/(b_nb_{n+1})\equiv\frac{1}{4}$, and $\alpha_{n,m}\equiv 4$.

Example 2. Consider (1.1) with $c_n\equiv 1$, $b_{2n}=2^{-1}4^{2-n}$, $b_{2n+1}=4^{n-1}$, $n=1,2,3,\cdots$. Then $q_{2n-1}=\frac{1}{2}$, $q_{2n}=\frac{1}{8}$, and it follows that $\alpha_{N,m}=4$ if m is odd, $\alpha_{N,m}=2$ if N is odd and m is even, while $\alpha_{N,m}=8$ if N and m are both even. Thus $\liminf_{m\to\infty}\alpha_{N,m}$ equals 2 if N is odd and 4 if N is even, so (1.9) fails to hold. It was shown in [5, Ex. 2.1] that (1.1) is oscillatory in this example. Indeed, this follows immediately from our next theorem.

These two examples lead us to ask what additional conditions are sufficient for (1.1) to be oscillatory if condition (1.9) is not satisfied. Theorems 1.7–1.9 address this question.

THEOREM 1.7. *If for some* $N>0$,

$$\liminf_{m\to\infty}\alpha_{N,m}\neq\limsup_{m\to\infty}\alpha_{N,m}, \tag{1.10}$$

then (1.1) *is oscillatory.*

(*Note.* If (1.10) holds for some $N>0$, then it holds for all $N>0$ by definition of $\alpha_{n,m}$.)

Proof. Suppose that (1.1) is nonoscillatory. Then by Theorem 1.1 there is a positive sequence s_n which satisfies the Riccati equation (1.4) $q_js_j+s_{j-1}^{-1}=1$ for all sufficiently large j, say $j\geqq N$. By the note following the statement of the theorem, we may take this to be the same value N as in the hypotheses of the theorem. Multiplying (1.4) by q_j^{-1} yields

$$q_j^{-1}=s_j+(q_js_{j-1})^{-1},\qquad j\geqq N, \tag{1.11}$$

so

$$q_j^{-1}q_{j+1}^{-1}=\left(s_j+(q_js_{j-1})^{-1}\right)\left(s_{j+1}+(q_{j+1}s_j)^{-1}\right). \tag{1.12}$$

From (1.12), we obtain

$$q_j^{-1}=q_{j+1}s_js_{j+1}\left(1+(q_js_js_{j-1})^{-1}\right)\left(1+(q_{j+1}s_{j+1}s_j)^{-1}\right),$$

so

$$\begin{aligned} q_j^{-1}&=\left(1+(q_js_js_{j-1})^{-1}\right)(q_{j+1}s_js_{j+1}+1)\\ &=\left(1+\beta_j^{-1}\right)(1+\beta_{j+1}),\qquad j\geqq N, \end{aligned} \tag{1.13}$$

where we define

$$\beta_j=q_js_js_{j-1}.$$

Note that $\beta_j>0$, $j\geqq N$.

From (1.8) and (1.13) we have

$$\alpha_{N,m}=4\prod_{j=0}^{m}4^{-1}\left(1+\beta_{N+j}^{-1}\right)\left(1+\beta_{N+j+1}\right),\qquad m\geqq 0,$$

$$=4\left(1+\beta_N^{-1}\right)\left(1+\beta_{N+m+1}\right)\prod_{i=1}^{m}4^{-1}\left(1+\beta_{N+i}^{-1}\right)\left(1+\beta_{N+i}\right),$$

from which some elementary algebra leads us to

$$\alpha_{N,m}=4\left(1+\beta_N^{-1}\right)\left(1+\beta_{N+m+1}\right)\prod_{i=1}^{m}\left(1+4^{-1}\beta_{N+i}^{-1}(\beta_{N+i}-1)^2\right),$$

for $m\geqq 0$. We rewrite this as

$$\alpha_{N,m}=4\left(1+\beta_N^{-1}\right)\left(1+\beta_{N+m+1}\right)\prod_{i=1}^{m}(1+A_i),\tag{1.14}$$

where

$$A_i=4^{-1}\beta_{N+i}^{-1}(\beta_{N+i}-1)^2\geqq 0,\qquad i\geqq 1.\tag{1.15}$$

Now from the hypotheses we must have $\liminf_{m\to\infty}\alpha_{N,m}<\infty$. Hence there exists a finite bound B such that

$$\alpha_{N,m_k}\leqq B\tag{1.16}$$

for some sequence of subscripts $m_k\to\infty$. Since $\beta_j>0$ for all j, (1.14) and (1.16) imply that

$$\prod_{i=1}^{m_k}(1+A_i)\text{ is bounded.}$$

Since $m_k\to\infty$ and $A_i\geqq 0$, it follows that

$$\prod_{i=1}^{\infty}(1+A_i)\text{ is bounded,}\tag{1.17}$$

which implies that

$$\sum_{i=1}^{\infty}A_i\text{ is finite.}\tag{1.18}$$

Therefore $A_i\to 0$ as $i\to\infty$. Thus, by (1.15) we have

$$\beta_j\to 1\quad\text{as } j\to\infty.\tag{1.19}$$

That is,

$$q_js_js_{j-1}\to 1\quad\text{as } j\to\infty.\tag{1.20}$$

Since $A_i\geqq 0$, (1.14), (1.17), and (1.19) imply that $\lim_{m\to\infty}\alpha_{N,m}$ exists, a contradiction which proves the theorem.

In the next theorem we refer to l^2, the space of square summable sequences. We remark, as in the note following the statement of Theorem 1.7, that if the hypotheses of Theorems 1.8 and 1.9 hold for some $N>0$, they hold for all $N>0$.

THEOREM 1.8. *If* $\liminf_{n\to\infty}\alpha_{N,m}<\infty$ *for some* $N>0$ *and if the sequence* $\{q_j^{-1}-4\}\notin l^2$, *then* (1.1) *is oscillatory.*

Proof. Assume (1.1) is nonoscillatory. Because of the assumption $\liminf\alpha_{N,m}<\infty$, we can proceed as in the proof of Theorem 1.7 to arrive at (1.18) and (1.19), and thus by (1.15) we have

$$\sum_{j=1}^{\infty}(\beta_j-1)^2/(4\beta_j)<\infty. \tag{1.21}$$

Since $\beta_j\to 1$ as $j\to\infty$, it follows from (1.21) that

$$(\beta_j-1)\in l^2. \tag{1.22}$$

Expanding the right side of (1.13) leads to

$$q_j^{-1}=1+\beta_j^{-1}+\beta_{j+1}+\beta_j^{-1}\beta_{j+1},$$

so

$$\begin{aligned} q_j^{-1}-4&=2\beta_j^{-1}-2+\beta_{j+1}-1+\beta_j^{-1}\beta_{j+1}-\beta_j^{-1}\\ &=2\beta_j^{-1}(1-\beta_j)+(\beta_{j+1}-1)+\beta_j^{-1}(\beta_{j+1}-1). \end{aligned} \tag{1.23}$$

By (1.19) and (1.22), each term on the right in (1.23) is in l^2, so $\{q_j^{-1}-4\}\in l^2$. Therefore, if $\{q_j^{-1}-4\}\notin l^2$, (1.1) must be oscillatory, as claimed.

COROLLARY 1.2. *If* $\liminf_{m\to\infty}\alpha_{N,m}<\infty$ *and* $\lim_{j\to\infty}q_j^{-1}\neq 4$, *then* (1.1) *is oscillatory.*

Proof. The condition $\lim_{j\to\infty}q_j^{-1}\neq 4$ implies that $\{q_j^{-1}-4\}\notin l^2$.

In connection with Theorem 1.8 and Corollary 1.2, we again call attention to Example 1 above, a nonoscillatory example with $\lim_{m\to\infty}\alpha_{N,m}=4$ and $q_n\equiv\frac{1}{4}$.

THEOREM 1.9. *If for some* N *the sequence* $\{\alpha_{N,m}\}$ *is eventually monotone increasing in* m, *then* (1.1) *is nonoscillatory.*

Proof. If $\alpha_{N,m}$ is eventually monotone increasing in m, there exists $M>0$ such that for all $m>M$, $\alpha_{N,m}\geqq\alpha_{N,m-1}$. From (1.8), this implies that $q_{N+m}\leqq\frac{1}{4}$ for all $m\geqq M$. Thus (1.1) is nonoscillatory by Theorem 1.3.

The following example shows that the monotonicity hypothesis of Theorem 1.9 cannot be replaced by the condition that $\alpha_{N,m}\to\infty$ as $m\to\infty$.

Example 3. For (1.1), let $b_n\equiv 1$, $c_{3n}=(4\sqrt{\sqrt{2}})^{-1}$, $c_{3n+1}=(\sqrt{2})^{-1}$ and $c_{3n+2}=(\sqrt{2\sqrt{2}})^{-1}$. Then $q_{3n}=(16\sqrt{2})^{-1}$, $q_{3n+1}=\frac{1}{2}$ and $q_{3n+2}=(2\sqrt{2})^{-1}$, so

$$\begin{aligned} \alpha_{2,3m+1}&=\frac{b_{3m+3}b_{3m+4}}{4c^2_{3m+3}}\cdot\frac{b_{3m+2}b_{3m+3}}{4c^2_{3m+2}}\cdot\frac{b_{3m+1}b_{3m+2}}{4c^2_{3m+1}}\cdot\alpha_{2,3m-2}\\ &=\frac{16\sqrt{2}}{4}\cdot\frac{2\sqrt{2}}{4}\cdot\frac{2}{4}\alpha_{2,3m-2}\\ &=2\alpha_{2,3m-2}. \end{aligned} \tag{1.24}$$

Thus $\alpha_{2,j}\to\infty$ as $j\to\infty$. Also,

$$\begin{aligned} \alpha_{2,3m+2}&=\frac{b_{3m+4}b_{3m+5}}{4c^2_{3m+4}}\cdot\alpha_{2,3m+1}\\ &=\frac{1}{2}\alpha_{2,3m+1}. \end{aligned} \tag{1.25}$$

Thus (1.24) and (1.25) imply that $\alpha_{2,j}\to\infty$ but not monotonically.

Next, define $a_n = b_n - c_n - c_{n-1}$. Then

$$a_{3n} = 1 - \left(4\sqrt{\sqrt{2}}\right)^{-1} - \left(\sqrt{2\sqrt{2}}\right)^{-1},$$

$$a_{3n+1} = 1 - (\sqrt{2})^{-1} - \left(4\sqrt{\sqrt{2}}\right)^{-1},$$

and

$$a_{3n+2} = 1 - (2\sqrt{2})^{-1} - (\sqrt{2})^{-1}.$$

So, for all n,

$$a_{3n} + a_{3n+1} + a_{3n+2} \cong 3 - 3.0238 = -.0238.$$

Therefore $\Sigma^{\infty} a_n = -\infty$. Since the sequence $\{b_n\}$ is bounded, equation (1.1) must be oscillatory by [5, Thm. 3.7] or [3, Thm. 4].

2. An extension theorem for the case $q_n < 1$. In this section we introduce a technique for extending known oscillation criteria for equation (1.1) by making use of the following theorem, where $q_n = c_n^2/(b_n b_{n+1})$, as before. We assume throughout that $q_n < 1$ for all sufficiently large n, since (1.1) must be oscillatory if $q_{n_k} \geqq 1$ for a sequence $n_k \to \infty$, by Theorem 1.4.

THEOREM 2.1. *Given* (1.1), *let* $B_n = q_n^{-1} - 1 > 0$, $n \geqq N$. *If the equation*

$$y_{n+1} + y_{n-1} = B_n y_n \tag{2.1}$$

is oscillatory, then (1.1) *is oscillatory also.*

To prove this theorem we will make use of the following simple comparison lemma.

LEMMA 2.1. *If* $\{u_n\}$, $n \geqq N$, *is a positive solution of*

$$u_n + \frac{1}{u_{n-1}} = \tilde{B}_n, \tag{2.2}$$

and if $B_n \geqq \tilde{B}_n > 0$ *for all* $n > N$, *then*

$$v_n + \frac{1}{v_{n-1}} = B_n \tag{2.3}$$

has a solution, with $v_n \geqq u_n$ *for* $n \geqq N$.

Proof. Given such a sequence u_n, let $v_N = u_N$ and define

$$v_{N+1} = B_{N+1} - \frac{1}{v_N} \geqq \tilde{B}_{N+1} - \frac{1}{u_N} = u_{N+1}.$$

Thus v_n satisfies (2.3) for $n = N+1$, and $v_{N+1} \geqq u_{N+1} > 0$. Proceeding inductively, we construct the required solution v_n.

Proof of Theorem 2.1. Assume (1.1) is nonoscillatory, and let x_n be a solution with $x_n > 0$, $n \geqq N$. Then by Theorem 1.1, $z_n = c_n x_{n+1}/x_n$ is a positive solution of (1.3) for $n > N$. If we take (1.3) for n and for $n+1$, multiply corresponding sides, and divide the result by c_n^2, we obtain

$$\frac{z_n z_{n+1}}{c_n^2} + 1 + \left(\frac{z_n z_{n+1}}{c_n^2}\right) \Big/ \left(\frac{z_{n-1} z_n}{c_{n-1}^2}\right) + \left(\frac{z_{n-1} z_n}{c_{n-1}^2}\right)^{-1} = \frac{1}{q_n}.$$

We write this as

$$r_n + 1 + \frac{r_n}{r_{n-1}} + \frac{1}{r_{n-1}} = \frac{1}{q_n}, \qquad n > N, \tag{2.4}$$

where $r_n = z_n z_{n+1}/c_n^2 > 0$. Thus r_n is a positive solution of

$$r_n + \frac{1}{r_{n-1}} = \frac{1}{q_n} - 1 - \frac{r_n}{r_{n-1}} = B_n - \varepsilon_n \quad \text{for } n > N, \tag{2.5}$$

where $\varepsilon_n = r_n/r_{n-1} > 0$. By applying Lemma 2.1, with $\tilde{B}_n = B_n - \varepsilon_n$, we see that there exists a sequence $v_n \geqq r_n > 0$, $n > N$, satisfying

$$v_n + \frac{1}{v_{n-1}} = B_n.$$

This equation is of the form (1.3) with $c_n \equiv 1$ and with b_n replaced by B_n. Since v_n is a positive solution, we may apply Theorem 1.1 to conclude that (2.1) is nonoscillatory, which completes the proof.

We now apply some known oscillation criteria to (2.1) to obtain new criteria for (1.1).

For example, by [5, Cor. 3.3], if

$$\liminf_{n\to\infty} \sum_{k=1}^{n} (B_k - 2) = -\infty \tag{2.6}$$

then (2.1) is oscillatory. Since $B_k = q_k^{-1} - 1$, we have the following oscillation result for (1.1).

THEOREM 2.2. *Equation* (1.1) *is oscillatory if for some N,*

$$\liminf_{n\to\infty} \sum_{k=N}^{n} \left(q_k^{-1} - 3\right) = -\infty. \tag{2.7}$$

For another result, we apply Theorem 1.4 to (2.1), which tells us that (2.1) is oscillatory if $Q_{n_k} \geqq 1$ for a sequence $n_k \to \infty$, where

$$Q_n = \frac{C_n^2}{B_n B_{n+1}} = \frac{1}{B_n B_{n+1}} = \left(q_n^{-1} - 1\right)^{-1}\left(q_{n+1}^{-1} - 1\right)^{-1}. \tag{2.8}$$

Thus, by Theorem 2.1, (1.1) is oscillatory if $q_n < 1$ and

$$\left(q_{n_k}^{-1} - 1\right)\left(q_{n_k+1}^{-1} - 1\right) \leqq 1, \tag{2.9}$$

for some sequence $n_k \to \infty$. But since $0 < q_n < 1$ for all $n \geqq N$, some simple algebra shows that condition (2.9) is equivalent to

$$q_{n_k} + q_{n_k+1} \geqq 1. \tag{2.10}$$

Thus we have the following refinement of Theorem 1.4.

THEOREM 2.3. *If $q_{n_k} + q_{n_k+1} \geqq 1$ for some sequence $n_k \to \infty$, then* (1.1) *is oscillatory.*

Similarly, Theorem 1.2 leads to the following result.

THEOREM 2.4. *If $q_n < 1$ and $q_n + q_{n+1} + (3-\varepsilon) q_n q_{n+1} \geqq 1$ for all $n \geqq N$, for some $\varepsilon > 0$, then* (1.1) *is oscillatory.*

Any other known oscillation criteria could be applied to (2.1) in this same way to obtain further results like Theorems 2.2–2.4. We leave the details to the interested reader.

3. Extensions of §2. In this section, we will assume that $q_n + q_{n+1} < 1$ for all $n \geqq N$, for some $N > 0$. If this were not the case, (1.1) would be oscillatory by Theorem 2.3.

The method of §2 can be extended in two ways. First, since (2.1) is of the form (1.1) with $C_n \equiv 1$, we can apply Theorem 2.1 to equation (2.1) to obtain the following result.

THEOREM 3.1. *Given* (2.1), *let* $\bar{B}_n = Q_n^{-1} - 1$, $n \geqq N$, *where* $Q_n = (B_n B_{n+1})^{-1}$. *If the equation*

$$(3.1) \qquad u_{n+1} + u_{n-1} = \bar{B}_n u_n$$

is oscillatory, then (2.1) *is oscillatory also.*

(*Note.* Since $\bar{B}_n = Q_n^{-1} - 1$ and $B_n = q_n^{-1} - 1$, our assumption $q_n + q_{n+1} < 1$ implies that $B_n B_{n+1} > 1$, which in turn implies that $\bar{B}_n > 0$.)

As in §2, we can now apply known oscillation criteria to equation (3.1) to obtain new criteria for (2.1) and hence for (1.1). For example, by Theorem 1.4, (3.1) is oscillatory if $\bar{B}_n > 0$ for $n \geqq N$ and

$$(3.2) \qquad \bar{Q}_{n_k} = \left(\bar{B}_{n_k} \bar{B}_{n_k+1}\right)^{-1} \geqq 1 \quad \text{for some sequence } n_k \to \infty.$$

Since $\bar{B}_n = Q_n^{-1} - 1 = B_n B_{n+1} - 1$ and $B_n = q_n^{-1} - 1$, we have

$$(3.3) \qquad \begin{aligned} \bar{Q}_n &= (\bar{B}_n \bar{B}_{n+1})^{-1} = (B_n B_{n+1} - 1)^{-1} (B_{n+1} B_{n+2} - 1)^{-1} \\ &= \left[(q_n^{-1} - 1)(q_{n+1}^{-1} - 1) - 1\right]^{-1} \left[(q_{n+1}^{-1} - 1)(q_{n+2}^{-1} - 1) - 1\right]^{-1}. \end{aligned}$$

Hence, condition (3.2) becomes

$$(3.4) \qquad \left(\frac{1 - q_{n_k} - q_{n_k+1}}{q_{n_k} q_{n_k+1}}\right) \left(\frac{1 - q_{n_k+1} - q_{n_k+2}}{q_{n_k+1} q_{n_k+2}}\right) \leqq 1.$$

Thus, by applying Theorems 2.1 and 3.1, the following criterion is readily verified.

THEOREM 3.2. *If* $q_n + q_{n+1} < 1$ *for* $n \geqq N$ *and* (3.4) *holds for a sequence* $n_k \to \infty$, *then* (1.1) *is oscillatory.*

In general, since $\bar{C}_n \equiv 1$ in (3.1), any oscillation criterion in terms of the coefficients of (3.1) becomes an oscillation criterion for (1.1) given in terms of the expression

$$(3.5) \qquad \bar{B}_n = B_n B_{n+1} - 1 = \frac{1 - q_n - q_{n+1}}{q_n q_{n+1}}.$$

A second way of extending the results of §2 is again to assume that (1.1) is nonoscillatory and proceed as in the proof of Theorem 2.1 to obtain (2.4). However, (2.4) may be written in the form

$$(3.6) \qquad (1 + r_n)\left[1 + \frac{1}{r_{n-1}}\right] = \frac{1}{q_n}, \qquad n > N.$$

From (3.6) we obtain

$$(3.7) \qquad (1 + r_n)\left[1 + \frac{1}{r_{n-1}}\right](1 + r_{n+1})\left[1 + \frac{1}{r_n}\right] = \frac{1}{q_n q_{n+1}}.$$

By the inequality $(1 + a)(1 + a^{-1}) \geqq 4$ for $a > 0$, (3.7) implies

$$(3.8) \qquad 4(1 + r_{n+1})\left[1 + \frac{1}{r_{n-1}}\right] \leqq \frac{1}{q_n q_{n+1}}, \qquad n > N.$$

Then

$$r_{n+1}+\frac{1}{r_{n-1}}\leqq\frac{1}{4q_nq_{n+1}}-1-\frac{r_{n+1}}{r_{n-1}}. \tag{3.9}$$

Thus, r_n is a positive solution of an equation of the form

$$r_{n+1}+\frac{1}{r_{n-1}}=\frac{1}{4q_nq_{n+1}}-1-\delta_n, \qquad n>N, \tag{3.10}$$

where $\delta_n\geqq r_{n+1}/r_{n-1}>0$. We write this second-order nonlinear equation as

$$r_{n+1}+\frac{1}{r_{n-1}}=E_n-\delta_n, \qquad n>N, \tag{3.11}$$

where $E_n=(4q_nq_{n+1})^{-1}-1$, and consider the related first-order equation

$$u_n+\frac{1}{u_{n-1}}=E_{2n}-\delta_{2n}. \tag{3.12}$$

Then the sequence $u_n=r_{2n+1}$ is a positive solution of (3.12) for $n>(N-1)/2$. By Lemma 2.1, the equation

$$v_n+\frac{1}{v_{n-1}}=E_{2n} \tag{3.13}$$

also has a positive solution. We then apply Theorem 1.1 to conclude that

$$y_{n+1}+y_{n-1}=E_{2n}y_n \tag{3.14}$$

is nonoscillatory. Thus if (3.14) is oscillatory, (1.1) must be oscillatory also.

Similarly, $u_n=r_{2n}$ is a positive solution of

$$u_n+\frac{1}{u_{n-1}}=E_{2n-1}-\delta_{2n-1}. \tag{3.15}$$

By Lemma 2.1, the equation

$$v_n+\frac{1}{v_{n-1}}=E_{2n-1}$$

must also have a positive solution v_n. Again, an application of Theorem 2.1 implies that the following equation is nonoscillatory.

$$z_{n+1}+z_{n-1}=E_{2n-1}z_n. \tag{3.16}$$

Thus if (3.16) is oscillatory, so is (1.1).

As in the first part of this section, one can now apply various known oscillation criteria to (3.14) or (3.16) to obtain new criteria for (1.1).

As an example, by [5, Cor. 3.3], if

$$\liminf_{n\to\infty}\sum_{k=N}^{n}(E_{2k}-2)=-\infty, \tag{3.17}$$

then (3.14) is oscillatory. Similarly, if

$$\liminf_{n\to\infty}\sum_{k=N}^{n}(E_{2k-1}-2)=-\infty, \tag{3.18}$$

then (3.16) is oscillatory. However, if (3.14) or (3.16) is oscillatory, so is (1.1). At least one of (3.17) or (3.18) will be true if

$$\liminf_{n\to\infty} \sum_{k=N}^{n} (E_k - 2) = -\infty. \tag{3.19}$$

Thus we have the following theorem.

THEOREM 3.3. *If condition* (3.19) *holds then* (1.1) *is oscillatory, where* $E_k = (4q_kq_{k+1})^{-1} - 1$.

From (1.8), $(4q_kq_{k+1})^{-1} = \alpha_{k,1}$, which means Theorem 3.3 can be restated as follows.

THEOREM 3.4. *If*

$$\liminf_{n\to\infty} \sum_{k=N}^{n} (\alpha_{k,1} - 3) = -\infty,$$

then (1.1) *is oscillatory.*

Using the $\alpha_{n,m}$ notation, observe that Theorem 3.4 is the same as Theorem 2.2, except that Theorem 2.2 has $\alpha_{k,0}$ instead of $\alpha_{k,1}$. We conjecture that Theorem 3.4 is true for $\alpha_{k,m}$ for any $m \geqq 0$.

A more ambitious question is as follows. In essence, Theorem 2.1 says that any known result on the oscillation of (2.1) involving B_n can be rephrased with B_n replaced by $(1/q_n - 1) = (\alpha_{n,0} - 1)$ and then applied to (1.1) to yield sufficient conditions for the oscillation of that equation. Our conjecture is that B_n can be replaced by $(\alpha_{n,m} - 1)$, for any $m \geqq 0$.

REFERENCES

[1] F. V. ATKINSON, *Discrete and Continuous Boundary Problems*, Academic Press, New York, 1964.

[2] T. FORT, *Finite Differences and Difference Equations in the Real Domain*, Oxford Univ. Press, London, 1948.

[3] D. B. HINTON AND R. T. LEWIS, *Spectral analysis of second order difference equations*, J. Math. Anal. Appl., 63 (1978), pp. 421–438.

[4] J. W. HOOKER AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451–462.

[5] J. W. HOOKER, M. K. KWONG, AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations* II, J. Math. Anal. Appl., 107 (1985), pp. 182–196.

[6] W. T. PATULA, *Growth and oscillation properties of second order linear difference equations*, this Journal, 10 (1979), pp. 55–61.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

005

ON OSCILLATIONS OF SOME RETARDED DIFFERENTIAL EQUATIONS*

O. ARINO[†], G. LADAS[‡] AND Y. G. SFICAS[§]

Abstract. Consider the delay differential equation

$$y'(t)+py(t-\tau)-qy(t-\sigma)=0 \tag{$*$}$$

where p, q, τ, and σ are positive constants.

THEOREM. *Assume that $\sigma \leqq \tau$, $q<p$, and $q(t-\sigma) \leqq 1$. Then every nonoscillatory solution of $(*)$ tends to zero as $t \to \infty$. Furthermore, assume that $(p-q)\tau e>1$. Then every solution of $(*)$ oscillates.*

The above result was extended to equations with several delays.

Finally we obtained sufficient conditions for the oscillation of delay differential equations with oscillating coefficients.

Key words. oscillation, retarded differential equation, delay differential equation

AMS(MOS) subject classifications. Primary 34K15; secondary 34C10

1. Introduction. Recently, there has been a lot of interest in establishing computable sufficient conditions for the oscillation of all solutions of linear delay differential equations. See, for example [2], [4], [8], [9] and the references cited in [9].

For the most part the literature is devoted to equations of the form

$$x'(t)=-\sum_{i=1}^{n} p_i(t)x(t-\tau_i(t))$$

where all the coefficients p_i are positive. The case where both positive and negative coefficients may be present was recently considered by Ladas and Sficas [5].

The aim of the present paper is to provide a significant extension of the results in [5] and to develop some methods for studying equations with positive and negative coefficients. In particular, we combine the methods used separately in [2] and [5] and utilize more cleverly the integral form introduced in [5, formula 10].

2. Differential inequalities. In this section we study the properties of solutions of certain differential inequalities and obtain some useful results which can be used as tools in the study of oscillation theory.

LEMMA 1. *Let z be a nonnegative solution*[1] *of the delay differential inequality*

$$z'(t)+az(t-\tau) \leqq 0, \qquad t \geqq t_0 \tag{1}$$

where a and τ are positive constants. Then the following statements hold:

(i) $z \in L^1(t_0, \infty)$;

(ii) *$z(t)$ decreases to zero as $t \to \infty$*;

* Received by the editors January 2, 1985; accepted for publication October 1, 1985.

[†] Département de Mathématiques et d'Informatique, Université de Pau, Avenue Louis-Sallenave, 64000 Pau, France.

[‡] Department of Mathematics, University of Rhode Island, Kingston, Rhode Island 02881-0816.

[§] Department of Mathematics, University of Ioannina, Ioannina, Greece.

[1] That is, $z(t) \geqq 0$ and continuous for $t_0-\tau \leqq t \leqq t_0$ and satisfies (1) for $t \geqq t_0$.

(iii) *if* $z(t_0)>0$ *then there exist positive constants* α *and* A *such that*

$$z(t)\geqq Ae^{-\alpha t}, \qquad t\geqq t_0. \tag{2}$$

Proof. As $z'(t)\leqq -az(t-\tau)\leqq 0$ for $t\geqq t_0$, it follows that z is a decreasing function. Integrating (1) from t_1 to t_2 with $t_0\leqq t_1\leqq t_2$, we find that

$$a\int_{t_1}^{t_2} z(s-\tau)\,ds\leqq z(t_1)-z(t_2)\leqq z(t_0) \tag{3}$$

which implies that $z\in L^1(t_0,\infty)$. And since $z(t)$ is also decreasing, it follows that $\lim_{t\to\infty} z(t)$ exists and is equal to zero. Next, we turn to the proof of (iii). Replacing t_1 by t and letting $t_2\to\infty$ in (3), we find

$$z(t)\geqq a\int_t^\infty z(s-\tau)\,ds. \tag{4}$$

Using the decreasing nature of $z(t)$ we obtain,

$$z(t)\geqq a\int_t^\infty z(s-\tau)\,ds\geqq a\int_t^{t+\tau/2} z(s-\tau)\,ds\geqq \frac{a\tau}{2}z\left(t-\frac{\tau}{2}\right)$$

that is,

$$z(t)\geqq Cz\left(t-\frac{\tau}{2}\right), \tag{5}$$

where $C=a\tau/2>0$. Then $z(t_0+\tau/2)\geqq Cz(t_0)>0$. Set

$$I_n=\left[t_0+n\frac{\tau}{2},t_0+(n+1)\frac{\tau}{2}\right], \qquad n=0,1,2,\cdots.$$

Let $t\geqq t_0$ be given. Then $t\in I_n$ for some n. Using the estimate (5) we find by iteration

$$z(t)\geqq C^{n+1}z(t_0).$$

Assume $C>1$. Thus, using the fact that $2(t-t_0)/\tau\leqq n+1$, we have

$$z(t)\geqq e^{(n+1)\ln C}z(t_0)\geqq \exp\left[\frac{2(t-t_0)}{\tau}\ln C\right]z(t_0).$$

On the other hand, if $C\leqq 1$, using the fact that $n\leqq 2(t-t_0)/\tau$, we obtain

$$z(t)\geqq \exp\left[\frac{2(t-t_0)}{\tau}\ln C\right]Cz(t_0).$$

In either case, inequality (2) is valid and the proof of Lemma 1 is complete.

The proof of the following result makes use of the notion of nonautonomous exponents [1], [2].

PROPOSITION 1. *Let* z *be a nonnegative solution of the delay differential inequality* (1). *Assume that*

$$a\tau>\frac{1}{e}. \tag{6}$$

Then $z(t)=0$ *for* $t\geqq t_0$.

Proof. If $z(t_0)=0$ then, as z is nonnegative and decreasing, it follows that $z(t)=0$ for $t\geqq t_0$. Next, assume, for the sake of contradiction, that $z(t_0)>0$. Then, from Lemma 1 (iii), $z(t)>0$ for $t\geqq t_0$. Set

$$z(t)=\exp\left[-\int_0^t \lambda(s)\,ds\right], \qquad t\geqq t_0. \tag{7}$$

Substituting (7) into (1) we find that

$$\lambda(t) \geqq a \exp\left[\int_{t-\tau}^{t} \lambda(s)\,ds\right], \qquad t \geqq t_0. \tag{8}$$

Also from (7), and in view of (2), there is a positive constant β such that

$$\int_0^t \lambda(s)\,ds \leqq \beta t, \qquad t \geqq t_0. \tag{9}$$

In view of (8), it follows by induction that for $n = 1, 2, \cdots$

$$\lambda(t) \geqq b_n, \qquad t \geqq t_0 + (n-1)\tau \tag{10}$$

where $b_0 = 0$ and $b_n = a\exp(b_{n-1}\tau)$. We now claim that the sequence $\{b_n\}$ is increasing and $\lim_{n\to\infty} b_n = +\infty$. In fact, using (6) and the fact that $\exp(x) \geqq ex$ for all x we find

$$b_n = a\exp(bb_{n-1}\tau) \geqq aeb_{n-1}\tau \geqq b_{n-1}$$

that is, $\{b_n\}$ is increasing. If, contrary to the claim, $\{b_n\}$ were bounded, then $b_\infty \equiv \lim_{n\to\infty} b_n > 0$ and

$$b_\infty = a\exp(b_\infty \tau) \geqq aeb_\infty \tau > b_\infty.$$

Thus our claim about $\{b_n\}$ is true and consequently, from (10), $\lim_{t\to\infty}\lambda(t) = \infty$. This contradicts (9) and the proof of the proposition is complete.

Remark 1. The main tools behind the proof of Proposition 1 are the integral form (4) of inequality (1), the exponential representation (7) of z, and the integral estimates (8) and (9). As we will see in the next result, the same tools can be used in the case of delay differential inequalities with several delays. Note also that in the following result the condition (6) of Proposition 1 has been replaced by the general assumption that all solutions of the corresponding equation oscillate. As it is known, in the case of equations with one delay the two statements are equivalent.

PROPOSITION 2. *Let z be a nonnegative solution of the delay differential inequality*

$$z'(t) + \sum_{j=1}^{n} a_j z(t-\tau_j) \leqq 0, \qquad t \geqq t_0 \tag{11}$$

where a_j and τ_j are positive constants for $j = 1, 2, \cdots, n$. Assume that all solutions of the DDE

$$z'(t) + \sum_{j=1}^{n} a_j z(t-\tau_j) = 0, \qquad t \geqq t_0 \tag{12}$$

oscillate. Then, $z(t) = 0$ for $t \geqq t_0$.

Proof. Inequality (11) implies that for each $j = 1, 2, \cdots, n$,

$$z'(t) + a_j z(t-\tau_j) \leqq 0, \qquad t \geqq t_0$$

and therefore the results (i), (ii), and (iii) of Lemma 1 are also true here. In particular, if $z(t_0) > 0$ then $z(t) > 0$ for $t \geqq t_0$. As in Proposition 1, the change of variables (7) yields

$$\lambda(t) \geqq \sum_{j=1}^{n} a_j \exp\left[\int_{t-\tau_j}^{t} \lambda(s)\,ds\right] \tag{8$'$}$$

which inductively leads to a sequence of estimates,

$$\lambda(t) \geqq b_k, \qquad t \geqq t_0+(n-1)\tau$$

where $b_0=0$, $b_k=\sum_{j=1}^{n} a_j \exp(\tau_j b_{k-1})$ for $k=1,2,\cdots,n$, and $\tau=\max_{1 \leqq j \leqq n}\tau_j$. We now claim that the sequence $\{b_n\}$ is increasing. Indeed, the characteristic equation

$$F(\lambda) \equiv \lambda + \sum_{j=1}^{n} a_j e^{-\lambda\tau_j}=0$$

of (12) has no real roots. As $F(0)=\sum_{j=1}^{\infty} a_j > 0$, it follows that $F(\lambda)>0$ for all $\lambda \in R$. In particular

$$F(-b_{k-1}) = -b_{k-1} + \sum_{j=1}^{n} a_j e^{b_{k-1}\tau_j} > 0.$$

Hence,

$$b_k = \sum_{j=1}^{n} a_j e^{b_{k-1}\tau_j} > b_{k-1}$$

which proves our claim. Next, we will show that

$$\lim_{k \to \infty} b_k = \infty. \tag{13}$$

Otherwise, $b_\infty \equiv \lim_{k \to \infty} b_k$ is a finite number and so

$$b_\infty = \sum_{j=1}^{n} a_j \exp(\tau_j b_\infty)$$

which implies that (12) has the nonoscillatory solution $z(t)=\exp(-b_\infty t)$. This contradiction establishes (13). Finally, as in the proof of Proposition 1, (13) implies that $\lim_{t \to \infty} \lambda(t)=\infty$ which contradicts (9). The proof of Proposition 2 is complete.

3. Positive and negative coefficients—autonomous equations. In [5], Ladas and Sficas studied the asymptotic behavior of the nonoscillatory solutions and the oscillation of all solutions of the DDE

$$x'(t)+px(t-\tau)-qx(t-\sigma)=0 \tag{14}$$

where the delays τ and σ and the coefficients p and q satisfy the hypothesis that

(H_1) p, q, τ, and σ are positive constants.

The aim in this section is to sharpen the conditions which were assumed in [5] and to extend the results to equations with more than two delays. It was shown in [5] that the hypothesis

(H_2) $q<p$ and $\sigma \leqq \tau$, is a necessary condition for all solutions of (14) to oscillate.

The following statement improves [5, Thm. 2]:

THEOREM 1. *In addition to the hypotheses* (H_1) *and* (H_2) *assume that*

(H_3) $q(\tau-\sigma) \leqq 1$.

Then every nonoscillatory solution of (14) *tends to zero as* $t \to \infty$.

Proof. It suffices to prove the theorem for the eventually positive solutions of (14). To this end, let $x(t)$ be a solution of (14) which is positive for $t \geqq t_0$. As in [5] we introduce the function

$$z(t)=x(t)-q\int_{t-\tau}^{t-\sigma} x(s)\,ds, \qquad t \geqq t_0+\tau. \tag{15}$$

Then

$$(16)\qquad z'(t)=-(p-q)x(t-\tau)<0,\quad t\geqq t_0+\tau$$

and so $z(t)$ is decreasing. We claim that $z(t)$ is bounded below. Otherwise, $\lim_{t\to\infty}z(t)=-\infty$ which implies that x, itself, is unbounded. But then, there must exist a point $t_1\geqq t_0+\tau$ such that $z(t_1)<0$ and $x(t_1)=\max_{s\leqq t_1}x(s)>0$. From (15) and ($H_3$) we obtain the contradiction that

$$0>z(t_1)=x(t_1)-q\int_{t_1-\tau}^{t_1-\sigma}x(s)\,ds\geqq x(t_1)-qx(t_1)(\tau-\sigma)$$
$$=x(t_1)[1-q(\tau-\sigma)]\geqq 0.$$

Thus $z(t)$ is bounded below and $\lim_{t\to\infty}z(t)\equiv l$ exists and is finite. Now integrating (16) from $t_1=t_0+2\tau$ to t and letting $t\to\infty$ we obtain

$$z(t_1)-l=-(p-q)\int_{t_1}^{\infty}x(s-\tau)\,ds$$

which proves that $x\in L^1(t_1,\infty)$. From (14), it follows that $x'\in L^1(t_1,\infty)$. Hence $\lim_{t\to\infty}x(t)$ exists and it has to be zero because $x\in L^1(t_1,\infty)$. The proof is complete.

Remark 2. (a) In place of (H_3) it was assumed in [5] that the more restrictive condition

$$\tau\leqq\frac{1}{q}-\frac{1}{p}$$

holds.

(b) The hypothesis (H_3) is the "best" one when the coefficients p and q are equal. Indeed, when $p=q$ all constants are solutions and conversely all solutions of (14) are asymptotically constant. In the latter case $\lim_{t\to\infty}x(t)$ is given in terms of the initial data, using the "first integral" $x(t)+p\int_{t-\sigma}^{t-\tau}x(s)\,ds\equiv C$, namely,

$$\lim_{t\to\infty}x(t)=\frac{C}{1-p(\tau-\sigma)}.$$

The following is an improved version of [5, Thm. 3]. Its proof makes use of Theorem 1; otherwise it is identical to the proof given in [5] and will be omitted.

THEOREM 2. *Assume* (H_1), (H_2), (H_3) *and* (H_4) $(p-q)\tau e>1$. *Then every solution of* (14) *oscillates.*

Example 1. The DDE

$$x'(t)+e^4x(t-4)-2e^2x(t-2)=0$$

satisfies the hypotheses (H_1), (H_2) and (H_4), but not (H_3). Therefore it is not surprising that $x(t)=e^t$ is a nonoscillatory solution which does not tend to zero as $t\to\infty$.

Next, we extend Theorems 1 and 2 to equations with several delays.

THEOREM 3. *Consider the delay differential equation*

$$(17)\qquad x'(t)+\sum_{i=1}^{n}p_ix(t-\sigma_i)-\sum_{j=1}^{m}q_jx(t-\tau_j)=0$$

where the coefficients p_i, q_j *and the delays* σ_i, τ_j *are positive constants for* $i=1,2,\cdots,n$ *and* $j=1,2,\cdots,m$. *Assume that the following hypotheses are satisfied.*

(H_5) *There exist a positive number $p \leqq n$ and a partition of $\{1,2,\cdots,m\}$ into p disjoint subsets $J_1, J_2,\cdots, J_p$ such that $j \in J_i$ implies $\tau_j \leqq \sigma_i$ and $\sum_{k\in J_i} q_k < p_i$.*

(H_6) $\sum_{i=1}^{p}\sum_{k\in J_i} q_k(\sigma_i - \tau_k) < 1$.

Then every nonoscillatory solution of (17) *tends to zero as $t \to \infty$. Furthermore, if in addition to the above hypotheses we assume that the equation*

$$z'(t)+\sum_{i=1}^{p}\Big(p_i - \sum_{k\in J_i} q_k\Big) z(t-\sigma_i)=0 \tag{18}$$

has only oscillatory solutions then the same is true with (17).

Proof. As the negative of a solution of (17) is also a solution, we first prove that every eventually positive solution $x(t)$ of (17) tends to zero as $t\to\infty$. The key idea is, once more, the introduction of the function

$$z(t)=x(t)-\sum_{i=1}^{p}\sum_{k\in J_i} q_k \int_{t-\sigma_i}^{t-\tau_k} x(s)\,ds. \tag{19}$$

We have

$$z'(t)=\sum_{i=1}^{p}\Big(\sum_{k\in J_i} q_k - p_i\Big)x(t-\sigma_i) - \sum_{i=p+1}^{n} p_i x(t-\sigma_i) \tag{20}$$

and so eventually, $z'(t)<0$ and $z(t)$ is decreasing. Now, as in the proof of Theorem 1, we can deduce from (H_6) that z is bounded below and integrating (20) we conclude that $x \in L^1(t_0,\infty)$. Then from (17), it follows that $x' \in L^1(t_0,\infty)$, and therefore x tends to a limit as $t\to\infty$. And this limit is necessarily zero because $x\in L^1(t_0,\infty)$.

Next, in addition to (H_5) and (H_6), we assume that every solution of (18) is oscillatory. We should prove that every solution of (17) oscillates. Otherwise, (17) has an eventually positive solution $x(t)$. And, as we have already proved, $\lim_{t\to\infty} x(t)=0$. From (19) and (20), we see that $z(t)$ is decreasing to zero which implies that $z(t)$ is eventually positive. In view of (19), we have, $0<z(t)\leqq x(t)$. Thus, (20) yields

$$z'(t)+\sum_{i=1}^{p}\Big(p_i - \sum_{k\in J_i} q_k\Big) z(t-\sigma_i)\leqq 0. \tag{21}$$

On the basis of our assumptions, Proposition 2 applied to (21) implies that eventually $z(t)=0$. This contradiction completes the proof of Theorem 3.

4. Nonautonomous equations. Consider the DDE

$$x'(t)+p(t)x(t-\tau)=0 \tag{22}$$

where $p(t)$ is continuous on $[0,\infty)$ and $\tau>0$ is a constant. Set

$$p(t)=p^+(t)-p^-(t)$$

where p^+ and p^- are the positive and negative parts of p respectively.

Lemma 2. *Assume that*

$$p^- \in L^1[0,\infty). \tag{H_7}$$

Then every nonoscillatory solution of (22) *tends to a (finite) limit as $t\to\infty$.*

Proof. Choose a $t_0 > t$ such that

$$\int_{t_0}^{\infty} p^-(s)\,ds = \alpha < 1.$$

Now, we will prove that if x is an eventually positive solution of (22), then x is bounded above. Assuming the contrary, we could find a sequence $\{t_n\}$ such that $t_n \geqq t_0$

$$\lim_{n\to\infty} t_n = +\infty, \quad \lim_{n\to\infty} x(t_n) = +\infty, \quad x(t_n) = \max_{t \leqq t_n} x(t).$$

Let $t_1 \geqq t_0$ be so large that $x(t-\tau)>0$ for $t \geqq t_1$. Then from (22) we have

$$x'(t) \leqq p^-(t)x(t-\tau), \qquad t \geqq t_1,$$

and by integration

$$x(t)-x(t_1) \leqq \alpha \max_{s \leqq t} x(s).$$

If we replace t by t_n we find

$$x(t_n) \leqq \frac{x(t_1)}{1-\alpha}$$

which contradicts the asumption that $\lim_{n\to\infty} x(t_n) = +\infty$. Thus x is bounded and so in view of (H_7), $p^-(t)x(t-\tau) \in L^1[t_1,\infty)$. Integrating (22) from t_1 to t_2, with $t_1<t_2$, we get

$$\int_{t_1}^{t_2} p^+(s)x(s-\tau) \leqq \int_{t_0}^{\infty} p^-(s)x(s-\tau) + 2 \sup_{s \geqq t_0} x(s).$$

Therefore, $p^+(t)x(t-\tau) \in L^1[t_1,\infty)$ and so $x' \in L^1[t_1,\infty)$ which implies that x tends to a finite limit as $t \to \infty$.

Example 2. The DDE

$$y'(t) - \frac{\cos t}{2-\cos t} y\left(t-\frac{\pi}{2}\right) = 0$$

has the nonoscillatory solution $y(t) = 2+\sin t$ which does not have a limit as $t \to \infty$. As expected, the hypothesis (H_7) is not satisfied in this example.

LEMMA 3. *Assume* (H_7) *and that*

$$(\mathrm{H}_8) \qquad \int_0^{\infty} p^+(s)\,ds = \infty.$$

Then every solution of (22) *either oscillates or tends to zero as* $t \to \infty$.

Proof. Assume, for the sake of contradiction, that (22) has a positive solution $x(t)$ which does not tend to zero as $t \to \infty$. In view of Lemma 2, $\lim_{t\to\infty} x(t) = x(\infty)$ exists and since it is not zero it follows that

$$(23) \qquad x(\infty) > 0.$$

Then, for t sufficiently large, say $t \geqq t_1$, we have

$$\frac{1}{2}x(\infty) \leqq x(t) \leqq \frac{3}{2}x(\infty).$$

In view of (22) this leads to

$$x'(t) + \frac{1}{2}x(\infty)p^+(t) - \frac{3}{2}x(\infty)p^-(t) \leqq 0$$

for $t \geqq t_1+\tau$. Integrating from $t_1+\tau$ to t we obtain

$$(24) \qquad x(t)-x(t_1+\tau)+\frac{1}{2}x(\infty)\int_{t_1+\tau}^{t} p^+(s)\,ds - \frac{3}{2}x(\infty)\int_{t_1+\tau}^{t} p^-(s)\,ds \leqq 0.$$

Taking limits in (24) as $t\to\infty$ and using the hypotheses (H_7) and (H_8), we find that $y(\infty)=-\infty$ which contradicts (23) and completes the proof.

The following result gives sufficient conditions for all solutions of (22) to oscillate. A similar result was established in [6] under the (strong) assumption that $\lim_{n\to\infty}(v_n-\theta_n)=\infty$.

THEOREM 4. *Assume that the following conditions hold*:

(i) (H_7) *and*

$$(\mathrm{H}_9)\qquad \varliminf_{t\to\infty}\int_{t-\tau}^{t}p^+(s)\,ds>\frac{1}{e}.$$

(ii) *There exist two sequences* $\{\theta_n\}_{n=1}^{\infty}$ *and* $\{v_n\}_{n=1}^{\infty}$ *such that for* $n=1,2,\cdots$,

$$\theta_n<v_n\leqq\theta_{n+1},\quad v_n-\theta_n\geqq\frac{3\tau}{2},\quad \theta_{n+1}-v_n<\frac{\tau}{2}$$

and

$$p(t)\ is\begin{cases}\geqq 0 & for\ t\in\bigcup_{n=1}^{\infty}(\theta_n,v_n),\\ <0 & for\ t\in\bigcup_{n=1}^{\infty}(v_n,\theta_{n+1}).\end{cases}$$

(iii) *There exists a positive constant k such that*

$$\varliminf_{t\to\infty}\int_{t-\tau/2}^{t}p^+(s)\,ds>\frac{1}{k}\quad for\ t\in\bigcup_{n=1}^{\infty}(v_n,\theta_{n+1})$$

and

$$(25)\qquad p^+(t)\geqq kp^-\left(t-\frac{\tau}{2}\right)\quad for\ t\ sufficiently\ large.$$

Then every solution of (22) *oscillates.*

Proof. Otherwise, (22) has a solution $x(t)>0$ for $t\geqq t_0$ where t_0 is sufficiently large. And by Lemma 3 [which holds because condition (H_9) implies (H_8)],

$$(26)\qquad \lim_{t\to\infty}x(t)=0.$$

Let $t\in\bigcup_{n=1}^{\infty}[v_n,\theta_{n+1}]$. Thus from (22) and the fact that $x(t)$ decrease in the interval $[t-3\tau/2,t-\tau/2]$ we find

$$(27)\qquad x\left(t-\frac{\tau}{2}\right)>x\left(t-\frac{\tau}{2}\right)-x(v_n)=\int_{t-\tau/2}^{v_n}p(x)x(s-\tau)\,ds$$
$$=\int_{t-\tau/2}^{t}p^+(s)x(s-\tau)\,ds\geqq x(t-\tau)\int_{t-\tau/2}^{t}p^+(s)\,ds.$$

Now for t sufficiently large, say $t\geqq t_1$ and for $t\in\bigcup_{n=1}^{\infty}[v_n,\theta_{n+1}]$ we find, from (27) and (iii),

$$(28)\qquad x\left(t-\frac{\tau}{2}\right)>\frac{1}{k}x(t-\tau).$$

And for all $t\geqq t_1$ the following inequality holds:

(29)

$$x'(t)+\left[p^+(t)-kp^-\left(t-\frac{\tau}{2}\right)\right]x(t-\tau)+kp^-\left(t-\frac{\tau}{2}\right)x(t-\tau)-kp^-(t)x\left(t-\frac{\tau}{2}\right)\geqq 0.$$

In fact (29) reduces to (22) for $t \in \bigcup_{n=1}^{\infty}(\theta_n, v_n)$ and for $t \in \bigcup_{n=1}^{\infty}(v_n, \theta_{n+1})$ it follows from (22) using (28) and (H_9). Set

$$z(t) = x(t) - k\int_{t-\tau/2}^{t} p^-(s)\,x\left(s-\frac{\tau}{2}\right)ds.$$

Thus from (29) we have for $t \geqq t_1$

$$z(t) \leqq x(t) \tag{30}$$

and

$$z'(t) + \left[p^+(t) - kp^-\left(t-\frac{\tau}{2}\right)\right]x(t-\tau) \leqq 0. \tag{31}$$

From (30) and because of (26) and (H_7), $\lim_{t\to\infty} z(t) = 0$. Hence $z(t) > 0$ and (30), (31), and (iii) yield the inequality

$$z'(t) + \left[p^+(t) - kp^-\left(t-\frac{\tau}{2}\right)\right]z(t-\tau) \leqq 0. \tag{32}$$

Since

$$\varliminf_{t\to\infty}\int_{t-\tau}^{t}\left[p^+(s) - kp^-\left(s-\frac{\tau}{s}\right)\right]ds = \varliminf_{t\to\infty}\int_{t-\tau}^{t} p^+(s)\,ds > \frac{1}{e},$$

it follows from [7] that (32) cannot have an eventually positive solution. But $z(t)$ is positive and this contradiction completes the proof.

Example 3. This is an example of a DDE which satisfies the hypotheses of Theorem 4 and therefore every solution of the equation oscillates. Consider the DDE

$$x'(t) + p(t)x(t-4\pi) = 0, \qquad t \geqq 0$$

where

$$p(t) = \begin{cases} \mu\sin^2 t & \text{for } t \in [10n\pi, 10n\pi + 9\pi), \\ -\mu\dfrac{\sin^2 t}{t^2} & \text{for } t \in [10n\pi + 9\pi, 10(n+1)\pi), \end{cases} \qquad n = 0, 1, 2, \cdots,$$

and μ is a positive constant satisfying the condition

$$\mu\frac{3\pi}{2} > \frac{1}{e}. \tag{33}$$

In this example, $\tau = 4\pi$, $\theta_n = 10n\pi$, $v_n = 10n\pi + 9\pi$, $v_n - \theta_n = 9\pi > 3\tau/2 = 6\pi$ and $\theta_{n+1} - v_n = \pi < \tau/2 = 2\pi$ and so condition (ii) of Theorem 4 is satisfied. Condition (H_7) is clearly satisfied and condition (H_9) is true because of (33). Indeed, examining the value of the integral in (H_9) for the different possible locations of t and $t-\tau$ in the intervals $(\theta_n, v_n]$ and $(v_n, \theta_{n+1}]$ we find that the integral takes its smallest value when $t = \theta_{n+1}$ and in this case, in view of (33),

$$\int_{t-\tau}^{t} p^+(s)\,ds = \int_{10n\pi+6\pi}^{10n\pi+9\pi} \mu\sin^2 ds = \mu\frac{3\pi}{2} > \frac{1}{e}.$$

Inequality (25) is true as can be seen by examining it for the different locations of t and $t-\tau/2$ in the intervals $(\theta_n, v_n]$ and $(v_n, \theta_{n+1}]$. For example, when $t \in (\theta_n, v_n]$ and $t - \tau/2$ in $(v_n, \theta_{n+1}]$,

$$p^+(t) - kp^-\left(t-\frac{\tau}{2}\right) = \sin^2 t - k\frac{\sin^2(t-2\pi)}{(t-2\pi)^2} = \sin^2 t\left[1 - \frac{k}{(t-2\pi)^2}\right] \geqq 0$$

for t sufficiently large and $k>0$. Finally, the first inequality in (iii) of Theorem 4 is satisfied with k any constant greater than $1/\mu\pi$. Indeed, the smallest value of the integral

$$\int_{t-\tau/2}^{t} p^{+}(s)\,ds$$

is obtained when $t=10(n+1)\pi$ and in this case

$$\int_{t-\tau/2}^{t} p^{+}(s)\,ds=\int_{10n\pi+8P\pi}^{10(n+1)\pi} \mu\sin^2 s\,ds=\mu\pi>\frac{1}{k}.$$

REFERENCES

[1] O. Arino and I. Györi, *Asymptotic integration of functional differential systems which are asymptotically autonomous*, Publ. Intern. de Pau, 1983.

[2] O. Arino, I. Györi and A. Jawhari, *Oscillation criteria in delay equations*, J. Differential Equations, 53 (1984), pp. 115–123.

[3] O. Arino, G. Ladas and Y. G. Sficas, *Oscillations of nonautonomous retarded differential equations*, Publ. Intern. de Pau, 1984.

[4] B. R. Hunt and J. A. Yorke, *When all solutions of* $x'=-\Sigma_{i=1}^{n}q_i(t)x(t-T_i(t))$ *oscillate*, J. Differential Equations, 53 (1984), pp. 139–145.

[5] G. Ladas and Y. G. Sficas, *Oscillations of Delay Differential Equations with positive and negative coefficients*, Proc. of the International Conference on the Qualitative Theory of Differential Equations, Edmonton, Alberta, Canada, June 18–20, 1984.

[6] G. Ladas, Y. G. Sficas and I. P. Stavroulakis, *Functional differential inequalities and equations with oscillating coefficients*, Proc. Vth International Conference on Trends in Theory and Practice of Nonlinear Differential Equations, Marcel Dekker, New York, 1984.

[7] G. Ladas and I. P. Stavroulakis, *On Delay Differential Inequalities of First Order*, Funkcial Ecvac., 25 (1982), pp. 105–113.

[8] ———, *Oscillations caused by several retarded and advanced arguments*, J. Differential Equations, 44 (1982), pp. 134–152.

[9] B. G. Zhang, *A survey of the oscillation of solutions of first order differential equations with deviating argument*, Proc. International Conference on Trends in Nonlinear Analysis, Arlington, TX, June 18–22, 1984.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

006

AN ABSTRACT DELAY-DIFFERENTIAL EQUATION MODELLING SIZE DEPENDENT CELL GROWTH AND DIVISION*

M. GYLLENBERG† AND H. J. A. M. HEIJMANS‡

Abstract. A two-phase model for the growth of a single cell population structured by size is formulated and analysed. The model takes the form of a delay-differential equation in a Banach space. Using positivity arguments, we describe the spectrum of the infinitesimal generator of the semigroup associated with solutions. Under a certain condition on the growth rate of individual cells the semigroup is compact after finite time. This enables us to determine the ultimate behavior of solutions and prove the existence of a stable size distribution.

Key words. structured populations, cell cycle, first order partial differential equation with delay and transformed argument, delay-differential equation in a Banach space, strongly continuous semigroup, generation expansion, positive operator, Riesz operator, spectral theory, stable size distribution

AMS(MOS) subject classifications. Primary 34K30, 92A15, 35R10

1. Introduction. In this paper we study a mathematical model for the dynamics of a population of single cells which can be distinguished from each other according to their size and the particular cell cycle phase they are in. Models for populations of dividing organisms incorporating size or age-size structure have been formulated, among others, by Bell and Anderson (1967) and Sinko and Streifer (1971) and have recently been investigated by Diekmann, Heijmans and Thieme (1984) and Heijmans (to appear a, b). We refer to the book of Metz and Diekmann (to appear) for a general exposition of the dynamics of physiologically structured populations. There exists a vast literature on models for progress through the cell cycle and its various phases, see for instance the book of Eisen (1979). Tyson and Hannsgen (1984) and Hannsgen and Tyson (1984) have studied cell cycle models which also take the cell size distribution into account.

We consider a model in which we assume that the cell cycle consists of two distinct phases. The first phase is of variable length. The cells in this phase cannot divide—they increase in size and, provided they do not die, they will eventually enter the second phase. This second phase, which is assumed to have constant duration, can be considered as an idealization of the mitotic period. At the conclusion of this phase cells split into two equal parts and the newborn daughter cells start the cycle in the first phase. It is assumed that the cells in the first phase are fully characterized by their size. By this we mean that for instance the growth, death and transition rates are functions of size, and of size only. Moreover, we assume that in the second phase the growth and death rates are functions of size, but the fission rate is a (delta-) function of the time elapsed since entering the second phase.

Our model could be considered as a generalization of the one-phase model studied by Diekmann et al. (1984) since if one formally puts the duration of the second phase equal to zero our fundamental equation (2.1) reduces to the corresponding equation in the above-mentioned paper.

* Received by the editors April 22, 1985; accepted for publication October 1, 1985.

† Helsinki University of Technology, SF-02150 Espoo 15, Finland.

‡ Centre for Mathematics and Computer Science, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands.

The model could easily be modified (without essentially affecting the results) to allow for more complicated cell cycles and asymmetric division, see Gyllenberg (to appear).

Diekmann et al. (1984) showed that under reasonable hypotheses the population will ultimately grow or decay exponentially and they gave conditions on the individual growth rate under which the size distribution converges towards a so-called stable size distribution. In this paper we shall prove that the two-phase model exhibits a similar asymptotic behavior, if we adapt the condition on the growth rate.

It turns out to be mathematically convenient to write the model as a delay-differential equation in a Banach space. Our main tools will be the theory of strongly continuous semigroups and spectral theory, in particular of positive operators.

The organization of the paper is as follows. In §2 we write down the balance equation for the size distribution of the population of the first phase. This equation, which is a first-order hyperbolic PDE with time delay, transformed argument and singular coefficients, is then reformulated as an abstract linear delay-differential equation in a Banach space. In §3 we prove well-posedness of the abstract problem and associate a strongly continuous semigroup of bounded linear operators with the solution. In §4 we represent the solution as a generation expansion and give conditions under which the semigroup is compact after finite time. In §5 we study the related eigenvalue problem and characterize the spectrum of the infinitesimal generator of the semigroup. In §6 we use the results of the preceding sections to state and prove the main result on the asymptotic behavior of solutions.

2. The model. The starting point of our investigation is the balance equation for the size distribution of cells in the first phase

$$(2.1)\quad \frac{\partial}{\partial t}n(t,x)+\frac{\partial}{\partial x}(g(x)n(t,x))$$
$$=-\mu(x)n(t,x)-b(x)n(t,x)+\frac{2p(y^{-1}(x))b(y^{-1}(x))}{y'(y^{-1}(x))}n(t-r,y^{-1}(x)).$$

Here t denotes time and x denotes size. The unknown n is the size distribution of cells in the first phase, i.e. the integral $\int_{x_1}^{x_2} n(t,x)\,dx$ represents the number of cells in the first phase with size between x_1 and x_2 at time t. The functions g,μ and b (which are assumed to be known) are the rates at which cells of size x grow, die and transit to the second phase, respectively. $r>0$ is the constant duration of the second phase, $y(x)$ is the size of a newborn cell whose mother entered the second phase (exactly r time units before) with size x and $p(x)$ is the fraction of cells who survive the second phase given that they entered it with size x.

The left-hand side of (2.1) is the derivative along characteristic curves and describes an individual's motion in the time-size continuum due to growth. The first term on the right-hand side describes the loss due to deaths and the second the loss due to transition to the second phase. The last term describes the birth of cells from mother cells completing their second phase: of those cells that entered the second phase r time units ago with size $y^{-1}(x)$ a fraction $p(y^{-1}(x))$ will successfully complete the phase and give rise to two new cells of size x. The factor $1/y'(y^{-1}(x))$ may seem strange. It is due to the fact that cells giving birth to daughters in the size interval $(x,x+dx)$ left the first phase with size in the interval

$$\left(y^{-1}(x),y^{-1}(x)+\frac{dx}{y'(y^{-1}(x))}\right).$$

If one formally puts $r=0$, one should take $y(x)=\frac{1}{2}x$ and the last term in (2.1) reduces to the corresponding term in the model of Diekmann et al. (1984), namely $4b(2x)n(t,2x)$.

We assume that cells cannot enter the second phase before they have reached a minimal size x_0 and that $y(x_0)\leqq x_0$. It follows that cells with size less than $\alpha:=y(x_0)$ cannot exist. This fact is expressed by the boundary condition

$$n(t,\alpha)=0 \tag{2.2}$$

which supplements (2.1).

We assume further that each cell in the first phase must either die or transit to the second phase before it reaches a maximal size (normalized to $x=1$) of the first phase. This requires that the integral $\int_\alpha^x [b(s)/g(s)]\,ds$ diverges as $x\uparrow 1$ and that the source term in (2.1) is interpreted as zero for $x\geqq\beta:=y(1)$ where β is assumed to be smaller than 1. We once and for all make the convention that all functions containing $y^{-1}(x)$ as an argument are given the value zero for $x\geqq\beta$. The possible sizes a cell in the first phase can have thus lie in the interval $(\alpha,1)$, which should be chosen as the domain of x in (2.1). In order to obtain a well-posed problem, an initial function ν should be prescribed on $[-r,0]\times[\alpha,1]$:

$$n(t,x)=\nu(t,x), \qquad -r\leqq t\leqq 0, \quad \alpha\leqq x<1. \tag{2.3}$$

Concerning the growth, death and transition rates and the other given functions, we assume (compare Diekmann et al. (1984)):

(H_y) $y\in C^1[x_0,1]$, $y'>0$, $\alpha:=y(x_0)\leqq x_0$ and $\beta:=y(1)<1$.

(H_p) $p\in C[x_0,1]$, $0<p(x)\leqq 1$, $x\in[x_0,1]$.

(H_g) $g\in C[\alpha,1]$, $g(x)>0$, $x\in[\alpha,1]$.

(H_μ) $\mu\in C[\alpha,1]$, $\mu(x)\geqq 0$, $x\in[\alpha,1]$.

(H_b) $b\in C[\alpha,1]$, $b(x)=0$, $x\in[\alpha,x_0]$, $b(x)>0$, $x\in(x_0,1)$,

$$\lim_{x\uparrow 1}\int_{x_0}^x b(s)\,ds=\infty \quad\text{and}\quad \frac{b(x)}{g(x)}E(x)\leqq M<\infty, \qquad x\in[\alpha,1].$$

In the last condition we have used the notation

$$E(x)=\exp\left(-\int_\alpha^x \frac{b(s)+\mu(s)}{g(s)}\,ds\right). \tag{2.4}$$

$E(x)/E(y)$ is the probability that a cell of size y remains in the first phase at least until it reaches size x and $\int_{x_1}^{x_2}(b(s)/g(s))E(s)\,ds$ is the probability that a cell with size x_0 enters the second phase when its size is between x_1 and x_2. By our assumptions the (possibly defective) probability density bE/g is not only an L^1-function but also bounded and continuous.

We point out that some of the assumptions could be weakened at the cost of some minor technical difficulties. For instance, if $y(x_0)>x_0$ we could redefine x_0 in the following way. Let $x_n=y(x_{n-1})$, $n=1,2,\cdots$. Since $y(1)<1$, $x_n\to x_0'$ as $n\to\infty$ where x_0' is the smallest fixed point of y. x_0' could then be taken as the new x_0. A similar procedure has been carried out in Heijmans (to appear, b). Guided by Diekmann et al. (1984) we substitute

$$n(t,x)=\frac{E(x)}{g(x)}u(t,x) \tag{2.5}$$

into (2.1) and obtain

$$\frac{\partial u}{\partial t}(t,x)+g(x)\frac{\partial u}{\partial x}(t,x)=k(x)u(t-r,y^{-1}(x)), \tag{2.6}$$

where

$$k(x)=\begin{cases}2\dfrac{g(x)p(y^{-1}(x))b(y^{-1}(x))}{E(x)y'(y^{-1}(x))g(y^{-1}(x))}E(y^{-1}(x)), & x\in[\alpha,\beta),\\ 0, \quad x\in(\beta,1). \end{cases} \tag{2.7}$$

The boundary condition (2.2) becomes

$$u(t,\alpha)=0 \tag{2.8}$$

and the initial condition (2.3) changes into

$$u(t,x)=\phi(t,x), \qquad t\in[-r,0], \quad x\in[\alpha,1], \tag{2.9}$$

where

$$\phi(t,x)=[g(x)/E(x)]\nu(t,x).$$

We shall look for solutions which are continuous functions of t with values in the Banach space $X=L^1[\alpha,1]$. Therefore we rewrite the problem (2.6), (2.8), (2.9) as the following abstract delay equation:

$$\frac{du(t)}{dt}=Bu(t)+Lu(t-r), \qquad t>0, \tag{2.10}$$

$$u(t)=\phi(t), \qquad t\in[-r,0]. \tag{2.11}$$

Here B is the unbounded closed linear operator defined by $B\psi=-g\psi'$ for all ψ in the domain $\mathscr{D}(B)=\{\psi\in X\,|\,\psi$ is absolutely continuous on $[\alpha,1]$, $\psi(\alpha)=0\}$ and L is the operator defined for all ψ in X by $(L\psi)(x)=k(x)\psi(y^{-1}(x))$. It follows from (H_b) that L is a bounded linear operator on X. ϕ is a given initial function in

$$C=C([-r,0];X).$$

The rest of the paper is devoted to the investigation of so-called mild solutions of the abstract problem (2.10)–(2.11).

3. Existence and uniqueness and the corresponding semigroup. It is obvious that the operator B defined at the end of §2 generates a strongly continuous semigroup $\{S(t)\}_{t\geqq 0}$ of linear operators on X. In fact, let

$$G(x)=\int_\alpha^x \frac{d\xi}{g(\xi)} \tag{3.1}$$

and define

$$X(t,x)=G^{-1}(G(x)+t), \qquad 0\leqq G(x)+t\leqq G(1). \tag{3.2}$$

(Note that G^{-1} is well defined on $[0,G(1)]$ because $g>0$.) Then $S(t)$ is given for every $\psi\in X$, every $t\geqq 0$ and almost every $x\in[\alpha,1]$ by

$$(S(t)\psi)(x)=\begin{cases}\psi(X(-t,x)) & \text{if } G(x)-t>0,\\ 0 & \text{if } G(x)-t\leqq 0.\end{cases} \tag{3.3}$$

Note that $S(t)=0$ for $t \geqq G(1)$. Observe that $X(t,x)$ is the solution of the initial value problem

$$\frac{dX}{dt}=g(X), \qquad X(0,x)=x \tag{3.4}$$

and hence $X(t,x)$ represents the size of a cell at time t which had size x at time zero.

If there exists a continuously differentiable function u satisfying (2.10), (2.11), it satisfies the following integral equation (variation of constants formula)

$$u(t)=S(t)\phi(0)+\int_0^t S(t-s)Lu(s-r)\,ds \tag{3.5}$$

for $t>0$. Any continuous function u which satisfies (3.5), (2.11) is called a *mild solution* of the initial value problem.

Travis and Webb (1974) have investigated existence, uniqueness and semigroup properties of a class of functional differential equations in Banach spaces. Some of their basic results can be applied to the present problem. As a special case of Proposition 2.1 of Travis and Webb (1974) we have

PROPOSITION 3.1. *For each $\phi \in C$ there exits a unique mild solution $u(\phi)$: $[-r,\infty)\to X$ of the initial value problem* (2.10), (2.11).

If u is a continuous function $[-r,\infty)\to X$ we denote by u_s $(s \geqq 0)$ the element of C defined by

$$u_s(\theta)=u(s+\theta), \qquad \theta \in [-r,0].$$

For each $t \geqq 0$ we define $T(t)$: $C \to C$ by $T(t)\phi=u(\phi)_t$, $\phi \in C$, where $u(\phi)$ is the unique mild solution of (2.10), (2.11) given by Proposition 3.1. The results of Travis and Webb (1974, Prop. 3.1.) give us the following:

PROPOSITION 3.2. *$\{T(t)\}_{t \geqq 0}$ is a strongly continuous semigroup of linear operators on C. The infinitesimal generator A of $\{T(t)\}_{t \geqq 0}$ is given by*

$$\mathscr{D}(A)=\{\phi \in C \mid \phi' \in C, \phi(0) \in \mathscr{D}(B), \phi'(0-)=B\phi(0)+L\phi(-r)\},$$
$$(A\phi)(\theta)=\phi'(\theta), \qquad \theta \in [-r,0].$$

One of our main objectives is to describe the large time behavior of mild solutions. Such information can be obtained from spectral properties of $T(t)$. If the semigroup is compact after finite time, then by a well-known spectral mapping theorem (cf. Pazy (1983, Chap. 2)) the spectrum of $T(t)$ is completely determined by the spectrum of its infinitesimal generator A. In the next section we give conditions under which $T(t)$ is indeed compact for t large enough and in §5 we use positivity arguments to give a rather precise characterization of the spectrum of A. It turns out that the same condition which ensures compactness of $T(t)$ guarantees the existence of a strictly dominant real eigenvalue of A. A combination of these results enables us to determine the asymptotic behavior of solutions.

4. Generation expansion and compactness of the semigroup. In the theory of linear autonomous differential-delay equations in finite-dimensional Euclidean spaces $\mathbb{R}^n$ the semigroup associated with the solution acts on the space $C([-r,0];\mathbb{R}^n)$ and it is a relatively easy consequence of Ascoli's theorem that the semigroup is compact for $t \geqq r$ (cf. Hale (1977, Chap. 7)). In our case $\mathbb{R}^n$ is replaced by the infinite-dimensional Banach space X and the proof of Lemma 1.1 of Hale (1977, Chap. 7) does not carry over to this case, simply because the Heine–Borel theorem fails in infinite-dimensional spaces. Neither can we use the compactness results of Travis and Webb (1974) since their results depend heavily on the assumption that the operator B generates a semigroup which which is compact for *all* $t>0$ and this condition is not satisfied in the problem under consideration.

In order to prove compactness of the semigroup corresponding to a related problem (without delay), Diekmann et al. (1984) wrote down a generation expansion for the solution. Here we shall use similar methods. The larger dimensionality of our state space makes the compactness proof a bit more involved than in the above-mentioned paper.

In the problem under consideration where we have to take account of individuals present at negative time, it makes sense to define also the -1st generation. We write

$$u(t;\phi)=\sum_{i=-1}^{\infty} u^i(t;\phi), \tag{4.1}$$

where

$$u^{-1}(t;\phi)=\begin{cases}\phi(t), & -r\leqq t\leqq 0,\\ 0, & t>0,\end{cases} \tag{4.2}$$

$$u^0(t;\phi)=S(t)\phi(0)+\int_0^t S(t-\tau)Lu^{-1}(\tau-r;\phi)\,d\tau, \qquad t\geqq 0 \tag{4.3}$$

and the higher generations are obtained by iteration of the integral operator.

$$u^{i+1}(t;\phi)=\int_0^t S(t-\tau)Lu^i(\tau-r;\phi)\,d\tau, \qquad t\geqq 0, \quad i\geqq 0. \tag{4.4}$$

Let for $i\geqq 0$ and $t\in[r,\infty)$ the operator family $T^i(t)$: $C\to C$ be defined by

$$T^i(t)\phi=u^i(\phi)_t. \tag{4.5}$$

Now let $i\geqq 0$, $t\geqq r$ and $\theta\in[-r,0]$, then

$$\begin{aligned}(T^{i+1}(t)\phi)(\theta)=u^{i+1}(t+\theta;\phi)&=\int_0^{t+\theta} S(t+\theta-\tau)Lu^i(\tau-r;\phi)\,d\tau\\ &=\int_{-\theta}^t S(t-s)Lu^i(s+\theta-r;\phi)\,ds\\ &=\int_{-\theta}^t S(t-s)L(T^i(s-r)\phi)(\theta)\,ds\\ &=\int_r^t S(t-s)L(T^i(s-r)\phi)(\theta)\,ds.\end{aligned}$$

For a bounded operator F: $X\to X$ we define the bounded operator $\tilde{F}$: $C\to C$ by $(\tilde{F}\phi)(\theta)=F(\phi(\theta))$ for all $\phi\in C$. Thus we can write

$$T^{i+1}(t)=\int_r^t \tilde{S}(t-s)\tilde{L}T^i(s-r)\,ds, \qquad t\geqq r, \quad i\geqq 0. \tag{4.6}$$

We note that $T(t)=\sum_{i=0}^{\infty}T^i(t)$, $t\geqq r$. If we can prove that $T^1(t)$ is compact for $t\geqq r$ then it follows from (4.6) that $T^{i+1}(t)$ is compact for $t\geqq r$ and $i\geqq 1$ and this finally yields that $T(t)$ is compact, $t\geqq r+G(1)$, since $T^0(t)=0$ if $t\geqq r+G(1)$. Therefore the rest of this section is concerned with a proof of the compactness of $T^1(t)$, $t\geqq r$.

The following version of Ascoli's theorem can be found in Martin (1976, Thm. II, 3.2.).

LEMMA 4.1. *A set V in C is precompact if the following conditions are satisfied*:

i) *V is bounded.*

ii) *The family V is equicontinuous.*

iii) *For each $\theta \in [-r,0]$ the subset $\{\phi(\theta) \mid \phi \in V\}$ of X is precompact.*

An easy calculation shows that

$$(4.7)\quad u^0(t,x;\phi)=\phi(0,X(-t,x))+\int_0^t k(X(-\tau,x))\phi\big(t-\tau-r,y^{-1}(X(-\tau,x))\big)\,d\tau,$$

$$(4.8)\quad u^1(t,x;\phi)=\int_0^t k(X(-\tau,x))\Big\{\phi\big(0,X(-t+\tau+r,y^{-1}(X(-\tau,x)))\big) + \int_0^{t-\tau-r} k\big(X(-\sigma,y^{-1}(X(-\tau,x)))\big) \cdot\phi\big(t-\sigma-\tau-2r,y^{-1}\big(X(-\sigma,y^{-1}(X(-\tau,x)))\big)\big)\,d\sigma\Big\}\,d\tau.$$

In trying to prove compactness of $T^1(t)$ it becomes clear that we need some relation between g and y. We shall make the following assumption:

Assumption 4.2.

$$g(x)y'(x)<g(y(x)),\qquad x_0\leqq x\leqq 1.$$

Below we shall give an interpretation of this inequality.

THEOREM 4.3. *If Assumption* 4.2 *is satisfied, then the semigroup $T(t)$ is compact for $t\geqq r+G(1)$.*

Proof. We have already explained that it suffices to show that $T^1(t)$ is compact for $t\geqq r$. Instead of (4.8) we can write

$$u^1(t,x;\phi)=u_1^1(t,x;\phi)+u_2^1(t,x;\phi),$$

where

$$u_1^1(t,x;\phi)=\int_0^t k(X(-\tau,x))\cdot\phi\big(0,X(-t+\tau+r,y^{-1}(X(-\tau,x)))\big)\,d\tau,$$

$$u_2^1(t,x;\phi)=\int_0^t k(X(-\tau,x))\Big\{\int_{-r}^{t-\tau-2r} k\big(X(s-t+\tau+2r,y^{-1}(X(-\tau,x)))\big) \cdot\phi\big(s,y^{-1}\big(X(s-t+\tau+2r,y^{-1}(X(-\tau,x)))\big)\big)\,ds\Big\}\,d\tau$$

where in this second expression we have substituted

$$s=t-\sigma-\tau-2r.$$

Let $T_j^1(t)$: $C\to C$ for $t\geqq r$, $j=1,2$ be given by

$$\big(T_j^1(t)\phi\big)(\theta)=u_j^1(t+\theta;\phi),\qquad \theta\in[-r,0].$$

Here we shall prove that $T_2^1(t)$ is compact for $t\geqq r$. The easier proof of compactness of $T_1^1(t)$, $t\geqq r$ is omitted.

Let $t\geqq r$ be fixed and let for $R>0$ the subset C_R of C be given by $C_R=\{\phi\in C \mid \|\phi\|_C\leqq R\}$. We will show that $V=\{T_2^1(t)\phi \mid \phi\in C_R\}$ obeys the conditions of Lemma 4.1. Obviously V is bounded. Now we replace τ by the variable

$$z=y^{-1}\big(X(s-t+\tau+2r,y^{-1}(X(-\tau,x)))\big).$$

Then

$$X(-s+t-\tau-2r,y(z))=y^{-1}(X(-\tau,x))=\xi.$$

Differentiation with respect to τ yields:

$$-g(\xi)+\frac{g(\xi)}{g(y(z))}\cdot y'(z)\frac{dz}{d\tau}=-\frac{g(y(\xi))}{y'(\xi)}.$$

Therefore, if Assumption 4.2 is satisfied, $dz/d\tau$ never becomes zero and replacing τ by z in the expression of u_2^1 one obtains

$$u_2^1(t,x;\phi)=\iint\limits_{\Omega(t,x)} Q(s,z;t,x)\phi(s,z)\,ds\,dz,$$

where $\iint_{\Omega(t,x)}ds\,dz$ is uniformly continuous in t and x in bounded subsets of the (t,x)-plane and $Q(s,z;\ t,x)$ is uniformly continuous in s,z,t and x in bounded subsets of the (s,z,t,x)-plane. At this point the reader will have no difficulty in seeing that V indeed obeys conditions (ii) and (iii) of Lemma 4.1. □

In the case where there is no delay, which has been studied by Diekmann et al. (1984), the function y is given by $y(x)=\frac{1}{2}x$ and Assumption 4.2 reduces to $\frac{1}{2}g(x)<g(\frac{1}{2}x)$, $x_0\leqq x\leqq 1$, and this is indeed the condition imposed in that paper in order to establish compactness of the semigroup.

To see the biological meaning of Assumption 4.2, consider two identical cells in the first phase with size $x>x_0$. Assume that one of the cells immediately enters the second phase. It will divide after r time units. Assume further that the two daughter cells will remain in the first phase for t time units. $t+r$ time units after our initial moment each daughter cell will have size $X(t,y(x))$. The other cell is assumed to behave differently. It first grows for t time units reaching size $X(t,x)$, then enters the second phase and finally at time $t+r$ divides into two daughter cells of size $y(X(t,x))$ each. Assumption 4.2 guarantees that

$$y(X(t,x))<X(t,y(x)). \tag{4.9}$$

This can be seen as follows. Differentiation of $G(x)-G(y(x))$ shows that this expression is increasing in x if Assumption 4.2 is satisfied. Now for $t>0$ and x, $\alpha\leqq x\leqq X(-t,1)$ we have that $x<X(t,x)$ and therefore

$$G(x)-G(y(x))<G(X(t,x))-G(y(X(t,x)))=t+G(x)-G(y(X(t,x)));$$

hence

$$G(y(X(t,x)))<t+G(y(x))$$

which implies (4.9). This thought experiment shows that the combination of growth and division provides a dispersion mechanism for cell size, which is essential for proving compactness and also, as we shall see in the following section, for proving some sort of strong positivity.

If $\beta<x_0$, then every cell has to pass size x_0 in each cycle. If Assumption 4.2 fails for all $x\in[x_0,1]$ (which corresponds to the case where individual cells grow exponentially throughout the cell cycle), then $\tau:=G(x)-G(x_0)+r+G(x_0)-G(y(x))$ is constant. But τ is the time elapsed between the event when the mother cell passes size x_0 and the event when the two daughter cells pass size x_0. Thus τ can be considered as the effective cycle time. In the case of exponential individual growth the cycle time τ is the same for all cells; it does not depend on size; there is no dispersion.

Finally we point out that Assumption 4.2 implies

$$y(x)<x \quad \text{for all } x\in(x_0,1) \tag{4.10}$$

which is a strengthening of (H_y). To see this, let $x\in(x_0,1)$ and take $t>0$ such that $X(t,x_0)=x$. Then (4.9) implies

$$y(x)=y(X(t,x_0))<X(t,y(x_0))\leqq X(t,x_0)=x. \tag{4.11}$$

5. The spectrum of $\boldsymbol{A}$. In this section we combine ideas similar to those of Travis and Webb (1974), Hale (1977, Chap. 7) and Heijmans (to appear, a) to describe the spectrum of the generator A and, in particular, to prove the existence of a strictly dominant, algebraically simple real eigenvalue. Assumption 4.2 is not presupposed unless this is explicitly stated.

Let us first introduce some notation. The norm of a Banach space Z is denoted by $\|\cdot\|_Z$. Z^* stands for the dual space of Z. We let $\langle\Phi,\phi\rangle_Z$ be the duality pairing of $\phi\in Z$, $\Phi\in Z^*$. For an operator T defined on a domain $\mathscr{D}(T)\subset Z$ with values in Z we let $\sigma(T)$, $P\sigma(T)$ and $\rho(T)$ denote the spectrum, point spectrum and resolvent set of T respectively. $r(T)$ is the spectral radius, $\mathscr{N}(T)$ the kernel and $\mathscr{R}(T)$ the range of T.

By definition, $\lambda\in\rho(A)$ if and only if the equation

$$(\lambda I-A)\phi=\psi \tag{5.1}$$

has a unique solution $\phi\in\mathscr{D}(A)$ for all ψ in C and ϕ depends continuously on ψ. By Proposition 3.2 each ϕ in $\mathscr{D}(A)$ is continuously differentiable on $[-r,0]$ and $A\phi=\phi'$. Hence (5.1) can be rewritten as

$$\lambda\phi(\theta)-\phi'(\theta)=\psi(\theta), \qquad \theta\in[-r,0] \tag{5.2}$$

and it follows that every solution ϕ of (5.2) is given by

$$\phi(\theta)=e^{\lambda\theta}\phi(0)+\int_\theta^0 e^{\lambda(\theta-s)}\psi(s)\,ds, \qquad \theta\in[-r,0]. \tag{5.3}$$

In particular,

$$\phi(-r)=e^{-\lambda r}\phi(0)+\int_{-r}^0 e^{-\lambda(r+s)}\psi(s)\,ds. \tag{5.4}$$

On the other hand, Proposition 3.2 also tells us that

$$\phi'(0)=B\phi(0)+L\phi(-r) \tag{5.5}$$

for all $\phi\in\mathscr{D}(A)$. Combining (5.2), (5.4) and (5.5), one obtains

$$\Delta(\lambda)\phi(0)=\psi(0)+H(\lambda)\psi, \tag{5.6}$$

where for each $\lambda\in\mathbf{C}$ the operator $\Delta(\lambda)$ with domain $\mathscr{D}(\Delta(\lambda))=\mathscr{D}(B)$ and values in X is defined by

$$\Delta(\lambda)=\lambda I-B-e^{-\lambda r}L \tag{5.7}$$

and $H(\lambda)$ is defined on all of C by

$$H(\lambda)\psi=L\left(\int_{-r}^{0}e^{-\lambda(r+s)}\psi(s)\,ds\right),\qquad \psi\in C. \tag{5.8}$$

We can now prove the following.

PROPOSITION 5.1. (a) $\lambda\in\sigma(A)$ *if and only if* $0\in\sigma(\Delta(\lambda))$.

(b) $\lambda\in P\sigma(A)$ *if and only if* $0\in P\sigma(\Delta(\lambda))$. *Moreover*, $\dim\mathcal{N}(\lambda I-A)=\dim\mathcal{N}(\Delta(\lambda))$.

Proof. (a) Above we have shown that if $\phi\in\mathcal{D}(A)$ is a solution of (5.1), then $\phi(0)$ satisfies (5.6). Conversely, if $\phi(0)\in X$ satisfies (5.6) then the function ϕ given by (5.3) belongs to $\mathcal{D}(A)$ and is a solution of (5.1). To complete the proof of (a), it suffices to show that the right-hand side of (5.6) covers X as ψ ranges over C. In order to see this, consider $\psi\in C$ given by $\psi(s)=f(s)w$ where $w\in X$ and the scalar function f defined on $[-r,0]$ satisfies (i) $f(0)=1$, (ii) $\int_{-r}^{0}e^{-\lambda s}f(s)\,ds=0$. It is obvious that $\psi(0)+H(\lambda)\psi=w$.

b) Suppose $\lambda\in P\sigma(A)$ and let $\phi\in C$, $\phi\neq0$ satisfy $A\phi=\lambda\phi$. Then $\phi(\theta)=\phi(0)e^{\lambda\theta}$ and $\Delta(\lambda)\phi(0)=0$. From $\phi\neq0$ it follows that $\phi(0)\neq0$ and therefore $0\in P\sigma(\Delta(\lambda))$. Similarly, $0\in P\sigma(\Delta(\lambda)\Rightarrow\lambda\in P\sigma(A)$. The second relation follows immediately. □

Proposition 5.1 characterizes the spectrum and the point spectrum of A acting in the space $C=C([-r,0],X)$ in terms of the operator $\Delta(\lambda)$ acting in the simpler space X. Below we shall investigate the spectral properties of $\Delta(\lambda)$ with the aid of yet another operator and eventually obtain a rather precise description of $\sigma(A)$.

Consider the equation

$$\Delta(\lambda)w=f, \tag{5.9}$$

that is,

$$\lambda w(x)+g(x)w'(x)-e^{-\lambda r}k(x)w(y^{-1}(x))=f(x) \tag{5.10}$$

where $f\in X$. We are looking for solution $w\in\mathcal{D}(\Delta(\lambda))=\mathcal{D}(B)$. Following Heijmans (to appear, a), we transform (5.10) into an integral equation by means of the following substitution:

$$w(x)=e^{-\lambda G(x)}v(x). \tag{5.11}$$

Then (5.10) takes the form

$$v'(x)-k_\lambda(x)v(y^{-1}(x))=\frac{f(x)}{g(x)}e^{\lambda G(x)}, \tag{5.12}$$

where by definition

$$k_\lambda(x)=\begin{cases}\dfrac{k(x)}{g(x)}e^{-\lambda[G(y^{-1}(x))-G(x)+r]}, & x\in[\alpha,\beta),\\ 0, & x\in[\beta,1).\end{cases} \tag{5.13}$$

Since w as a member of $\mathcal{D}(B)$ should be continuous and vanish at $x=\alpha$ the same must be true for v. We therefore look for solutions $v\in Y$ of (5.12) where Y is the Banach space

$$Y=\{v\in C[\alpha,1]\mid v(\alpha)=0\}. \tag{5.14}$$

Integration of (5.12) yields

$$v-K(\lambda)v=U(\lambda)f \tag{5.15}$$

where for $\lambda\in\mathbb{C}$ the operators $K(\lambda)$: $Y\to Y$ and $U(\lambda)$: $X\to X$ are defined by

$$(5.16)\qquad [K(\lambda)v](x)=\int_\alpha^{(x,\beta)^-} k_\lambda(\xi)v(y^{-1}(\xi))\,d\xi,\qquad v\in Y,\quad x\in[\alpha,1],$$

$$(5.17)\qquad [U(\lambda)f](x)=\int_\alpha^x \frac{f(\xi)}{g(\xi)}e^{\lambda G(\xi)}d\xi,\qquad f\in X.$$

The advantage of the formulation using the operators $K(\lambda)$ and $U(\lambda)$ is that these are compact (the proof of this fact is standard) and that $K(\lambda)$ has useful positivity properties. Observe that the range of $U(\lambda)$ lies in Y.

We can now prove the following theorem concerning some relations between the spectra of A, $\Delta(\lambda)$ and $K(\lambda)$.

THEOREM 5.2. *The following conditions are equivalent.*

(a) $\lambda\in\sigma(A)$.

(b) $\lambda\in P\sigma(A)$.

(c) $0\in\sigma(\Delta(\lambda))$.

(d) $0\in P\sigma(\Delta(\lambda))$.

(e) $1\in\sigma(K(\lambda))$.

(f) $1\in P\sigma(K(\lambda))$.

Moreover, if $K(\lambda)v=v$ *for some* $\lambda\in\mathbb{C}$ *and* $v\in Y$, *then* w *given by* (5.11) *belongs to* $\mathscr{D}(B)$ *and satisfies* $\Delta(\lambda)w=0$. *If* $0\notin\sigma(\Delta(\lambda))$, *then* $\Delta(\lambda)^{-1}$ *is compact.*

Proof. In Proposition 5.1 we have already proved (a)⇔(c) and (b)⇔(d).

Putting $f=0$ one observes by comparing (5.9) and (5.15) that (d)⇔(f) and that the eigenvector v belonging to the eigenvalue 1 of $K(\lambda)$ corresponds to the eigenvector w belonging to the eigenvalue 0 of $\Delta(\lambda)$.

(e)⇔(f) follows directly from the compactness of $K(\lambda)$.

Since trivially (b)⇒(a) it remains to show that (c)⇒(e). To this end, suppose that $1\notin\sigma(K(\lambda))$ which means that $I-K(\lambda)$ is invertible. For each $f\in X$ there exists therefore a unique solution $v\in Y$ of (5.15). But then w defined by (5.11) satisfies (5.9). Hence $0\in\sigma(\Delta(\lambda))$.

To prove compactness of $\Delta(\lambda)^{-1}$ for $0\in\rho(\Delta(\lambda))$ observe that since $U(\lambda)$ is compact and $(I-K(\lambda))^{-1}$ is bounded, the mapping $f\to v$ defined by (5.15) is compact. The transformation $v\to w$ defined by (5.11) is obviously bounded, hence $\Delta(\lambda)^{-1}$: $f\to w$ is compact. □

One important consequence of Theorem 5.2 is that the spectrum of A consists solely of eigenvalues ((a)⇔(b)). We emphasize that in order to establish this result we have not used compactness of $T(t)$, which would also imply the equivalence of (a) and (b).

Theorem 5.2 gives two entirely different characterizations of $\sigma(A)$—one in terms of $\Delta(\lambda)$, the other in terms of $K(\lambda)$. These characterizations will also be used for different purposes in the analysis to follow. $\Delta(\lambda)$ will prove to be of great importance in determining the algebraic and analytic properties of the eigenvalues and the resolvent operator of A. $K(\lambda)$ turns out to play a fundamental role in the investigation of the location of the eigenvalues in the complex plane.

We start by writing down an explicit expression for the resolvent operator $R(\lambda,A)$: $C\to C$ of A. It follows from (5.3) and (5.6) that for $\lambda\in\rho(A)(\neq\varnothing$, since by a standard result for semigroups $\lambda\in\rho(A)$ for all λ with $\operatorname{Re}\lambda$ large enough)

$$(5.18)\qquad (R(\lambda,A)\psi)(\theta)=e^{\lambda\theta}\left\{\Delta(\lambda)^{-1}(\psi(0)+H(\lambda)\psi)+\int_\theta^0 e^{-\lambda s}\psi(s)\,ds\right\}.$$

Hence $R(\lambda,A)=R_1(\lambda)+R_2(\lambda)$, where for $\lambda\in\rho(A)$, the bounded operators $R_1(\lambda)$ and $R_2(\lambda)$ are given by

$$(R_1(\lambda)\psi)(\theta)=e^{\lambda\theta}\cdot\Delta(\lambda)^{-1}(\psi(0)+H(\lambda)\psi), \qquad \psi\in C,$$

$$(R_2(\lambda)\psi)(\theta)=\int_\theta^0 e^{\lambda(\theta-s)}\psi(s)\,ds, \qquad \psi\in C.$$

From the boundedness of $H(\lambda)$: $C\to X$ and the compactness of $\Delta(\lambda)^{-1}$ if $\lambda\in\rho(A)$ it follows that $R_1(\lambda)$: $C\to C$ is compact if $\lambda\in\rho(A)$. Furthermore $R_2(\lambda)$ is quasinilpotent, i.e. $r(R_2(\lambda))=0$ if $\lambda\in\mathbb{C}$. This can be shown as follows. Define the norm $\|\cdot\|_\gamma$ on C as follows: $\|\psi\|_\gamma=\sup_{-r\leqq\theta\leqq 0}\|e^{-\gamma\theta}\psi(\theta)\|_X$. (This norm is equivalent to the original norm $\|\cdot\|_C$.) Now let $\gamma\in\mathbb{R}$ be such that $\gamma+\mathrm{Re}\lambda>0$. A straightforward calculation shows that $\|R_2(\lambda)\psi\|_\gamma\leqq(1/(\gamma+\mathrm{Re}\lambda))\|\psi\|_\gamma$. Therefore $r(R_2(\lambda))\leqq 1/(\gamma+\mathrm{Re}\lambda)$ for all $\gamma>-\mathrm{Re}\lambda$ and this yields the result. As a sum of a compact and a quasinilpotent operator $R(\lambda,A)$ is a *Riesz operator* (cf. Dowson (1978)). The following result was proved by Lay (1970, Thm. 4.6).

THEOREM 5.3. *Let Z be an infinite-dimensional Banach space and let T be a closed operator on Z with nonempty resolvent set. Suppose that there exists an $\alpha\in\rho(T)$ such that $R(\alpha,T)$ is a Riesz operator. Then $\sigma(T)$ is a countable set of poles of $R(\lambda,T)$ of finite rank with ∞ the only possible point of accumulation.*

As a consequence we have the following result.

COROLLARY 5.4. *If $\lambda_0\in\sigma(A)$ then λ_0 is a pole of $R(\lambda,A)$ with residue of finite rank.*

Remark 5.5. (a) For all $\lambda_0\in\sigma(A)$ we have that λ_0 is a pole of order p of $(\Delta(\lambda))^{-1}$ iff λ_0 is a pole of order p of $R(\lambda,A)$. This follows easily from (5.18) and the fact that $e^{\lambda\theta}$, the operator $H(\lambda)$ and the operator from C to X given by $\psi\to\int_\theta^0 e^{-\lambda s}\psi(s)\,ds$ define entire functions.

(b) If Assumption 4.2 is satisfied, then the semigroup $T(t)$ is compact after finite time, and therefore the *Browder essential spectrum* (see e.g. Webb (1985) for a definition) $\sigma_{\mathrm{ess}}(T(t))=\{0\}$, $t>0$, and now Corollary 5.4 follows immediately from Proposition 4.13 of Webb (1985) which says among other things

$$\{e^{\lambda t}\,|\,\lambda\in\sigma_{\mathrm{ess}}(A)\}\subset\sigma_{\mathrm{ess}}(T(t)), \qquad t>0.$$

The operator $K(\lambda)$ is very similar to an operator studied by Heijmans (to appear, a). Using essentially the same methods, based on the positivity of $K(\lambda)$ for $\lambda\in\mathbb{R}$, one can prove the following result. For readers consulting the above mentioned reference we mention that $K(\lambda)$ corresponds to T_λ and that $x_0>\alpha$ and $x_0=\alpha$ respectively correspond to the cases $a>0$ and $a=0$ of that paper.

LEMMA 5.6. *There exists a $\lambda_d\in\mathbb{R}$ such that*

i) 1 *is an algebraically simple eigenvalue of* $K(\lambda_d)$.

ii) *The associated eigenvector $v_d\in Y$ is strictly positive on* $(\alpha,1]$.

iii) *All elements $\lambda\in\sigma(A)$ satisfy* $\mathrm{Re}\lambda\leqq\lambda_d$.

Let X_+ be the subset of X consisting of all functions which are nonnegative a.e.; then X_+ defines a cone in X and with the induced ordering X is a Banach lattice (see e.g. Schaefer (1974)). Define C_+ as

$$C_+=\{\phi\in C\,|\,\phi(\theta)\in X_+,\theta\in[-r,0]\}.$$

With the ordering induced by the cone C_+ the space C becomes a Banach lattice as well.

Now we can prove the following important result.

THEOREM 5.7. *The eigenvalue* λ_d *of* A *is algebraically simple. The eigenvector* ϕ_d *satisfies* $\phi_d(\theta,x)=e^{\lambda_d\theta}w_d(x)$ *where* $w_d\in Y$ *and* $w_d(x)>0$, $x\in(\alpha,1]$. *The dual eigenvector* Φ_d, *determined by* $A^*\Phi_d=\lambda_d\Phi_d$ *is strictly positive, i.e.* $\phi\in C_+$, $\phi\neq 0$ *implies that* $\langle\Phi_d,\phi\rangle_c>0$.

Proof. As in Theorem 5.4 of Heijmans (to appear, a) we can show that λ_d is a simple pole of $\Delta(\lambda)^{-1}$ and that the eigenvalue 0 of $\Delta(\lambda_d)$ has geometric multiplicity one. Combined with Proposition 5.1(b) and Remark 5.5(a) this yields the algebraic simplicity of the eigenvalue λ_d of A. Let v_d be given by Lemma 5.6 and w_d by (5.11); then $\Delta(\lambda_d)w_d=0$. Now $\phi_d\in C_+$ given by $\phi_d(\theta)=e^{\lambda_d\theta}w_d$ satisfies indeed the conditions stated in the theorem.

An easy calculation shows that $R(\lambda,A)$ defines a positive operator with respect to the cone C_+ if $\lambda>\lambda_d$. Now let $\lambda_0>\lambda_d$ be fixed. Then $r(R(\lambda_0,A))=1/(\lambda_0-\lambda_d)$ and a standard result from positive operator theory says that $R(\lambda_0,A)^*\Phi_d=(1/(\lambda_0-\lambda_d))\Phi_d$ for some positive functional $\Phi_d\neq 0$. Since $R(\lambda_0,A)^*=R(\lambda_0,A^*)$ (cf. Taylor and Lay (1980)) we obtain that $A^*\Phi_d=\lambda_d\Phi_d$. Now suppose that Φ_d is not strictly positive, i.e., there is a $\psi\in C_+$, $\psi\neq 0$ such that $\langle\Phi_d,\psi\rangle_c=0$. Then $\psi\in\mathcal{N}(\lambda_dI-A^*)^\perp=\mathcal{R}(\lambda_dI-A)$; hence $\lambda_d\phi-A\phi=\psi$ for some $\phi\in C$, hence

$$\Delta(\lambda_d)\phi(0)=\psi(0)+H(\lambda_d)\psi+X_+\backslash\{0\}.$$

A calculation very similar to the one performed in the proof of Theorem 5.4 of Heijmans (to appear, a) shows that

$$\mathcal{R}(\Delta(\lambda_d))\cap X_+=\{0\}$$

and this is a contradiction. Therefore Φ_d is strictly positive. □

An important question is whether or not the eigenvalue λ_d is strictly dominant, i.e. $\mathrm{Re}\lambda<\lambda_d$ if $\lambda\in\sigma(A)$, $\lambda\neq\lambda_d$. If Assumption 4.2 is satisfied, this would immediately imply that there exists a positive ε such that $\mathrm{Re}\lambda<\lambda_d-\varepsilon$ if $\lambda\in\sigma(A)$, $\lambda\neq\lambda_d$, because if this were not true then there would exist a sequence $\lambda_n\in\sigma(A)$ such that $\mathrm{Re}\lambda_n$ is strictly increasing and $\mathrm{Re}\lambda_n\to\lambda_d$, $n\to\infty$. But if $t>0$ is such that $T(t)$ is compact, then $e^{\lambda_nt}\in\sigma(T(t))$ and $|e^{\lambda_nt}|=e^{\mathrm{Re}\lambda_n\cdot t}\to e^{\lambda_dt}$, $n\to\infty$ which implies that $\sigma(T(t))$ has an accumulation point different from zero contradicting the compactness of $T(t)$.

The answer to the question concerning the strict dominance depends strongly on the dependence of the kernel $k_\lambda(x)$ of the operator $K(\lambda)$ on λ. As in Heijmans (to appear, a) we can prove the following result.

THEOREM 5.8. *If Assumption* 4.2 *is satisfied, then there is an* $\varepsilon>0$ *such that* $\mathrm{Re}\lambda<\lambda_d-\varepsilon$ *if* $\lambda\in\sigma(A)\backslash\{\lambda_d\}$.

If Assumption 4.2 is false for every $x\in[x_0,1]$, then

$$G(\gamma^{-1}(\xi))-G(\xi)=c,\qquad \alpha\leqq\xi\leqq\beta,$$

where c is a constant, and we find that

$$K(\lambda)=e^{-\lambda(r+c)}K(0),$$

where the operator $K(0)$ does not depend on λ.

It follows immediately that in this case

$$\lambda\in\sigma(A)\Rightarrow\lambda+k\cdot\frac{2\pi i}{r+c}\in\sigma(A),\qquad k\in\mathbb{Z},$$

and Theorem 5.8 is certainly not true in this case. If Assumption 4.2 is fulfilled on a nonempty subset of $[x_0,1]$ then the situation is more complicated but one can prove that Theorem 5.8 is still true (see Diekmann, Heijmans and Thieme (1985)).

As in Heijmans (to appear, a) it is possible to compute the so-called characteristic equation from which all eigenvalues of A can be calculated in principle. Here we shall only do this for the special case $\beta < x_0$.

Let $\beta < x_0$ and let v be a solution of

$$K(\lambda)v=v.$$

Then $v(x)$ is constant for $\beta \leqq x \leqq 1$ and we may take $v(x)=1$, $\beta \leqq x \leqq 1$. Then

$$v(x)=\int_\alpha^x k_\lambda(\xi)v(y^{-1}(\xi))\,d\xi=\int_\alpha^x k_\lambda(\xi)\,d\xi, \qquad \alpha \leqq x < \beta.$$

Since v has to be continuous in $x=\beta$, we obtain

$$1=\int_\alpha^\beta k_\lambda(\xi)\,d\xi \tag{5.19}$$

and this equation determines the elements of $\sigma(A)$ if $\beta < x_0$.

6. The stable size distribution. Throughout this section we assume that Assumption 4.2 is satisfied. Let λ_d be the strictly dominant eigenvalue of A, and let ϕ_d, Φ_d be given by Theorem 5.7. Since λ_d is a simple pole of $R(\lambda,A)$, we have the following decomposition of the state space C (cf. Taylor and Lay (1980)):

$$C=\mathscr{N}(\lambda_d I-A)\oplus\mathscr{R}(\lambda_d I-A), \tag{6.1}$$

where $\mathscr{N}(\lambda_d I-A)$ is the one-dimensional space spanned by the positive eigenvector ϕ_d. Let P be the orthogonal projection on $\mathscr{N}(\lambda_d I-A)$ according to this decomposition; then P is given by

$$P\phi=\langle\Phi_d,\phi\rangle_c\cdot\phi_d, \tag{6.2}$$

where we have normalized Φ_d, ϕ_d such that $\langle\Phi_d,\phi_d\rangle_c=1$. Let $\tilde{T}(t)$ be the restriction of $T(t)$ to $\mathscr{R}(\lambda_d I-A)$; then $r(\tilde{T}(t))\leqq e^{(\lambda_d-\varepsilon)t}$, $t\geqq 0$, where we have used Theorem 5.8. A standard result from semigroup theory says that for all $0<\eta<\varepsilon$ there exits an $M(\eta)\geqq 1$ such that $\|\tilde{T}(t)\psi\|\leqq M(\eta)e^{(\lambda_d-\eta)t}\|\psi\|$, for all $\psi\in\mathscr{R}(\lambda_d I-A)$. Let $\phi\in C$; then $\phi=P\phi+(I-P)\phi=\langle\Phi_d,\phi\rangle_c\cdot\phi_d+(I-P)\phi$ and therefore

$$T(t)\phi=\langle\Phi_d,\phi\rangle_c\cdot e^{\lambda_d t}\cdot\phi_d+\tilde{T}(t)(I-P)\phi, \qquad t\geqq 0$$

and the following result is obtained.

THEOREM 6.1. *For all $0<\eta<\epsilon$ there is a constant $M(\eta)\geqq 1$ such that for all $\phi\in C$*

$$\left\|T(t)\phi-\langle\Phi_d,\phi\rangle_c\cdot e^{\lambda_d t}\cdot\phi_d\right\|\leqq M(\eta)e^{(\lambda_d-\eta)t}\|\phi\|, \qquad t\geqq 0.$$

For obvious reasons we call ϕ_d the stable size distribution.

Finally we mention that there is an alternative way to reach the main results exploiting the positivity of the semigroup. Using known results from positive semigroup theory (cf. Greiner (1981)), Theorem 5.8 follows immediately. The main problem is now to establish the algebraic simplicity of λ_d. This can be done by showing that the semigroup is not only positive but also irreducible (cf. Schaefer (1974), Greiner (1981)). However, the technical difficulties arising in this approach seem to be greater than in the one we have adopted.

Acknowledgments. We gratefully acknowledge a number of valuable suggestions by Odo Diekmann and some stimulating discussions and correspondence with John Tyson on the biological aspects of the paper.

REFERENCES

G. I. Bell and E. C. Anderson (1967), *Cell growth and division*, I, *A mathematical model with applications to cell volume distributions in mammalian suspension cultures*, Biophys. J., 7, pp. 329–351.

O. Diekmann, H. J. A. M. Heijmans and H. R. Thieme (1984), *On the stability of the cell size distribution*, J. Math. Biology, 19, pp. 227–248.

______ (1985), *On the stability of the cell size distribution*, II *Time periodic developmental rates*, preprint.

H. R. Dowson (1978), *Spectral Theory of Linear Operators*, Academic Press, London–New York–San Francisco.

M. Eisen (1979), *Mathematical models in Cell Biology and Cancer Chemotherapy*, Lecture Notes in Biomathematics 30, Springer-Verlag, Berlin–Heidelberg–New York.

G. Greiner (1981), *Zur Perron–Frobenius Theorie stark stetiger Halbgruppen*, Math. Z., 177, pp. 401–423.

M. Gyllenberg (to appear), *The size and scar distributions of the yeast Saccharomyces cerevisiae*, J. Math. Biology.

J. K. Hale (1977), *Theory of Functional Differential Equations*, Springer-Verlag, Berlin–Heidelberg–New York.

K. B. Hannsgen and J. J. Tyson (1984), *Stability of the steady-state size distribution in a model of cell growth and division*, preprint.

H. J. A. M. Heijmans (to appear, a), *An eigenvalue problem related to cell growth*, J. Math. Anal. Appl.

______ (to appear, b), *The dynamical behavior of the age-size distribution of a cell population*, in Dynamics of Physiologically Structured Populations, J. A. J. Metz and O. Diekmann, eds., Lecture Notes in Biomathematics, Springer-Verlag, Berlin–Heidelberg–New York.

D. C. Lay (1970), *Spectral analysis using ascent, descent, nullity and defect*, Math. Ann., 184, pp. 197–214.

R. H. Martin, Jr. (1976), *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York–London–Sydney–Toronto.

J. A. J. Metz and O. Diekmann (to appear), *Dynamics of Physiologically Structured Populations*, Lecture Notes in Biomathematics, Springer-Verlag, Berlin–Heidelberg–New York.

A. Pazy (1983), *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York.

H. H. Schaefer (1974), *Banach Lattices and Positive Operators*, Springer, New York.

J. W. Sinko and W. Streifer (1971), *A model for populations reproducing by fission*, Ecology, 52, pp. 330–335.

A. E. Taylor and D. C. Lay (1980), *Introduction to Functional Analysis*, John Wiley, New York–Chichester–Brisbane–Toronto.

C. C. Travis and G. F. Webb (1974), *Existence and stability for partial differential equations*, Trans. Amer. Math. Soc., 200, pp. 395–418.

J. J. Tyson and K. B. Hannsgen (1984), *The distribution of cell size and generation time in a model of the cell cycle incorporating size control and random transitions*, preprint.

G. F. Webb (1985), *Theory of Nonlinear Age-dependent Population Dynamics*, Marcel Dekker, New York.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

007

WEAK SOLUTIONS OF AN INITIAL BOUNDARY VALUE PROBLEM FOR AN INCOMPRESSIBLE VISCOUS FLUID WITH NONNEGATIVE DENSITY*

JONG UHN KIM†

Abstract. The initial boundary value problem associated with the motion of an incompressible viscous fluid with variable density is studied. The lower bound of density is allowed to be zero and the local existence of a weak solution is established.

Key words. local existence, weak solutions, semi-Galerkin method, compactness property

AMS(MOS) subject classifications. Primary 35D05, 35Q10, 76D0

Introduction. In this paper we establish the local existence of weak solutions to the initial boundary value problem associated with the motion of a fluid with nonnegative density. We assume that the fluid is contained in a region Ω of space and that the fluid is nonhomogeneous, incompressible and viscous. Let $u(t,x)$, $\rho(t,x)$ and $p(t,x)$ denote the unknown velocity vector, the density and the pressure of the fluid at point x at time t. Then the governing equations are

$$\rho\frac{\partial u}{\partial t}+(\rho u\cdot\nabla)u-\Delta u=\rho f+\nabla p, \tag{0.1}$$

$$\nabla\cdot u=0, \tag{0.2}$$

$$\frac{\partial\rho}{\partial t}+(u\cdot\nabla)\rho=0, \tag{0.3}$$

where $f(t,x)$ is the known external force. The boundary condition is given by

$$u(t,x)=0 \quad \text{for } t>0,\ x\in\partial\Omega, \tag{0.4}$$

and the initial conditions are expressed by

$$u(0,x)=u_0(x), \qquad \rho(0,x)=\rho_0(x), \tag{0.5}$$

where $u_0(x)$ and $\rho_0(x)$ are given functions. Here Ω is an open bounded subset of R^3 with a smooth boundary. We assume that the coefficient of viscosity is uniformly a constant; it is taken to be 1 for convenience.

Kazhikhov [4] established the existence of weak solutions of (0.1) to (0.5) under the assumption that $0<\alpha\leqq\rho_0(x)\leqq M<\infty$, a.e. in Ω; for the definition of weak solution, see §1. His work was reviewed by Lions [6], who raised the question of existence of solutions in the case $0\leqq\rho_0(x)\leqq M$. In fact, Lions [6] proved the existence of weak solutions to a variant of the above system of equations for a penalized model when $0\leqq\rho_0(x)\leqq M<\infty$, a.e. in Ω. In this paper we show that (0.1) to (0.5) admit a local weak solution if we assume some more regularity on $u_0(x)$ and $f(t,x)$. For the precise statement, see Theorem 2.1 below. Our basic approach to this problem is parallel with

* Received by the editors March 25, 1985, and in final form September 12, 1985.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

that of Kazhikhov [4]: we employ the so-called "semi-Galerkin method". We use, however, different energy estimates which are derived capitalizing on the regularity of $u_0(x)$. These estimates are independent of lower bounds of $\rho_0(x)$ and are strong enough to use standard compactness properties of a sequence of approximating solutions. In Kazhikhov [4], energy estimates were obtained under the essential assumption that $0<\alpha\leqq\rho_0(x)\leqq M$, a.e. in Ω and a delicate compactness property was necessary; see Lions [6] for detailed exposition. In §1, we explain our notations and present some mathematical preliminaries. Detailed proof of our result is given in §2.

1. Notation and preliminaries. $x=(x_1,x_2,x_3)$ is the space variable in R^3. For $x,y\in R^3$, $x\cdot y=\sum_{i=1}^3 x_iy_i$. ∇ is the gradient and Δ is the Laplacian. When $f(x)=(f_1(x),f_2(x),f_3(x))$ is an R^3-valued function,

$$\nabla\cdot f=\sum_{i=1}^3\frac{\partial f_i}{\partial x_i},\quad |\nabla f|^2=\sum_{i,j=1}^3\left|\frac{\partial f_i}{\partial x_j}\right|^2,\quad \|f\|_{L^p(\Omega)}=\left(\sum_{i=1}^3\|f_i\|^p_{L^p(\Omega)}\right)^{1/p},$$

$$\|\nabla f\|_{L^p(\Omega)}=\left(\sum_{i,j=1}^3\left\|\frac{\partial f_i}{\partial x_j}\right\|^p_{L^p(\Omega)}\right)^{1/p}.$$

For $s\in R$, $H^s(\Omega)$ denotes the usual Sobolev space as defined in Lions and Magenes [7]. Let $W=\{\phi\in C_0^\infty(\Omega)^3\colon \nabla\cdot\phi=0\}$. Then $H_\sigma^s(\Omega)$ is defined to be the completion of W with respect to the norm of $H^s(\Omega)^3$. When E is a Banach space, $L^p(0,T;E)$ stands for the space of functions g, strongly measurable on $[0,T]$ with range in E such that

$$\|g\|_{L^p(0,T;E)}\overset{\text{def}}{=}\left(\int_0^T\|g(t)\|_E^p dt\right)^{1/p}<\infty,\quad \text{for } 1\leqq p<\infty$$

and

$$\|g\|_{L^\infty(0,T;E)}\overset{\text{def}}{=}\operatorname*{ess\,sup}_{t\in[0,T]}\|g(t)\|_E<\infty,\quad \text{for } p=\infty.$$

$C^k([0,T];E)$ is the space of functions which are k-times continuously differentiable on $[0,T]$ with range in E and when $k=0$, we omit the superscript. The norm is defined in the obvious way. If $E=R$, we denote it by $C^k([0,T])$.

Now we give the definition of weak solution (0.1) to (0.5).

DEFINITION 1.1. A weak solution of (0.1) to (0.5) is a pair of functions $u(t,x)$, $\rho(t,x)$ such that $u(t,x)=(u_1(t,x),u_2(t,x),u_3(t,x))\in L^2(0,T;H_\sigma^1(\Omega))$, $\rho(t,x)\in L^\infty([0,T]\times\Omega)$ and

$$(1.1)\quad -\int_0^T\int_\Omega\rho u\cdot\frac{\partial\Phi}{\partial t}dx\,dt-\sum_{j=1}^3\int_0^T\int_\Omega\rho u_j u\cdot\frac{\partial\Phi}{\partial x_j}dx\,dt+\sum_{j=1}^3\int_0^T\int_\Omega\frac{\partial u}{\partial x_j}\cdot\frac{\partial\Phi}{\partial x_j}dx\,dt$$
$$=\int_0^T\int_\Omega\rho f\cdot\Phi\,dx\,dt+\int_\Omega\rho_0(x)u_0(x)\cdot\Phi(0,x)\,dx,$$

$$(1.2)\quad \int_0^T\int_\Omega\rho\frac{\partial\psi}{\partial t}dx\,dt-\sum_{j=1}^3\int_0^T\int_\Omega\rho u_j\frac{\partial\psi}{\partial x_j}dx\,dt=\int_\Omega\rho_0(x)\psi(0,x)\,dx$$

hold for all $\Phi\in C^1([0,T];H_\sigma^1(\Omega))$ and $\psi\in C^1([0,T];H^1(\Omega))$ satisfying $\Phi(T,x)=0,\psi(T,x)=0$, a.e. in Ω, where $\rho_0(x)\in L^\infty(\Omega)$ and $u_0(x)\in H_\sigma^0(\Omega)$ are given functions.

Next we list some basic facts for later use. Let the operator P be the projection of $L^2(\Omega)^3$ onto $H_\sigma^0(\Omega)$ and consider the eigenvalue problem:

$$P\Delta\phi_k+\lambda_k\phi_k=0 \quad \text{in } \Omega, \qquad \phi_k=0 \quad \text{on } \partial\Omega. \tag{1.3}$$

LEMMA 1.2. *$P\Delta$ is a self-adjoint operator in $H_\sigma^0(\Omega)$ and its inverse is compact. There are countably many eigenvalues $(-\lambda_k)_{k=1}^\infty$ and corresponding eigenfunctions $\{\phi_k\}_{k=1}^\infty$ such that $0<\lambda_1\leqq\lambda_2\leqq\cdots,\lambda_k\to\infty$ as $k\to\infty$ and*

$$\int_\Omega \phi_j\cdot\phi_k\,dx=\delta_{jk}. \tag{1.4}$$

For detailed discussion on the convergence of eigenfunction expansions and the regularity of eigenfunctions, see Ladyzhenskaya [5]. The following lemma is a special case of the result in Aubin [1].

LEMMA 1.3. *Let E_0, E, E_1 be Banach spaces such that $E_0\subset E\subset E_1$, E_0 and E_1 are reflexive, and the injection $E_0\to E$ is compact. Define $\mathscr{F}=\{v\in L^{P_0}(0,T;E_0)$: $dv/dt\in L^{P_1}(0,T;E_1)\}$, where $0<T<\infty$ and $1<P_0, P_1<\infty$. The norm of $\mathscr{F}$ is defined by*

$$\|v\|_{L^{P_0}(0,T;E_0)}+\left\|\frac{dv}{dt}\right\|_{L^{P_1}(0,T;E_1)}.$$

Then the injection $\mathscr{F}\to L^{P_0}(0,T;E)$ is compact.

For the following lemmas, see Lions and Magenes [7].

LEMMA 1.4. *For any $s\in R$ and any $\varepsilon>0$, the injection $H^s(\Omega)\to H^{s-\varepsilon}(\Omega)$ is compact.*

LEMMA 1.5. *Let $f\in H^2(\Omega)\cap H_0^1(\Omega)$. Then it holds that*

$$\|f\|_{H^{1+\theta}(\Omega)}\leqq C_\theta\|\nabla f\|_{L^2(\Omega)}^{1-\theta}\|\Delta f\|_{L^2(\Omega)}^{\theta} \quad \textit{for } 0\leqq\theta\leqq 1 \tag{1.5}$$

where C_θ depends only on θ and Ω.

LEMMA 1.6. *For any $\delta>\frac{3}{2}$, $H^\delta(\Omega)\subset C(\bar\Omega)$ and $(f,g)\to fg$ is a continuous mapping from $H^\delta(\Omega)\times H_0^1(\Omega)$ into $H_0^1(\Omega)$ and from $H^\delta(\Omega)\times H^{-1}(\Omega)$ into $H^{-1}(\Omega)$.*

For $f\in H^\delta(\Omega)$, $g\in H^{-1}(\Omega)$, fg is defined by $\langle fg,\psi\rangle=\langle g,f\psi\rangle$ for all $\psi\in H_0^1(\Omega)$, where $\langle\ ,\ \rangle$ is the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. This multiplication coincides with the pointwise multiplication for $g\in L^2(\Omega)$.

2. Local existence. In this section we state and prove the main result:

THEOREM 2.1. *Suppose $f(t,x)\in L^\infty(0,T^*;L^2(\Omega)^3)$, $u_0(x)\in H_\sigma^1(\Omega)$, $\rho_0(x)\in L^\infty(\Omega)$ and $0\leqq\rho_0(x)\leqq M<\infty$, a.e. in Ω. Then there is a number $T\in(0,T^*]$ and a weak solution $u(t,x)$, $\rho(t,x)$ of* (0.1) *to* (0.5) *such that $\rho(t,x)\in L^\infty(0,T\times\Omega)$, $u(t,x)\in L^2(0,T;H^2(\Omega)^3)\cap L^\infty(0,T;H_\sigma^1(\Omega))$.*

The proof of this result is divided into several steps. First we consider the initial value problem:

$$\frac{\partial\rho}{\partial t}+(v\cdot\nabla)\rho=0, \qquad \rho(0,x)=\rho_0(x). \tag{2.1}$$

LEMMA 2.2. *Suppose $v(t,x)\in C([0,T];C^1(\bar\Omega)^3)$, $\nabla\cdot v=0$ for all $(t,x)\in[0,T]\times\bar\Omega$, $v=0$ for all $(t,x)\in[0,T]\times\partial\Omega$ and $\rho_0(x)\in C^1(\bar\Omega)$, $\alpha\leqq\rho_0(x)\leqq\beta$ for all $x\in\bar\Omega$, where $\alpha,\beta\in R$. Then* (2.1) *has a unique solution $\rho(t,x)$ in $C^1([0,T]\times\bar\Omega)$. Furthermore, $\alpha\leqq\rho(t,x)\leqq\beta$ holds for all $(t,x)\in[0,T]\times\bar\Omega$.*

LEMMA 2.3. *For each $n=1,2,\cdots$, let $v_n(t,x)\in C([0,T];C^1(\bar\Omega)^3)$, $\nabla\cdot v_n=0$ for all $(t,x)\in[0,T]\times\bar\Omega$ and $v_n=0$ for all $(t,x)\in[0,T]\times\partial\Omega$. Suppose that v_n converges to v in $C([0,T];C^1(\bar\Omega)^3)$, and denote by $\rho_n(t,x)$, $\rho(t,x)$ the unique solution of*

$$\frac{\partial\rho_n}{\partial t}+(v_n\cdot\nabla)\rho_n=0, \qquad \rho_n(0,x)=\rho_0(x) \tag{2.1n}$$

and the unique solution of (2.1), *respectively, where* $\rho_0(x)$ *is the same as in Lemma* 2.2. *Then* $\rho_n(t,x)$ *converges to* $\rho(t,x)$ *in* $C([0,T]\times\bar{\Omega})$.

Proof of Lemma 2.2. We use the classical method of characteristics to construct a solution. Let E be an open ball in R^3 such that $\bar{\Omega}\subset E$. We extend v to $w\in C([0,T];C^1(\bar{E})^3)$ so that $v\equiv w$ for all $(t,x)\in[0,T]\times\bar{\Omega}$. Consider the system:

$$\frac{dx}{dt}=w(t,x),\qquad x(0)=y\in\bar{\Omega}. \tag{2.2}$$

Then there is $0<\tilde{T}\leqq T$ such that (2.2) has a unique solution $x(t,y)$ in $C^1([0,\tilde{T}];C^1(\bar{\Omega})^3)$. If $y\in\partial\Omega$, then $x(t,y)=y$ for all $t\in[0,\tilde{T}]$ since $w(t,x)=v(t,x)=0$ for all $(t,x)\in[0,T]\times\partial\Omega$. If $y\in\Omega$, then $x(t,y)\in\Omega$ for all $t\in[0,\tilde{T}]$ by the uniqueness of solution. Hence, we may take $\tilde{T}=T$ and replace w by v in (2.2). Furthermore, $\det\{\partial x_i/\partial y_j\}=1$ for each $(t,y)\in[0,T]\times\bar{\Omega}$ since $v(t,x)$ is a divergence-free vector field. It is apparent that for each $t\in[0,T]$, the mapping S_t defined by S_t: $y\to x(t,y)$ is C^1-diffeomorphism of $\bar{\Omega}$ onto itself and that $y=S_t^{-1}x=y(t,x)\in C^1([0,T]\times\bar{\Omega})^3$. Now define $\rho(t,x)=\rho_0(y(t,x))$. Then $\rho(t,x)$ is the unique solution in $C^1([0,T]\times\bar{\Omega})$, which can be easily shown by the classical argument.

Proof of Lemma 2.3. Let $x_n(t,y)\in C^1([0,T];C^1(\bar{\Omega})^3)$ be the solution of

$$\frac{dx_n}{dt}=v_n(t,x_n),\qquad x_n(0,y)=y\in\bar{\Omega}. \tag{2.2n}$$

Then $x_n(t,y)$ converges to $x(t,y)$, the solution of (2.2), uniformly in $[0,T]\times\bar{\Omega}$; see Hale [3]. For each $t\in[0,T]$, define $y_n(t,\cdot)$ to be the inverse of the mapping $y\to x_n(t,y)$. Then it is easy to see that all the first order derivatives of $y_n(t,x)$ are uniformly bounded with respect to n and $(t,x)\in[0,T]\times\bar{\Omega}$. By Ascoli's theorem, we derive that $y_n(t,x)$ converges to $y(t,x)$ uniformly in $[0,T]\times\bar{\Omega}$, from which it follows that $\rho_n(t,x)$ converges to $\rho(t,x)$ uniformly in $[0,T]\times\bar{\Omega}$, from which it follows that $\rho_n(t,x)$ converges to $\rho(t,x)$ uniformly in $[0,T]\times\bar{\Omega}$.

Next we choose sequences of functions $\{\rho_{0m}(x)\}_{m=1}^{\infty}$, $\{f_m(t,x)\}_{m=1}^{\infty}$ such that $\rho_{0m}(x)\in C^1(\bar{\Omega})$, $1/m\leqq\rho_{0m}(x)\leqq M+(1/m)$ for all $x\in\bar{\Omega}$, $\rho_{0m}(x)\to\rho_0(x)$ in $L^2(\Omega)$, and $f_m(t,x)\in C([0,T^*];L^2(\Omega)^3)$, $\|f_m(t,x)\|_{L^2(\Omega)^3}\leqq\|f(t,x)\|_{L^\infty(0,T^*;L^2(\Omega)^3)}$ for all $t\in[0,T^*]$, $f_m(t,x)\to f(t,x)$ in $L^2(0,T^*;L^2(\Omega)^3)$ where $\rho_0(x)$ and $f(t,x)$ are given functions in Theorem 2.1. Recalling Lemma 1.2, we set

$$U_m(t,x)=\sum_{k=1}^{m}A_{mk}(t)\phi_k(x) \tag{2.3}$$

and consider the system of equations:

$$\frac{\partial\rho_m}{\partial t}+(U_m\cdot\nabla)\rho_m=0, \tag{2.4}$$

$$\begin{aligned}\sum_{k=1}^{m}b_{jk}^m(t)\frac{d}{dt}A_{mk}(t)+\sum_{k,l=1}^{m}C_{jkl}^m(t)A_{mk}(t)A_{ml}(t)+\lambda_jA_{mj}(t)&\\ =d_j^m(t),\qquad j=1,\cdots,m,&\end{aligned} \tag{2.5}$$

with initial conditions

$$\rho_m(0,x)=\rho_{0m}(x), \tag{2.6}$$

$$A_{mk}(0)=\int_{\Omega}u_0(x)\cdot\phi_k(x)\,dx,\qquad k=1,\cdots,m, \tag{2.7}$$

where $u_0(x)$ is given in Theorem 2.1 and

$$b_{jk}^m(t)=\int_\Omega \rho_m(t,x)\phi_k(x)\cdot\phi_j(x)\,dx, \tag{2.8}$$

$$C_{jkl}^m(t)=\int_\Omega \rho_m(t,x)\{(\phi_k\cdot\nabla)\phi_l\}(x)\cdot\phi_j(x)\,dx, \tag{2.9}$$

$$d_j^m(t)=\int_\Omega \rho_m(t,x)f_m(t,x)\cdot\phi_j(x)\,dx. \tag{2.10}$$

Now we find solutions of (2.4) to (2.7).

PROPOSITION 2.4. *There is a number* $T\in(0,T^*]$ *independent of m such that there are solutions* $\rho_m(t,x)\in C^1([0,T]\times\bar\Omega)$, $A_{mk}(t)\in C^1([0,T])$, $k=1,\cdots,m$ *of* (2.4) *to* (2.7) *satisfying*

$$\left\|\sqrt{\rho_m}\,\frac{\partial U_m}{\partial t}\right\|_{L^2(0,T;\,L^2(\Omega)^3)}\leqq K, \tag{2.11}$$

$$\|U_m\|_{L^2(0,T;\,H^2(\Omega)^3)}\leqq K, \tag{2.12}$$

$$\|U_m\|_{L^\infty(0,T;\,H_\sigma^1(\Omega))}\leqq K, \tag{2.13}$$

$$\left\|\frac{\partial\rho_m}{\partial t}\right\|_{L^\infty(0,T;\,H^{-1}(\Omega))}\leqq K, \tag{2.14}$$

$$\frac{1}{m}\leqq\rho_m(t,x)\leqq M+\frac{1}{m}\quad \textit{for all } (t,x)\in[0,T]\times\bar\Omega, \tag{2.15}$$

where K is a positive constant independent of m.

Proof. First we shall derive a priori estimates. Suppose $\rho_m(t,x)\in C^1([0,T]\times\bar\Omega)$ and $1/m\leqq\rho_m(t,x)\leqq M+1/m$ for all $(t,x)\in[0,T]\times\bar\Omega$, where T is a positive number which will be determined later on. Using this $\rho_m(t,x)$ we define b_{jk}^m, C_{jkl}^m, d^m by (2.8) to (2.10). Suppose $A_{mk}(t)\in C^1([0,T])$, $k=1,\cdots,m$, to be the solution of (2.5). Borrowing a technique from Beirao da Veiga [2], we multiply (2.5) by $(d/dt)A_{mj}(t)$ and $\varepsilon\lambda_jA_{mj}(t)$, $\varepsilon>0$ and sum over $j=1,\cdots,m$:

$$\int_\Omega\rho_m\left|\frac{\partial U_m}{\partial t}\right|^2dx+\int_\Omega\rho_m\{(U_m\cdot\nabla)U_m\}\cdot\frac{\partial U_m}{\partial t}\,dx+\frac{1}{2}\frac{d}{dt}\int_\Omega|\nabla U_m|^2dx$$
$$=\int_\Omega\rho_mf_m\cdot\frac{\partial U_m}{\partial t}\,Dx, \tag{2.16}$$

$$-\varepsilon\int_\Omega\rho_m\frac{\partial U_m}{\partial t}\cdot P\Delta U_m dx-\varepsilon\int_\Omega\rho_m((U_m\cdot\nabla)U_m)\cdot P\Delta U_m dx+\varepsilon\int_\Omega|P\Delta U_m|^2dx$$
$$=-\varepsilon\int_\Omega\rho_mf_m\cdot P\Delta U_m dx. \tag{2.17}$$

We observe that

$$\left|\int_\Omega\rho_m\frac{\partial U_m}{\partial t}\cdot P\Delta U_m dx\right|\leqq\frac{1}{4}\int_\Omega|P\Delta U_m|^2dx+(M+1)\int_\Omega\rho_m\left|\frac{\partial U_m}{\partial t}\right|^2dx, \tag{2.18}$$

(2.19)

$$\left|\int_\Omega \rho_m\{(U_m\cdot\nabla)U_m\}\cdot P\Delta U_m dx\right| \leqq \frac{1}{4}\int_\Omega |P\Delta U_m|^2 dx + (M+1)^2\int_\Omega |U_m|^2|\nabla U_m|^2 dx,$$

(2.20)

$$\left|\int_\Omega \rho_m\{(U_m\cdot\nabla)U_m\}\cdot\frac{\partial U_m}{\partial t} dx\right| \leqq \frac{1}{4}\int_\Omega \rho_m\left|\frac{\partial U_m}{\partial t}\right|^2 dx + (M+1)\int_\Omega |U_m|^2|\nabla U_m|^2 dx,$$

$$\begin{aligned}(2.21)\qquad \int_\Omega |U_m|^2|\nabla U_m|^2 dx &\leqq 3\|U_m\|^2_{L^\infty(\Omega)}\int_\Omega |\nabla U_m|^2 dx\\ &\leqq C_\delta\|\nabla U_m\|^{3-2\delta}_{L^2(\Omega)}\|\Delta U_m\|^{1+2\delta}_{L^2(\Omega)}\\ &\leqq C_\delta\|\nabla U_m\|^{3-2\delta}_{L^2(\Omega)}\|P\Delta U_m\|^{1+2\delta}_{L^2(\Omega)},\end{aligned}$$

where C_δ denotes positive constants depending only on Ω and $0<\delta<\frac{1}{2}$. Here we have used Theorem 2 of Ladyzhenskaya [5, p. 67] and Lemmas 1.5, 1.6. Combining (2.16) to (2.21), we obtain

(2.22)

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |\nabla U_m|^2 dx + \frac{1}{2}\int_\Omega \rho_m\left|\frac{\partial U_m}{\partial t}\right|^2 dx \leqq \int_\Omega \rho_m|f_m|^2 dx + C_\delta(M+1)\|\nabla U_m\|^{3-2\delta}_{L^2(\Omega)}\|P\Delta U_m\|^{1+2\delta}_{L^2(\Omega)},$$

$$(2.23)\quad \frac{1}{4}\varepsilon\int_\Omega |P\Delta U_m|^2 dx \leqq \varepsilon\int_\Omega \rho_m^2|f_m|^2 dx + \varepsilon(M+1)\int_\Omega \rho_m\left|\frac{\partial U_m}{\partial t}\right|^2 dx + \varepsilon(M+1)^2C_\delta\|\nabla U_m\|^{3-2\delta}_{L^2(\Omega)}\|P\Delta U_m\|^{1+2\delta}_{L^2(\Omega)},$$

where C_δ depends only on Ω and $0<\delta<\frac{1}{2}$. Now we fix $\varepsilon\in(0, 1/(4(M+1))]$ and take $\delta=\frac{1}{4}$. Then, (2.22) together with (2.23) implies

$$(2.24)\quad \frac{1}{2}\frac{d}{dt}\int_\Omega |\nabla U_m|^2 dx + \frac{1}{4}\int_\Omega \rho_m\left|\frac{\partial U_m}{\partial t}\right|^2 dx + \frac{1}{8}\varepsilon\int_\Omega |P\Delta U_m|^2 dx \leqq C\int_\Omega |f_m|^2 dx + C\|\nabla U_m\|^{10}_{L^2(\Omega)},$$

where C denotes positive constants depending only on M and Ω. From (2.24), it follows that

$$(2.25)\qquad \frac{d}{dt}\int_\Omega |\nabla U_m|^2 dx \leqq C + C\left(\int_\Omega |\nabla U_m|^2 dx\right)^5,$$

where C denotes positive constants depending only on M, Ω and $\|f\|_{L^\infty(0,T^*;\,L^2(\Omega)^3)}$. By making use of the well-known differential inequality, we conclude that there is a number $T\in(0,T^*]$ and $L>0$ such that

$$(2.26)\qquad \|\nabla U_m\|^2_{L^2(\Omega)} \leqq L \quad \text{for all } t\in[0,T],$$

where T and L depend only on M, Ω, $\|f\|_{L^\infty(0,T^*;L^2(\Omega)^3)}$, $\|u_0(x)\|_{H^1_\sigma(\Omega)}$ and are independent of m. From now on, we fix this T and L. Next we need the following fact.

LEMMA 2.5. *Suppose* $\rho_m(t,x) \in C^1([0,T]\times\overline{\Omega})$ *and* $1/m \leqq \rho_m(t,x) \leqq M+1/m$, *for all* $(t,x)\in[0,T]\times\overline{\Omega}$. *Then, the matrix* $\{b^m_{jk}(t)\}$ *defined by* (2.8) *is nonsingular and each component of its inverse belongs to* $C^1([0,T])$.

Proof. Suppose $\{b^m_{jk}(t_0)\}$ is singular for some $t_0 \in [0,T]$. Then, without loss of generality, we assume that the first row is a linear combination of other rows: $b^m_{1k}(t_0) = C_2 b^m_{2k}(t_0) + \cdots + C_m b^m_{mk}(t_0)$ holds for each $k=1,\cdots,m$, i.e., $\int_\Omega \rho_m(t_0,x)\eta(x)\cdot\phi_k(x)\,dx = 0$ holds for each $k=1,\cdots,m$, where $\eta=\phi_1 - C_2\phi_2 - \cdots - C_m\phi_m$. Therefore, $\int_\Omega \rho_m(t_0,x)|\eta(x)|^2\,dx=0$ holds, which implies $\eta(x)=0$ for all $x\in\Omega$ since $\rho_m(t_0,x)\geqq 1/m$ for all $x\in\Omega$. But this is impossible in view of (1.4). It is obvious that each component of the inverse of $\{b^m_{jk}(t)\}$ is continuously differentiable with respect to the time.

Now we proceed to prove the proposition. Let $\overline{B}_R$ be a closed ball in $C([0,T])^m$ with radius $R \geqq (L/\lambda_1)^{1/2}$, where T and L were fixed above and $-\lambda_1$ is the first eigenvalue of the operator $P\Delta$; see Lemma 1.2. Let $(A_{m1}(t),\cdots,A_{mm}(t))\in\overline{B}_R$ and set $U_m(t,x)=\sum_{k=1}^m A_{mk}(t)\phi_k(x)$. By Lemma 2.2, we find a solution $\rho_m(t,x)$ of (2.4), (2.6) in $C^1([0,T]\times\overline{\Omega})$. Using this $\rho_m(t,x)$, we find a solution $(\tilde{A}_{m1}(t),\cdots,\tilde{A}_{mm}(t))$ of (2.5), (2.7) in $C^1([0,T])^m \cap \overline{B}_R$ with the aid of (2.26) and Lemma 2.5. By means of (2.24), (2.26) and Lemma 2.3, we infer that the mapping $(A_{m1}(t),\cdots,A_{mm}(t)) \to (\tilde{A}_{m1}(t),\cdots,\tilde{A}_{mm}(t))$ is completely continuous from $\overline{B}_R$ into itself; see Hale [3]. Hence, it has a fixed point which, together with $\rho_m(t,x)$, is a solution of (2.4) to (2.7). (2.15) was shown in the proof of Lemma 2.2 and (2.13) follows from (2.26), which, combined with (2.24), also implies (2.11) and (2.12). Since U_m is divergence-free, (2.14) follows from (2.4).

Proof of Theorem 2.1. Thanks to the estimates (2.11) to (2.15), we can extract subsequences $\{U_m\}_{m=1}^\infty$ and $\{\rho_m\}_{m=1}^\infty$ such that $U_m \to u$ weakly in $L^2(0,T;H^2(\Omega)^3)$, $U_m\to u$ weak* in $L^\infty(0,T;H^1_\sigma(\Omega))$, $\rho_m\to\rho$ weak* in $L^\infty([0,T]\times\Omega)$ and $\partial\rho_m/\partial t\to\partial\rho/\partial t$ weak* in $L^\infty(0,T;H^{-1}(\Omega))$. By virtue of Lemmas 1.3, 1.4, $\rho_m\to\rho$ strongly in $L^2(0,T;H^{-1/2}(\Omega))$. Therefore, $\rho_m U_m \to \rho u$ in $\mathscr{D}^*((0,T)\times\Omega)^3$ and thus, $\rho_m U_m\to\rho u$ weak* in $L^\infty(0,T;L^2(\Omega)^3)$. In the meantime, $\{U_m\}_{m=1}^\infty$ is bounded in $L^q(0,T;H^{3/2+\delta}(\Omega)^3)$, $q=4/(1+2\delta)$, $0<\delta<\frac{1}{2}$, by interpolation. Hence, by Lemma 1.6, $\{U_m(\partial\rho_m/\partial t)\}_{m=1}^\infty$ is bounded in $L^2(0,T;H^{-1}(\Omega)^3)$. On the other hand, $\{\rho_m(\partial U_m/\partial t)\}_{m=1}^\infty$ is bounded in $L^2(0,T;L^2(\Omega)^3)$, which is obvious from (2.11). Consequently, $\{\partial(\rho_m U_m)/\partial t\}_{m=1}^\infty$ is bounded in $L^2(0,T;H^{-1}(\Omega)^3)$. Again by Lemmas 1.3, 1.4, $\rho_m U_m \to \rho u$ strongly in $L^2(0,T;H^{-1/2}(\Omega)^3)$. Since $U_m\to u$ weakly in $L^2(0,T;H^2(\Omega)^3)$, $\rho_m U_m U_{mk}\to\rho u v_k$ in $\mathscr{D}^*((0,T)\times\Omega)^3$ for each $k=1,2,3$ where $U_m=(U_{m1},U_{m2},U_{m3})$ and $u=(v_1,v_2,v_3)$. Thus it follows that $\rho_m U_m U_{mk}\to\rho u v_k$ weakly in $L^2(0,T;L^2(\Omega)^3)$, for each $k=1,2,3$. It is easy to see that $(U_m\cdot\nabla)\rho_m\to(u\cdot\nabla)\rho$ weakly in $L^2(0,T;H^{-1}(\Omega))$. Now let us choose arbitrary $\xi_j(t)\in C^1([0,T])$, $\xi_j(T)=0$, $j=1,\cdots,\nu$. Then, by (2.4) and (2.5), it holds that for each $m\geqq\nu$,

$$(2.27)\quad -\int_0^T\int_\Omega \rho_m U_m\cdot\frac{\partial}{\partial t}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt - \sum_{k=1}^{3}\int_0^T\int_\Omega \rho_m U_m U_{mk}\cdot\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt$$
$$+\sum_{k=1}^{3}\int_0^T\int_\Omega \frac{\partial U_m}{\partial x_k}\cdot\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt$$
$$=\int_\Omega \rho_{0m}U_{0m}\cdot\sum_{j=1}^{\nu}\xi_j(0)\phi_j\,dx + \int_0^T\int_\Omega \rho_m f_m\cdot\sum_{j=1}^{\nu}\xi_j\phi_j\,dx\,dt$$

where $U_m=(U_{m1},U_{m2},U_{m3})$ and $U_{0m}=\Sigma_{k=1}^{m}A_{mk}(0)\phi_k(x)$; see (2.6) and (2.7). By passing m to ∞, we obtain

$$\begin{aligned}(2.28)\qquad -\int_0^T\int_\Omega \rho u\cdot\frac{\partial}{\partial t}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt-\sum_{k=1}^{3}\int_0^T\int_\Omega \rho u v_k\cdot\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt\\ +\sum_{k=1}^{3}\int_0^T\int_\Omega\frac{\partial u}{\partial x_k}\cdot\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{\nu}\xi_j\phi_j\right)dx\,dt\\ =\int_\Omega\rho_0u_0\cdot\sum_{j=1}^{\nu}\xi_j(0)\phi_j\,dx+\int_0^T\int_\Omega\rho f\cdot\sum_{j=1}^{\nu}\xi_j\phi_j\,dx\,dt,\end{aligned}$$

where we used the convergence properties of $\{U_m\}_{m=1}^{\infty}$ and $\{\rho_m\}_{m=1}^{\infty}$ which were derived above; also recall the way we chose $\{\rho_{0m}\}_{m=1}^{\infty}$ and $\{f_m\}_{m=1}^{\infty}$. Noting that $\rho\in L^\infty([0,T]\times\Omega)$, $u\in L^\infty(0,T;H_\sigma^1(\Omega))\cap L^2(0,T;H^2(\Omega))$ and that every $\Psi(t,x)\in C^1([0,T];H_\sigma^1(\Omega))$, $\Psi(T,x)=0$, a.e. in Ω, and be approximated by a sequence $\{\Sigma_{k=1}^{\nu}\xi_{\nu k}\phi_k\colon\ \xi_{\nu k}(t)\in C^1([0,T]),\ \xi_{\nu k}(T)=0\}_{\nu=1}^{\infty}$ in $C^1([0,T];H_\sigma^1(\Omega))$, we conclude that u and ρ satisfy (1.1). By a similar argument, ρ satisfies (1.2). This completes the proof of Theorem 2.1.

REFERENCES

[1] J. P. Aubin, *Un théorème de compacité*, C. R. Acad. Sci., 256 (1963), pp. 5042–5044.

[2] H. Beirao da Veiga, *Diffusion on viscous fluids*, Ann. Sc. Norm. Sup. Pisa, 10 (1983), pp. 341–355.

[3] J. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York–London–Sydney, 1969.

[4] A. V. Kazhikov, *Solvability of the initial and boundary value problem for the equations of motion of an inhomogeneous viscous incompressible fluid*, Soviet. Phys. Dokl., 19, No. 6 (17), pp. 331–332.

[5] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York–London–Paris, 1969.

[6] J. L. Lions, *On some problems connected with Navier–Stokes equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York–San Francisco–London, 1978.

[7] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York–Heidelberg–Berlin, 1972.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

008

ETUDE DES ETATS STATIONNAIRES POUR UNE EQUATION DE SCHRÖDINGER NON LINEAIRE COMPORTANT UN TERME NON AUTONOME*

ALAIN BAMBERGER† AND LAURENCE HALPERN†

Abstract. This paper deals with the steady problem for a nonlinear one-dimensional Schrödinger equation of the following type:

$$i\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}+xu-p|u|^2u=f(x), \qquad x\in\mathbb{R},$$

arising in plasma physics.

We prove the existence of many steady states for large values of the nonlinearity parameter p.

Résumé. Nous étudions le problème stationnaire associé à une équation de Schrödinger non linéaire monodimensionnelle du type suivant:

$$i\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}+xu-p|u|^2u=f(x), \qquad x\in\mathbb{R},$$

intervenant en physique des plasmas.

Nous montrons l'existence d'un grand nombre d'états stationnaires lorsque le paramètre p de non linéarité est grand.

Key words. nonlinear Schrödinger equation

AMS(MOS) subject classification. Primary 35

Introduction. L'étude de l'absorption d'une onde électromagnétique par un plasma inhomogène conduit, avec nombre d'hypothèses simplificatrices, à une équation de Schrödinger monodimensionnelle non linéaire du type suivant:

$$(1) \qquad i\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}+xu-p|u|^2u=f(x); \qquad x\in\mathbb{R}$$

où u est une "onde sortante" en $-\infty$, u tend vers 0 en $+\infty$. La fonction f représente l'amplitude de l'onde magnétique incidente. Son support est "petit", concentré autour de 0 (voir J. C. Adam, A. Gourdin-Servenière et G. Laval [2], G. J. Morales et Y. C. Lee [8]).

Les auteurs de [2] ont testé numériquement la validité de ce modèle. Ils se sont intéressés aux solutions stationnaires, et ont notamment mis en évidence le fait que celles-ci n'étaient pas stables pour de grandes valeurs du paramètre p de non linéarité.

D'autres équations de Schrödinger non linéaires, du type

$$(2) \qquad i\frac{\partial u}{\partial t}+\Delta u+F(u)=0 \quad \text{dans } \mathbb{R}\times\mathbb{R}^n$$

ont été étudiées par les mathématiciens, en particulier pour une non linéarité de la forme $F(u)=|u|^{k-1}u$.

Ainsi J. Ginibre et G. Velo [6] ont démontré l'existence et l'unicité d'une solution locale pour $1<k<(n+2)/(n-2)$ $(n\geqq 3)$. Ils ont montré de plus que, pour $k<1+4/n$,

*Received by the editors September 24, 1984; accepted for publication August 18, 1985.

†Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau Cedex, France.

toute solution locale est globale, et pour $k<1+4/n$, R. T. Glassey [7] a prouvé qu'il existe des solutions qui explosent en un temps fini. D'autre part V. E. Zakharov et A. B. Shabat [9] ont mis en évidence l'existence de solitons en dimension 1.

Une étude globale de l'existence et du nombre de solitons pour l'équation (2) a été menée par H. Berestycki et P. L. Lions [4]. Ils ont établi l'existence d'une infinité de solutions radiales pour le problème

$$-\Delta u=g(u) \tag{3}$$

où g vérifie certaines hypothèses en 0 et $+\infty$.

En particulier en dimension 1, ils ont prouvé l'existence d'une "solution fondamentale" indéfiniment dérivable, paire, positive, décroissante pour $x>0$, tendant vers zéro à l'infini. Toutes les autres solutions sont obtenues par translation et symétrie par rapport à l'origine. T. Cazenave [5] a étudié la stabilité de ces solitons.

Pour notre part, nous nous intéressons aux solutions stationnaires de l'équation (1). Elle comporte un second membre et un terme "non autonome" xu. Aussi, les études précédentes qui utilisent de façon essentielle l'existence d'une intégrale première, ne semblent pas pouvoir s'appliquer dans cette étude.

Pour connaître l'influence du paramètre p de nonlinéarité sur le nombre de solutions de (1), nous avons introduit un problème simplifié:

Puisque le support de f est petit, nous l'avons concentré autour de 0, et nous avons donc remplacé le second membre par une discontinuité de la partie réelle de la dérivée en 0.

La condition d'onde sortante en $-\infty$ nous a conduits à supposer que le terme non linéaire est petit devant le terme xu pour x négatif. Nous avons donc fait intervenir la non-linéarité uniquement pour $x>0$. Pour ce problème nous avons montré qu'il y a toujours une solution stationnaire, et que le nombre de solutions stationnaires est croissant en fonction de p.

Il ne nous semble pas que ces simplifications changent fondamentalement les résultats.

Les problèmes envisagés et la démarche suivie. Les simplifications énoncées plus haut nous mènent aux deux problèmes suivants, que nous notons $\mathscr{P}_{\alpha,p}$ et $\mathscr{Q}_{\alpha,p}$:

Problème $\mathscr{P}_{\alpha,p}$: Trouver une fonction u qui est, pour $x<0$, solution du problème $\mathscr{P}_c^-$, et, pour $x>0$, solution du problème $\mathscr{P}_c^+$:

$$\mathscr{P}_c^-\begin{cases}(1.1) & u''+u=0;\ x<0,\\ (1.3) & u'+iu\to 0;\ x\to-\infty,\end{cases}\qquad \mathscr{P}_c^+\begin{cases}(1.2) & u''-u+p|u|^2u=0;\ x>0,\\ (1.4) & u\to 0;\ x\to+\infty,\end{cases}$$

et qui vérifie les conditions de transmission à l'origine:

$$u(0_+)-u(0_-)=0, \tag{1.5}$$
$$u'(0_+)-u'(0_-)=\alpha, \qquad \alpha>0. \tag{1.6}$$

Nous avons indexé par C les problèmes $\mathscr{P}_c^-$ et $\mathscr{P}_c^+$ pour signifier que nous cherchons des solutions à valeurs dans $\mathbb{C}$.

Problème $\mathscr{Q}_{\alpha,p}$: Trouver une fonction u qui est, pour $x<0$, solution du problème $\mathscr{Q}_c^-$, et, pour $x>0$, solution du problème $\mathscr{Q}_c^+$:

$$\mathscr{Q}_c^-\begin{cases}(2.1) & u''-xu=0;\ x<0,\\ (2.3) & u'+i(-x)^{1/2}u\to 0;\ x\to-\infty,\end{cases}\qquad \mathscr{Q}_c^+\begin{cases}(2.2) & u''-xu+p|u|^2u=0;\ x>0,\\ (2.4) & xu\to 0;\ x\to+\infty,\end{cases}$$

et qui vérifie les conditions de transmission à l'origine:

(2.5) $$u(0_+)-u(0_-)=0,$$

(2.6) $$u'(0_+)-u'(0_-)=\alpha, \qquad \alpha>0.$$

Rappelons que $u(0_+)$ (resp. $u(0_-)$) désigne la limite à droite (resp. à gauche) de u en 0.

Pour ces deux problèmes, nous avons imposé les mêmes conditions en zéro: continuité de la fonction et discontinuité de la partie réelle de sa dérivée. Les conditions (1.3) et (2.3) correspondent à une condition d'onde sortante en $-\infty$, les conditions (1.4) et (2.4) correspondent à une décroissance vers 0 de la fonction lorsque x tend vers $+\infty$.

Le problème $\mathscr{P}_{\alpha,p}$ est plus simple que le problème $\mathscr{Q}_{\alpha,p}$: le terme xu dans $\mathscr{Q}_{\alpha,p}$ a été remplacé par signe$(x)u$ dans $\mathscr{P}_{\alpha,p}$. Il nous a permis de mettre en évidence l'influence de la nonlinéarité, qui se traduit par l'existence de plusieurs solutions. Les résultats obtenus pour $\mathscr{P}_{\alpha,p}$ nous ont servi de guide pour l'étude de $\mathscr{Q}_{\alpha,p}$.

Nous avons adopté la démarche suivante: nous découplons les problèmes posés sur $\mathbb{R}_-$ et sur $\mathbb{R}_+$, et nous déterminons une relation entre $u(0_-)$ et $u'(0_-)$ (resp. $u(0_+)$ et $u'(0_+)$), condition nécessaire et suffisante pour qu'une solution u de l'équation differentielle posée sur $\mathbb{R}_-$ (resp. $\mathbb{R}_+$) vérifie les conditions à l'infini imposées. La première relation est linéaire: elle est obtenue par la résolution explicite d'une équation differentielle linéaire. La deuxième relation est non linéaire; elle provient d'un calcul explicite pour $\mathscr{P}_{\alpha,p}$, et sera calculée numériquement pour $\mathscr{Q}_{\alpha,p}$.

Ces deux relations, jointes aux conditions de transmission, constituent un système non linéaire complexe de quatre équations à quatre inconnues $u(0_-)$, $u'(0_-)$, $u(0_+)$, $u'(0_+)$. La forme du problème permet de découpler les parties réelles et imaginaires de u pour x positif, et de ramener en fait la discussion sur le nombre de solutions de $\mathscr{P}_{\alpha,p}$ (resp. $\mathscr{Q}_{\alpha,p}$) à la résolution d'un système non linéaire réel de deux équations à deux inconnues: la partie réelle de u et sa dérivée en 0_+. L'une des deux équations dépend du paramètre α^2p.

Nous représentons dans le plan $(u(0_+), u'(0_+))$ les deux équations, pour chacun des deux problèmes.

Pour le problème $\mathscr{P}_{\alpha,p}$, l'une des équations est représentée par une famille de cercles dont de rayon est croissant en fonction de p. Il apparait un seuil: si p est inférieur à ce seuil, $\mathscr{P}_{\alpha,p}$ admet deux solutions, et audelà $\mathscr{P}_{\alpha,p}$ n'a plus de solution. Ce phénomène de seuil est du à la simplification apportée en remplaçant xu par signe$(x)u$ dans $\mathscr{P}_{\alpha,p}$.

Pour le problème $\mathscr{Q}_{\alpha,p}$, l'une des relations est obtenue à partir du problème posé sur $\mathbb{R}_+$, et est représentée par une ellipse. L'autre relation est obtenue numériquement à partir d'un paramètre x_0, et est représentée par une spirale.

Lorsque le paramètre α^2p varie, la première relation entre $u(0_+)$ et $u'(0_+)$ est représentée par une famille d'ellipses, dont les axes croissent en fonction de α^2p: il y a toujours une solution, et plus la non-linéairité est importante, plus $\mathscr{Q}_{\alpha,p}$ admet de solutions.

La première partie de cette étude est consacrée au problème $\mathscr{P}_{\alpha,p}$. Des méthodes d'équations différentielles ordinaires permettent de décrire explicitement les solutions.

Ces méthodes ne sont plus applicables pour le problème $\mathscr{Q}_{\alpha,p}$ et nous serons amenés, dans la deuxième partie, à utiliser à la fois des méthodes mathématiques et des méthodes numériques.

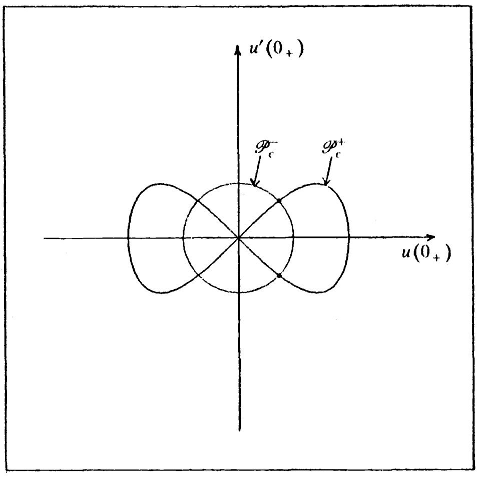


FIG. 1. *Nombre de solutions de $\mathscr{P}_{\alpha,p}$, α et p fixés.*

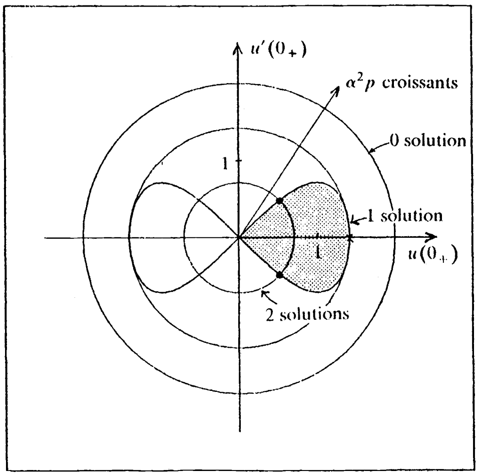


FIG. 2. *Nombre de solutions de $\mathscr{P}_{\alpha,p}$ lorsque α^2p varie.*

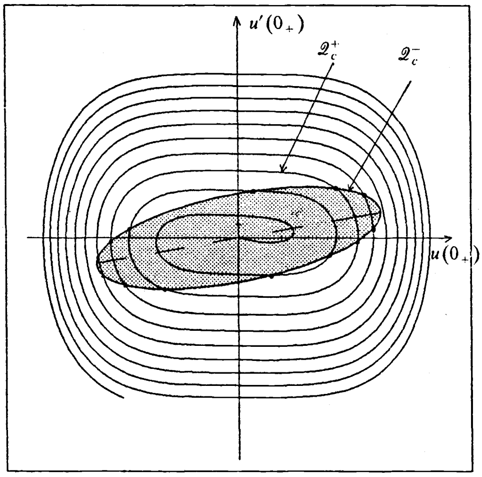


FIG. 3. *Nombre de solutions de $\mathscr{Q}_{\alpha,p}$, α et p fixés.*

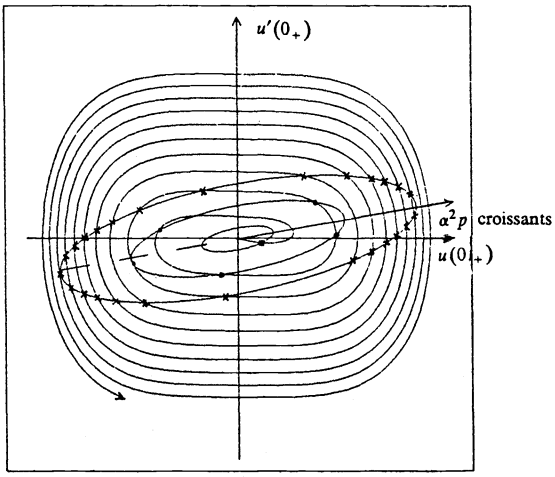


FIG. 4. *Nombre de solutions de $\mathscr{Q}_{\alpha,p}$ lorsque α^2p varie.*

1. Etude du problème $\mathscr{P}_{\alpha,p}$. Nous pouvons pour le problème $\mathscr{P}_{\alpha,p}$ décrire complètement et explicitement l'ensemble des solutions:

PROPOSITION 1.1. *Le problème $\mathscr{P}_{\alpha,p}$ admet, suivant les valeurs de p et α, zéro, une ou deux solutions*:

• $p=0$	*une solution*,
• $0<p<p^*=2/\alpha^2$	*deux solutions*,
• $p=p^*$	*une solution*,
• $p>p^*$	*pas de solution*.

De plus, ces solutions sont C^∞ sauf en $x=0$.

Le cas $p=0$ correspond à une équation linéaire. Le problème $\mathscr{P}_{\alpha,0}$ admet une seule solution, $u^*(0,x)$, donnée par

$$u^*(0,x)=\begin{cases} -\frac{\alpha}{2}(1+i)e^{-ix}, & x\leqq 0,\\ -\frac{\alpha}{2}(1+i)e^{-x}, & x\geqq 0.\end{cases}$$

Nous nous intéressons d'abord à $\mathscr{P}_c^+$, dont nous exprimerons les solutions à partir de celles d'un problème réel $\mathscr{P}_R^+$. Nous étudierons les orbites de l'équation différentielle associée dans le plan des phases. Nous montrerons en particulier qu'il existe une seule orbite non périodique, qui représente l'ensemble des solutions de $\mathscr{P}_R^+$, et nous expliciterons ces solutions.

1.1. Le problème $\mathscr{P}_c^+$. Introduisons le problème réel

$$(1.7)\qquad \mathscr{P}_R^+\begin{cases} -v''+v-v^3=0, & x\geqq 0,\\ v(x)\to 0, & x\to+\infty.\end{cases}$$

Pour passer de $\mathscr{P}_c^+$ à $\mathscr{P}_R^+$, nous aurons besoin d'estimations à priori. Pour l'équation (1.2) on obtient de façon classique la conservation du moment et de l'énergie. On a plus précisément ici le

LEMME 1.1. Estimations à priori. *Toute solution u de $\mathscr{P}_c^+$ appartenant à $\mathscr{C}^1(\mathbb{R}^+,\mathbb{R})$ possède un moment nul*:

$$(1.8)\qquad u'\bar{u}-\bar{u}'u=0,$$

ainsi qu'une énergie nulle:

$$(1.9)\qquad |u'|^2-|u|^2+\frac{p}{2}|u|^4=0.$$

Elle vérifie de plus

$$(1.10)\qquad u'(x)\to 0;\qquad x\to+\infty.$$

Démonstration. Nous multiplions (1.2) par $\bar{u}'$ et nous prenons la partie réelle

$$\frac{d}{dx}\left(|u'|^2-|u|^2+\frac{p}{2}|u|^4\right)=0,\quad \text{d'où}$$

$$(1.11)\qquad |u'|^2-|u|^2+\frac{p}{2}|u|^4=\text{cte}\quad \text{sur } \mathbb{R}_+.$$

D'autre part, nous multiplions (1.2) par $\bar{u}$ et nous prenons la partie imaginaire

$$\frac{d}{dx}(u'\bar{u}-\bar{u}'u)=0, \quad \text{d'où}$$

$$u'\bar{u}-\bar{u}'u=\text{cte} \quad \text{sur } \mathbb{R}_+. \tag{1.12}$$

Par hypothèse u tend vers 0 à l'infini, et on déduit de (1.11) que u' est bornée. La constante dans (1.12) est donc nulle. Montrons maintenant que u' tend vers 0 à l'infini.

D'après (1.8) la partie réelle et la partie imaginaire de u sont proportionnelles, nous pouvons donc raisonner sur $v=\operatorname{Re} u$.

La fonction réelle v est $\mathscr{C}^1$, tend vers 0 à l'infini, et d'après (1.11) v'^2 a une limite à l'infini. La fonction v' a donc une limite l quand x tend vers l'infini. Cette limite doit être nulle.

Nous en déduisons le

LEMME 1.2. Equivalence entre $\mathscr{P}_c^+$ et $\mathscr{P}_R^+$. *La fonction u est solution de $\mathscr{P}_c^+$ si et seulement si il existe un nombre complexe z_0 et une fonction réelle v tels que*

$$u=\frac{z_0}{\sqrt{p}}v, \qquad |z_0|=1, \tag{1.13}$$

v solution de $\mathscr{P}_R^+$.

Démonstration. Soit u une solution de $\mathscr{P}_c^+$. D'après la démonstration précédente, la partie imaginaire de u est proportionnelle à sa partie réelle

$$u=(1+\lambda i)u_1$$

et u_1 est solution de l'équation réelle

$$u_1''-u_1+p(1+\lambda^2)u_1^3=0.$$

La fonction u s'écrit donc sous la forme

$$u=\frac{1+\lambda i}{\sqrt{p(1+\lambda^2)}}v \tag{1.14}$$

où v est solution de $\mathscr{P}_R^+$.

Réciproquement, si v est solution de $\mathscr{P}_R^+$, la fonction u définie par (1.14) est solution de $\mathscr{P}_c^+$.

Nous définissons sur $\mathbb{R}$ le système différentiel

$$-v''+v-v^3=0, \qquad x\in\mathbb{R}. \tag{1.15}$$

Introduisons la fonction U (assimilable à une énergie potentielle), définie par

$$U(v)=-\frac{v^2}{2}+\frac{v^4}{4} \tag{1.16}$$

ainsi que la fonction E définie sur l'ensemble des fonctions C^1 par

$$E(v)=\frac{v'^2}{2}+U(v); \tag{1.17}$$

E est "l'énergie totale."

Nous avons pour ce système l'équivalent du lemme 1.1, c'est-à-dire que le système (1.15) est *conservatif*:

$$\frac{d}{dx}(E(v))=0. \tag{1.18}$$

Pour tracer les orbites du système, c'est-à-dire $\{(v,v');\ E(v)=E\}$, il est commode de tracer la courbe représentant les variations du potentiel U (voir figure 5).

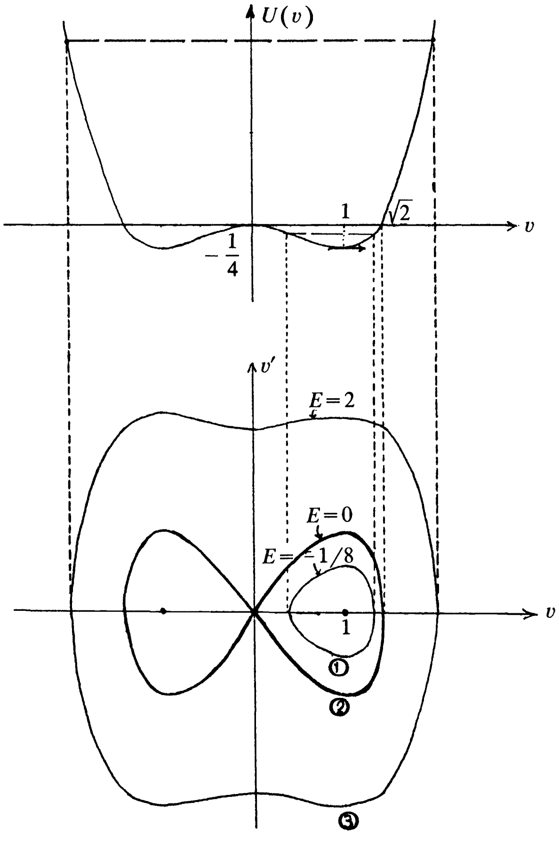


FIG. 5. *Orbites du système* (1.15).

Il y a deux sortes d'orbites périodiques: de type pour une énergie totale négative et de type pour une énergie totale positive. Il y a une orbite non périodique (à une symétrie par rapport à l'origine près), de type, correspondant à une énergie nulle. Un point représentatif situé sur cette orbite atteindra le point zéro sur une distance infinie. L'analyse de ces orbites est résumée dans le lemme 1.3.

LEMME 1.3. Description des solutions du système différentiel. *Le système* (1.15) *admet une infinité de solutions d'énergie E donnée. Elles sont toutes égales, à une translation et une symétrie près, à une fonction* V^E, *paire, telle que*

$$\frac{dV^E}{dx}(0)=0,\quad U(V^E(0))=E,\quad V^E(0)>1. \tag{1.19}$$

Les solutions de (1.15) *d'énergie nulle sont celles qui tendent vers* 0 *en* $+\infty$ *et* $-\infty$.

De plus V^0 est donnée explicitement par

$$V^0(x)=\begin{cases} 2\sqrt{2}\,\dfrac{e^x}{1+e^{2x}}, & x\geqq 0,\\ V^0(-x), & x\leqq 0.\end{cases} \tag{1.20}$$

Remarque. H. Berestycki et P. L. Lions ont montré, dans [3], le même type de résultat pour l'équation

$$-u''=g(u).$$

Leur approche est différente, mais les hypothèses faites sur g permettent d'utiliser la théorie des systèmes conservatifs, et d'assurer l'existence d'une orbite de type (1).

Démonstration du Lemme 1.3. Il ne nous reste qu'à établir l'expression (1.20):

V^0 est solution sur $\mathbb{R}_+$ du problème de Cauchy:

$$\begin{aligned} &-v''+v-v^3=0, \qquad x\geqq 0,\\ &v(0)=\sqrt{2},\\ &v'(0)=0.\end{aligned} \tag{1.21}$$

D'après (1.17) nous avons sur $\mathbb{R}_+$

$$v'^2-v^2+\frac{v^4}{2}=0. \tag{1.22}$$

Posons

$$w=\frac{v'}{v}. \tag{1.23}$$

La fonction w est solution sur $\mathbb{R}_+$ du problème de Cauchy:

$$\begin{aligned} &w'=w^2-1,\\ &w(0)=0.\end{aligned} \tag{1.24}$$

Cette équation de Riccati se résout explicitement sous la forme

$$w=\frac{1-e^{2x}}{1+e^{2x}}. \tag{1.25}$$

Il suffit maintenant d'intégrer (1.23) pour obtenir l'expression de V^0 sur $[0,+\infty[$. V^0 est obtenue alors par symétrisation sur $]-\infty,0]$. □

Nous pouvons maintenant énoncer le théorème d'existence pour $\mathscr{P}_R^+$.

THEOREME 1.1. Condition nécessaire et suffisante d'existence pour $\mathscr{P}_R^+$. *Une fonction v solution de l'équation différentielle* (1.15) *sur* $\mathbb{R}_+$ *est solution du problème* $\mathscr{P}_R^+$ *si et seulement si ses valeurs de Cauchy en* 0: $v_0=v(0)$ *et* $v_0'=v'(0)$ *sont liées par la relation*

$$v_0'^2-v_0^2+\frac{v_0^4}{2}=0. \tag{1.26}$$

Elle est alors donnée explicitement par

$$V(v_0,v_0';\,x)=V^0(x+x_0) \tag{1.27}$$

où x_0 est donné par

$$e^{x_0}=\frac{v_0}{\sqrt{2(1+v_0'/v_0)}}. \tag{1.28}$$ □

A un couple (v_0,v_0') de solutions de (1.26) sont associées quatre solutions de $\mathscr{P}_R^+$: $V(v_0,v_0';\,x)$, $V(v_0,-v_0';\,x)$, $-V(v_0,v_0';\,x)$, et $-V(v_0,-v_0',\,x)$.

Les deux premières sont obtenues à partir de V^0 par translation respectivement de x_0 et $-x_0$. Les deux autres sont ensuite obtenues par symétrie par rapport à l'axe des x. La figure 6 représente les variations de la solution fondamentale V^0 en fonction de x. La figure 7 représente les quatre solutions associées à un couple (v_0,v_0') de nombres positifs solution de (1.26).

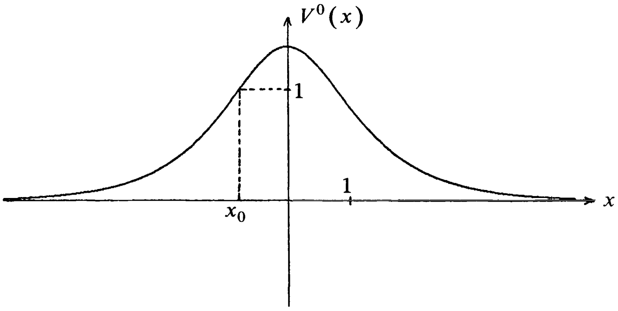


FIG. 6. *Représentation de V^0 sur* $\mathbb{R}$.

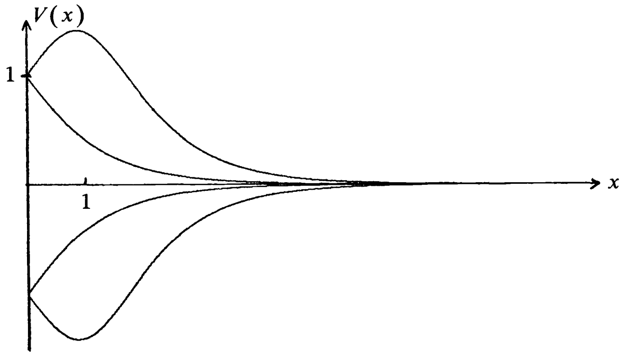


FIG. 7. *Solutions de $\mathscr{P}_R^+$ associées à* $(v_0,v_0')=(1,1/\sqrt{2})$.

Le théorème 1.1 nous donne une relation entre v_0 et v_0', condition nécessaire et suffisante pour que v soit solution de $\mathscr{P}_R^+$. Pour compléter notre étude, nous devons revenir au problème complexe $\mathscr{P}_{\alpha,p}$.

1.2. Le problème $\mathscr{P}_c^-$. L'étude de $\mathscr{P}_c^-$ nous conduira à une relation linéaire entre $u(0_-)$ et $u'(0_-)$. Les conditions de transmission la transformeront en une relation linéaire entre $u(0_+)$ et $u'(0_+)$. Le lemme 1.2 nous permettra d'en déduire une relation entre v_0 et v_0', valeurs de Cauchy de la solution de l'équation différentielle réelle associée à $\mathscr{P}_c^+$. Nous aboutirons ainsi à un système de deux équations réelles à deux inconnues v_0 et v_0'.

1.2.1. Le problème $\mathscr{P}_c^-$. Il peut être résolu explicitement:

LEMME 1.4. *Toutes les solutions u de $\mathscr{P}_c^-$ vérifient*

$$u' + iu = 0 \qquad \forall x \leqq 0, \tag{1.29}$$

$$u = u_0 \exp - ix \qquad \forall x \leqq 0. \tag{1.30}$$

En effet, d'après (1.1), u s'écrit

$$u = A^+ \exp ix + A^- \exp - ix,$$

et la condition d'onde sortante élimine la composante $\exp ix$.

Nous sommes maintenant en mesure d'écrire une caractérisation des solutions de $\mathscr{P}_{\alpha,p}$.

1.2.2. Détermination du système non linéaire reliant v_0 et v_0'. Notons (S) le système non linéaire liant v_0 et v_0':

$$\text{(S)}\begin{cases} v_0'^2 + v_0^2 = \alpha^2 p, \\ v_0'^2 - v_0^2 + \dfrac{v_0^4}{2} = 0, \\ v_0 \geqq 0. \end{cases} \tag{1.31}$$

Nous avons alors le:

THÉORÈME 1.2. Condition nécessaire et suffisante d'existence pour le problème $\mathscr{P}_{\alpha,p}$. *Le problème $\mathscr{P}_{\alpha,p}$ admet une solution u si et seulement si le système* (S) *admet une solution* (v_0, v_0'). *Elle est alors donnée par*:

$$u(x) = \begin{cases} \dfrac{1}{\alpha p}(v_0' - iv_0) V(v_0, v_0'; \, x), & x \geqq 0, \\ \dfrac{1}{\alpha p}(v_0' - iv_0) v_0 \exp - ix; & x \leqq 0. \end{cases} \tag{1.32}$$

Démonstration. Soit u une solution de $\mathscr{P}_{\alpha,p}$. D'après le lemme 1.4, $u'(0_-)$ et $u(0_-)$ sont liés par:

$$u'(0_-) + iu(0_-) = 0.$$

Nous transformons cette relation à l'aide des conditions de transmission:

$$u'(0_+) + iu(0_+) = \alpha. \tag{1.33}$$

D'autre part, d'après le lemme 1.2, u s'écrit pour $x \geqq 0$:

$$u = \frac{z_0}{\sqrt{p}} v$$

où v_0 et v_0' sont liés par

$$v_0'^2 - v_0^2 + \frac{v_0^4}{2} = 0.$$

L'équation (1.33) se traduit alors par une relation entre z_0, v_0 et v_0':

$$\frac{z_0}{\sqrt{p}}\left(v_0'+iv_0\right)=\alpha.$$

Si nous exprimons que z_0 est un nombre complexe de module 1, il vient

$$v_0'^2+v_0^2=\alpha^2 p, \tag{1.34}$$

$$z_0=\frac{1}{\alpha\sqrt{p}}\left(v_0'-iv_0\right). \tag{1.35}$$

Pour x positif, u s'écrit alors

$$u(x)=\frac{1}{\alpha p}\left(v_0'-iv_0\right)V\left(v_0,v_0';\ x\right)$$

où V est la solution du problème réel associé à (v_0,v_0'), et, pour x négatif

$$u(x)=\frac{1}{\alpha p}\left(v_0'-iv_0\right)v_0\exp(-ix).$$

Les considérations de parité sur v développées après le théorème 1.1 permettent enfin de supposer que v_0 est positif.

Réciproquement, si (v_0,v_0') est une solution de (S), la fonction u définie par (1.32) est solution de $\mathscr{P}_{\alpha,p}$.

Pour achever la démonstration de la proposition 1, il nous faut dénombrer les solutions du système (S) de deux équations à deux inconnues. La figure 8 représente les variations des fonctions:

$$Y^2+X^2=\alpha^2 p,$$
$$Y^2-X^2+\frac{X^4}{2}=0.$$

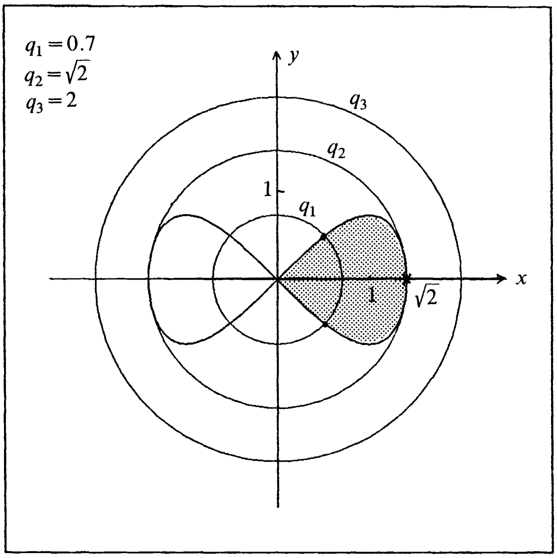


FIG. 8. *Solutions du système* (S) *pour différentes valeurs du paramètre* $q=\sqrt{\alpha^2 p}$.

Les points d'intersection des deux courbes représentatives donnent les solutions du système (S).

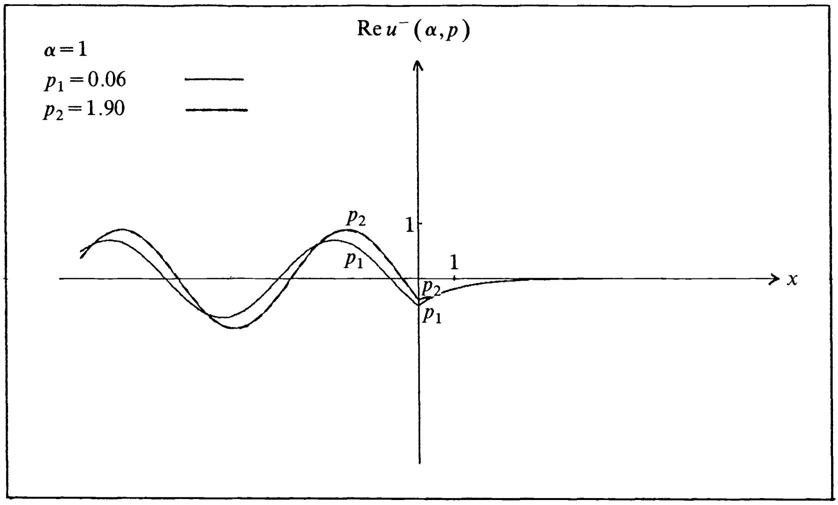
Nous pouvons calculer explicitement les solutions du système (S) en fonction du paramètre $p\alpha^2$.

Notons

$$V_0(p)=\sqrt{2}\left(1-\left(1-\frac{p\alpha^2}{2}\right)^{1/2}\right)^{1/2}, \tag{1.36}$$

$$V_0'(p)=\left(1-\frac{p\alpha^2}{2}\right)^{1/4}V_0(p) \tag{1.37}$$

si $p\alpha^2 \leqq 2$.



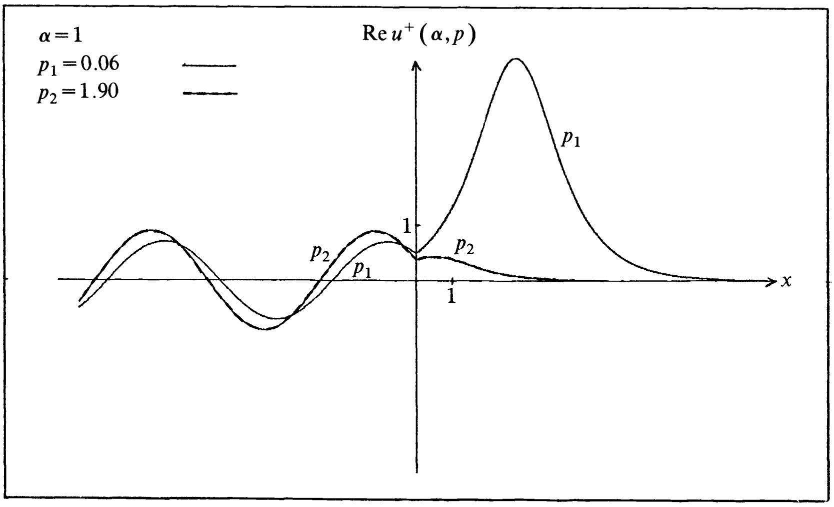


FIG. 9. *Représentation des parties réelles de u^+ et u^-.*

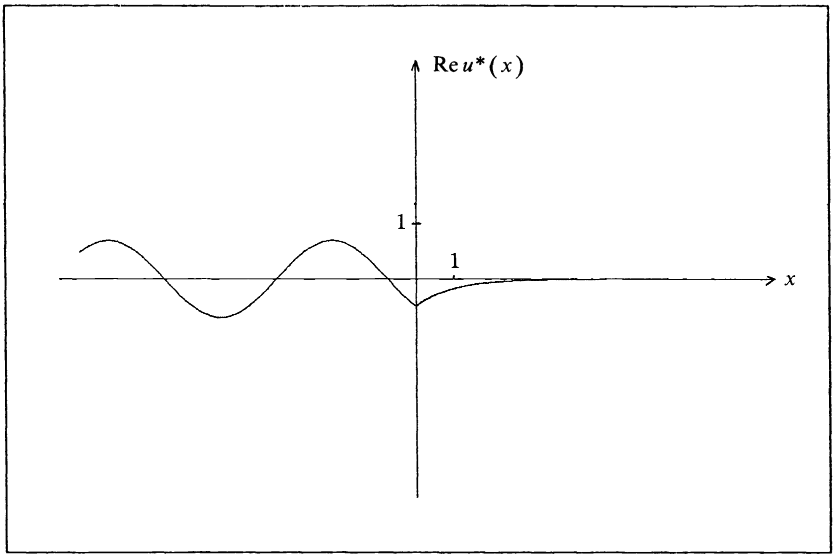


FIG. 10. *Représentation de la partie réelle de u^*, solution du problème linéaire.*

Nous avons alors le:

LEMME 1.5. *Le système* (S) *admet, suivant les valeurs de $p\alpha^2$, zéro, une ou deux solutions*:

$$\begin{array}{ll} 0<p\alpha^2<2, & 2 \text{ solutions} : (V_0(p), V_0'(p)) \text{ et } (V_0(p), -V_0'(p)), \\ p\alpha^2=2, & 1 \text{ solution} : (\sqrt{2}, 0), \\ p\alpha^2>2, & 0 \text{ solution}. \end{array}$$

L'expression des solutions de $\mathscr{P}_{\alpha,p}$ donnée au théorème 1.2 permet de différencier les deux solutions associées au même paramètre $\alpha^2 p$ par le signe de leurs parties réelles.

Nous notons $u^+(\alpha,p)$ la solution de partie réelle positive et $u^-(\alpha,p)$ la solution de partie réelle négative, et nous représentons leurs parties réelles pour $\alpha=1$ et différentes valeurs de p (voir figures 9 et 10).

Remarquons d'abord que la valeur de la fonction pour x négatif varie peu lorsque p est proche de zéro ou de deux.

De même, pour x positif, la partie réelle de u^- dépend peu de p, et lorsque p tend vers 0, $u^-(\alpha,p)$ tend vers la solution u^* du problème linéaire. Par contre, lorsque p devient petit, $u^+(\alpha,p)$ admet un maximum pour x positif qui croit en $1/\sqrt{p}$. La fonction u^- semble donc être la "bonne" solution, et u^+ la solution "parasite."

L'étude de $\mathscr{P}_{\alpha,p}$ fait ainsi apparaître qu'une non-linéarité, même faible, peut modifier totalement la nature de l'équation stationnaire. Nous allons maintenent examiner le problème dérivant directement de (1), en réintroduisant le terme non autonome xu. Nous verrons que l'effet de seuil disparait, mais qu'en revanche le nombre de solutions est croissant avec p.

2. Etude du problème $\mathscr{Q}_{\alpha,p}$. Comme pour le problème $\mathscr{P}_{\alpha,p}$, nous allons écrire un système de deux relations à deux inconnues, reliant les valeurs en zéro d'une fonction v

et de sa dérivée, où v est solution de l'équation réelle associée à $\mathcal{Q}_c^+$:

$$-v'' + xv - v^3 = 0. \tag{2.7}$$

La première relation sera obtenue comme précédemment par l'intégration explicite de $\mathcal{Q}_c^-$ et les conditions de transmission. Par contre la caractérisation des solutions de (2.7) est beaucoup plus complexe, car il nous semble impossible de résoudre explicitement cette équation. Nous ferons intervenir un paramètre x_0, et pour chaque valeur de x_0 nous définissons une fonctionnelle sur $[x_0, +\infty[$. Nous caractérisons ainsi les solutions de (2.7) qui tendent vers zéro à l'infini comme les solutions coincidant sur $[x_0, +\infty[$ avec le minimum de cette fonctionnelle pour un certain x_0. Nous obtiendrons ainsi la deuxième relation entre v_0 et v_0', par l'intermédiaire du paramètre x_0.

2.1. Etude mathématique. Nous commençons par caractériser les solutions du problème linéaire $\mathcal{Q}_c^-$.

2.1.1. Etude du problème $\mathcal{Q}_c^-$. Nous pouvons exprimer la solution u de $\mathcal{Q}_c^-$ à partir de la fonction d'Airy Ai définie dans [1] par:

$$\begin{aligned} &z'' - xz = 0, \qquad x \in \mathbb{R}, \\ &z(0) = 3^{-2/3}/\Gamma(2/3), \\ &z'(0) = -3^{-1/3}/\Gamma(1/3). \end{aligned} \tag{2.8}$$

LEMME 2.1. *Toutes les solutions de $\mathcal{Q}_c^-$ vérifient*

$$\forall v < 0, \quad u(x) = \lambda Ai(xe^{2i\pi/3}), \qquad \lambda \in \mathbb{C}. \tag{2.9}$$

En $x = 0$, u et u' sont liés par la relation

$$u'(0_-) + Ke^{2i\pi/3}u(0_-) = 0 \tag{2.10}$$

où $K = -\mathrm{Ai}'(0)/\mathrm{Ai}(0)$.

Ce lemme est une conséquence immédiate des propriétés de la fonction d'Airy données dans [1].

Nous introduisons maintenant le problème réel:

$$Q_R^+ \begin{cases} -v' + xv - v^3 = 0, & x \in \mathbb{R}_+, \\ xv \to 0, & x \to +\infty, \end{cases} \tag{2.11}$$

et nous cherchons à décrire l'ensemble de ses solutions. Pour celà nous notons que, pour x fixé, l'application:

$$v \to xv - v^3$$

est le gradient d'une fonction convexe sur

$$I_x = \left[-\sqrt{\frac{x}{3}}, +\sqrt{\frac{x}{3}}\right]. \tag{2.12}$$

Pour tout x_0 positif, nous introduisons un problème aux limites sur $]x_0, +\infty[$ tel que, pour tout x, la solution $v(x)$ appartienne à I_x. Nous montrerons que toute solution de $\mathcal{Q}_R^+$ est solution d'un unique problème de ce type, et nous en caractériserons les solutions comme minima d'une fonctionnelle convexe.

Nous la prolongerons ensuite sur $[0, x_0]$ et nous aurons *toutes* les solutions de $\mathcal{Q}_R^+$, paramétrisées par x_0.

2.1.2. Comportement asymptotique des solutions de $\mathscr{Q}_R^+$. Pour tout x_0 positif, nous introduisons le problème

$$(2.13)\qquad \mathscr{Q}_{x_0}\begin{cases} -v''+xv-v^3=0, \\ v(x_0)=\sqrt{\dfrac{x_0}{3}}, \qquad xv(x)\to 0 \quad \text{quand } x\to+\infty, \\ \forall x\geqq x_0, \quad v(x)\in I_x. \end{cases}$$

Avant de décrire ses solutions, nous montrons qu'elles sont les restrictions à $[x_0, +\infty[$ des solutions de $\mathscr{Q}_R^+$ (voir figure 11).

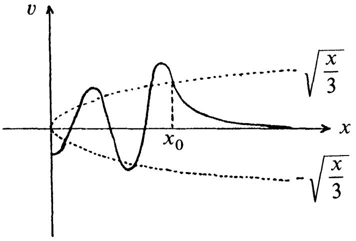


FIG. 11. *Illustration du lemme* 2.2.

LEMME 2.2. *Soit v une solution de $\mathscr{Q}_R^+$. Alors il existe un x_0 positif tel que la restriction de v à $[x_0, +\infty[$ soit solution de $\mathscr{Q}_{x_0}$.*

Démonstration. Comme dans 1, si v est solution de $\mathscr{Q}_R^+$, $v^{(n)}(x)$ tend vers 0 pour tout n lorsque x tend vers $+\infty$. Puisque v est $\mathscr{C}^\infty$, il existe alors un x_0 positif tel que $v^2(x_0)=x_0/3$ et, pour tout x supérieur à x_0, $v^2(x)\leqq x/3$. Alors v est solution de $\mathscr{Q}_{x_0}$.

Le lien entre $\mathscr{Q}_R^+$ et $\mathscr{Q}_{x_0}$ ainsi établi, nous étudions le problème $\mathscr{Q}_{x_0}$, et nous introduisons un problème de minimisation.

Notons:

$$(2.14)\qquad V(x_0)=\{v\in H^1(]x_0,+\infty[),\ \sqrt{x}\,v\in L^2(]x_0,+\infty[)\}.$$

$V(x_0)$ est un espace de Hilbert pour la norme définie par

$$(2.15)\qquad \|v\|^2=|v|_1^2+\|\sqrt{x}v\|_0^2.$$

Soit

$$(2.16)\qquad \mathscr{J}_{x_0}(v)=\frac{1}{2}\int_{x_0}^{+\infty}v'^2(x)\,dx+\frac{1}{2}\int_{x_0}^{+\infty}\left(xv^2(x)-\frac{v^4(x)}{2}\right)dx,$$

une fonctionnelle définie sur $V(x_0)$.

Nous définissons un convexe fermé de $V(x_0)$ par

$$(2.17)\qquad K(x_0)=\left\{v\in V(x_0);\ v(x_0)=\sqrt{\frac{x_0}{3}}\ ;\ v(x)\in I_x, \forall x\geqq x_0\right\}$$

et un problème de minimisation avec contraintes par:

$$(2.18)\qquad \mathscr{M}_{x_0}:\ \inf_{v\in V(x_0)}\mathscr{J}_{x_0}(v).$$

L'introduction de ce problème est justifiée par le

LEMME 2.3. Existence et unicité pour le problème $\mathcal{Q}_{x_0}$. *Le problème* $\mathcal{Q}_{x_0}$ *admet une solution unique* $\mathcal{C}^\infty$. *C'est une fonction positive, décroissante, convexe. De plus c'est la solution du problème de minimisation* $\mathcal{M}_{x_0}$.

Démonstration. La fonctionnelle $\mathcal{J}_{x_0}$ est strictement convexe, différentiable, coercive, sur la convexe fermé $K(x_0)$. Un théorème usuel d'optimisation prouve qu'elle admet un minimum unique sur $K(x_0)$, caractérisé par

$$\forall w \in K(x_0), \quad \mathcal{J}_{x_0}'(v)(w-v) \geqq 0,$$

soit ici

$$(2.19) \qquad \forall w \in K(x_0), \quad \int_{x_0}^{+\infty} (-v''+xv-v^3)(w-v) \geqq 0.$$

D'autre part, $|v|$ est dans $K(x_0)$ et $\mathcal{J}_{x_0}(v)=\mathcal{J}_{x_0}(|v|)$. L'unicité assure alors la positivité de v.

Posons maintenant:

$$(2.20) \qquad \Omega_0 = \left\{ x \in [x_0, +\infty[;\ v(x) = \sqrt{\frac{x}{3}} \right\},$$

$$(2.21) \qquad \Omega_+ = \left\{ x \in]x_0, +\infty[;\ v(x) < \sqrt{\frac{x}{3}} \right\}.$$

Si φ appartient à $\mathcal{D}_+(\,]x_0, +\infty[\,)$, et si ε est un nombre positif assez petit:

$$w = v - \varepsilon\varphi \in K(x_0).$$

Appliquons (2.19) à w:

$$\forall x \in \Omega_+, \quad -v''+xv-v^3=0,$$
$$\forall x \in \Omega_0, \quad -v''+xv-v^3 \leqq 0.$$

Nous en déduisons que v'' est positive, et donc que v' est croissante, sur $[x_0, +\infty[$. Or v' tend vers 0 à l'infini. Donc v' est négative, et v est décroissante sur $]x_0, +\infty[$. Par suite Ω_0 se réduit à x_0, et

$$(2.22) \qquad \forall x \in]x_0, +\infty[, \quad -v''+xv-v^3=0.$$

Ainsi v est dans $H^2(]x_0, +\infty[)$, et donc dans $\mathcal{C}^1(]x_0, +\infty[)$; v est $\mathcal{C}^\infty$ sur $]x_0, +\infty[$ et par continuité (2.22) est valable aussi en x_0. La caractérisation (2.19) montre enfin que toute solution de $\mathcal{Q}_{x_0}$ est solution de $\mathcal{M}_{x_0}$.

Notons maintenant, pour tout x_0:

(2.23) $v(x_0,\gamma;\ x)$ une solution $\mathcal{C}^\infty$ du problème de Cauchy,

$$\begin{cases} -v''+xv-v^3=0 & \text{sur }]x_0, +\infty[, \\ v(x^0) = \sqrt{\dfrac{x_0}{3}}, & v'(x_0)=\gamma, \end{cases}$$

(2.24) $v^*(x_0;\ x)$ la solution du problème de minimisation $\mathcal{M}_{x_0}$,

(2.25) $\gamma^*(x_0)$ sa dérivée en x_0.

Nous pouvons énoncer le

THÉORÈME 2.1. Condition nécessaire d'existence pour le problème $\mathcal{Q}_R^+$. *Soit v une solution du problème $\mathcal{Q}_R^+$. Il existe un x_0 positif tel que*

$$v=v^*(x_0;\ \cdot) \quad ou\ v=-v^*(x_0;\ \cdot)\ sur\ ]x_0,+\infty[.$$

Ce théorème est une conséquence des deux lemmes précédents.

Remarquons que $\gamma^*(x_0)$ n'est pas donné explicitement en fonction de x_0. Par la suite, nous calculerons $u^*(x_0;\ \cdot)$ numériquement, et nous aurons besoin de certaines propriétés de $\gamma^*(x_0)$ que nous établissons maintenant.

2.1.3. Propriétés de $\gamma^*(x_0)$—Comportement asymptotique lorsque x tend vers $+\infty$. Toutes les démonstrations seront basées sur une estimation à priori.

LEMME 2.4. Egalité d'énergie. *Si une fonction v est solution sur $]y,z[$ de l'équation différentielle* (2.7), *on a sur $]y,z[$ la relation*:

$$\frac{d}{dx}\left(v'^2-xv^2+\frac{v^4}{2}\right)=-v^2. \tag{2.26}$$

La démonstration est analogue à celle du lemme 1.1.

Nous établissons maintenant deux propriétés de $v^*(x_0;\ \cdot)$.

LEMME 2.5.

i) *Encadrement de la solution $v^*(x_0;\ \cdot)$ du problème aux limites par des solutions $v(x_0,\gamma;\ \cdot)$ de problèmes de Cauchy.*

●*Si $\gamma_1<\gamma^*(x_0)$ il existe un $x_1>x_0$, tel que $v(x_0,\gamma_1;\ x_1)=0$.*

●*Si $\gamma_1>\gamma^*(x_0)$ il existe un $x_2>x_0$, tel que $v'(x_0,\gamma_2;\ x_2)=0$.*

ii) *Encadrement de $\gamma^*(x_0)$ en fonction de x_0. Pour tout $x_0>0$ on a l'encadrement*

$$-\frac{1}{\sqrt{3}}\left(x_0+\frac{1}{4}x_0^{-1/2}\right)\leqq\gamma^*(x_0)\leqq-\frac{x_0}{3}\sqrt{\frac{5}{2}}. \tag{2.27}$$

En particulier $\gamma^*(x_0)$ tend vers $-\infty$ lorsque x_0 tend vers $+\infty$.

La figure 12 illustre la partie (i) du lemme 2.5.

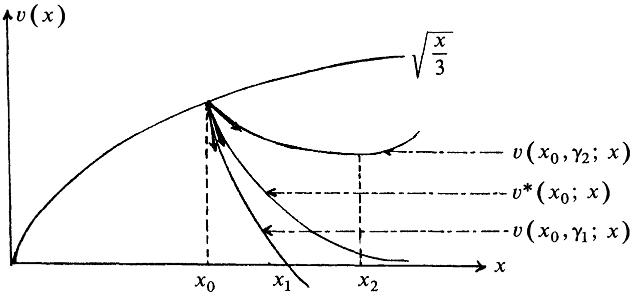


FIG. 12. *Représentation de $v(x_0,\gamma_1;\ \cdot)$, $v^*(x_0;\ \cdot)$, $v(x_0,\gamma_2;\ \cdot)$ pour $\gamma_1<\gamma^*(x_0)<\gamma_2$.*

Démonstration. i) Nous établissons le premier résultat; le deuxième se démontre de même. Soit donc $\gamma_1<\gamma^*(x_0)$, et raisonnons par l'absurde: supposons que $v(x_0,\gamma_1;\ x)$ ne s'annule jamais sur $]x_0,+\infty[$. D'après l'unicité de la solution du problème $\mathcal{Q}_{x_0}$, les deux fonctions $v^*(x_0;\ \cdot)$ et $v(x_0,\gamma_1;\ \cdot)$ prennent la même valeur en un point x' tel que $x'>x_0$ (voir figure 13):

$$v^*(x_0;\ x')=v(x_0,\gamma_1;\ x').$$

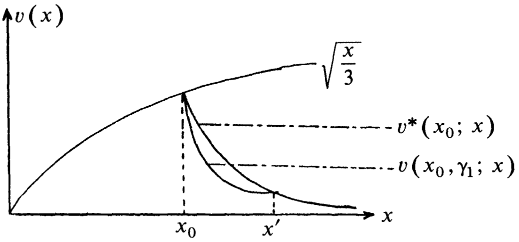


FIG. 13

Posons

$$w(x)=v^*(x_0;\ x)-v(x_0,\gamma_1;\ x).$$

w s'annule en x_0 et x', sa dérivée s'annule donc en un point x'' tel que:

$$x_0<x''<x',\quad w'(x'')=0.$$

Or

$$w''(x)=x\big(v^*(x_0;\ x)-v(x_0,\gamma_1;\ x)\big)-\Big(\big(v^*(x_0;\ x)\big)^3-\big(v(x_0,\gamma_1;\ x)\big)^3\Big).$$

w'' est positive sur $]x_0,x'[$, et $w'(x_0)=\gamma^*(x_0)-\gamma_1>0$. w' est alors strictement positive sur $]x_0,x'[$, et ne peut s'annuler en x''.

ii) La majoration résulte du lemme 2.4. En effet, la fonction

$$x\to\big(v^{*\prime}(x_0;\ x)\big)^2-x\big(v^*(x_0;\ x)\big)^2+\tfrac{1}{2}\big(v^*(x_0;\ x)\big)^4$$

est décroissante et tend vers 0 à l'infini. Elle est donc positive en x_0:

$$\gamma^{*2}(x_0)-\frac{5x_0^2}{18}\geqq 0.$$

La minoration provient de la remarque suivante: la fonction v définie sur $]x_0,+\infty[$ par

$$v(x)=\frac{x_0^{3/4}}{\sqrt{3}}\exp\left(\frac{2}{3}x_0^{3/2}x^{-1/4}\exp\left(-\frac{2}{3}x^{3/2}\right)\right)$$

est une sous-solution. Sa dérivée en x_0 est donc inférieure à $\gamma^*(x_0)$.

Nous pouvons de plus préciser le comportement asymptotique des fonctions $v^*(x_0;\ \cdot)$.

Effectuons un changement d'échelle et une translation; posons pour celà:

$$\text{(2.28)}\qquad \begin{aligned} &v^*(x_0;\ x)=\sqrt{x_0}\,w_{x_0}(y),\\ &y=\sqrt{x_0}\,(x-x_0). \end{aligned}$$

w_{x_0} est une fonction définie sur $[0, +\infty[$, solution du problème aux limites:

$$(2.29)\qquad \begin{aligned} &-w''+\left(1+\frac{y}{x_0^{3/2}}\right)w-w^3=0,\\ &w(0)=\frac{1}{\sqrt{3}},\qquad w(x)\to 0 \quad \text{quand } x\to+\infty. \end{aligned}$$

De plus la pente de w_{x_0} en 0 est déterminée:

$$(2.30)\qquad w'_{x_0}(0)=\frac{\gamma^*(x_0)}{x_0}.$$

Nous avons tracé ci-dessous les représentations de $v^*(x_0;\ \cdot)$ sur $[x_0, x_0+A]$ et de w_{x_0} sur $[0, A]$ pour différentes valeurs de x_0. Il apparait nettement sur la fig. 14 que la pente de v^* en $x_0, \gamma^*(x_0)$ décroit et tend vers $-\infty$ lorsque x_0 croit et tend vers $+\infty$. La figure 15 semble montrer que la famille w_{x_0} est croissante sur $[0, +\infty[$ et converge uniformément lorsque x_0 tend vers $+\infty$. C'est ce résultat que nous allons établir.

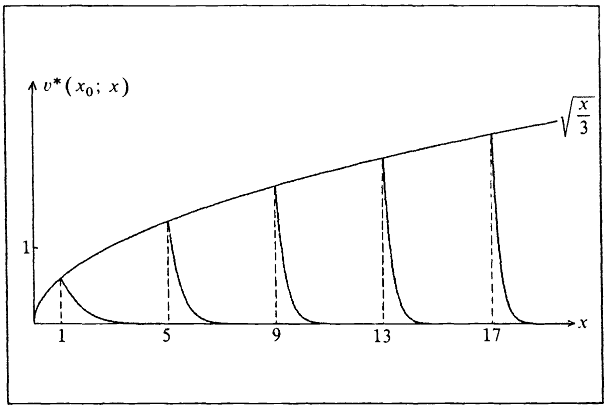


FIG. 14. *Représentation de $v^*(x_0;\ \cdot)$ sur $[x_0,\ +\infty[$ pour différentes valeurs de x_0.*

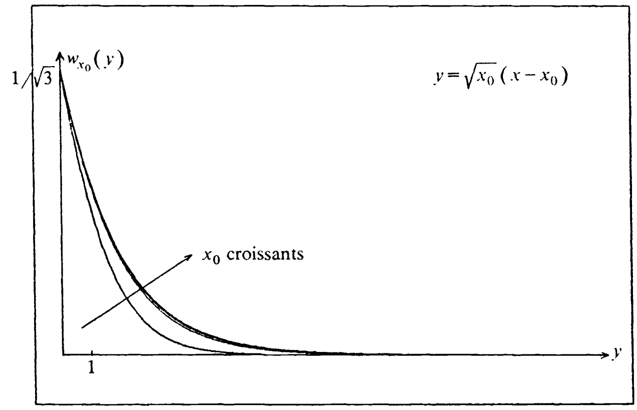


FIG. 15. *Représentation de w_{x_0} sur $[0,\ +\infty[$ pour différentes valeurs de x_0.*

LEMME 2.6. Comportement asymptotique de $v^*(x_0;\ \cdot)$ lorsque x_0 tend vers l'infini. *La famille de fonctions w_{x_0} est croissante avec x_0 et converge uniformément vers $V(\frac{1}{3}, -\frac{1}{3}\sqrt{\frac{5}{2}}\ ;\ \cdot)$ lorsque x_0 tend vers $+\infty$. En particulier*

i) *la fonction $x_0 \to \gamma^*(x_0)/x_0$ est croissante et tend vers $-\frac{1}{3}\sqrt{\frac{5}{2}}$ lorsque x_0 tend vers $+\infty$;*

ii) *sur $]x_0, +\infty[$, $v^*(x_0;\ x) \leqq \lambda\sqrt{x_0}\, e^{-\sqrt{x_0}(x-x_0)}$, où λ est une constante.*

Rappelons que $V(1/\sqrt{3},\ -\frac{1}{3}\sqrt{\frac{5}{2}}\ ;\ \cdot)$ est définie en (1.27) comme la solution du problème:

$$-v''+v-v^3=0, \qquad x \geqq 0,$$
$$v(0)=\frac{1}{\sqrt{3}},$$
$$v'(0)=-\frac{1}{3}\sqrt{\frac{5}{2}}\,.$$

Démonstration. 1. Montrons d'abord que la famille des fonctions w_{x_0} est croissante avec x_0. Soient donc deux réels positifs x_0 et x_1. Nous allons établir l'inégalité:

$$\forall y \geqq 0, \quad \frac{1}{\sqrt{x_0}} v^*\left(x_0;\ x_0+\frac{y}{\sqrt{x_0}}\right) \geqq \frac{1}{\sqrt{x_1}} v^*\left(x_1;\ x_1+\frac{y}{\sqrt{x_1}}\right). \tag{2.31}$$

Notons

$$\mu=\sqrt{\frac{x_1}{x_0}} \tag{2.32}$$

et, pour $x \geqq x_1$,

$$v(x)=\mu v^*(x_0;\ x_0+\mu(x-x_1)) \tag{2.33}$$

μ est supérieur à 1, et la fonction v vérifie

$$v(x_1)=v^*(x_1;\ x_1),$$
$$0 \leqq v(x) \leqq \sqrt{\frac{x}{3}}, \qquad x \geqq x_1,$$
$$-v''+xv-v^3=(1-\mu^3)(x-x_1)v(x) \geqq 0, \qquad x \geqq x_1.$$

v est donc une sur-solution pour le problème de minimisation $\mathscr{M}_{x_1}$, et

$$\forall x \geqq x_1, \quad v(x) \geqq v^*(x_1;\ x);$$

donc,

$$\forall y \geqq 0, \quad v\left(x_1+\frac{y}{\sqrt{x_1}}\right) \geqq v^*\left(x_1;\ x_1+\frac{y}{\sqrt{x_1}}\right).$$

L'inégalité (2.31) est établie.

2. Montrons maintenant que, pour tout x_0, la fonction w_{x_0} est bornée par $V(1/\sqrt{3},\ -\frac{1}{3}\sqrt{\frac{5}{2}}\ ;\ \cdot)$. Il suffit pour celà de prouver que la fonction φ définie sur $[x_0, +\infty[$ par

$$\varphi(x)=\sqrt{x_0}\, V\left(\frac{1}{\sqrt{3}},\ -\frac{1}{3}\sqrt{\frac{5}{2}}\ ;\ \sqrt{x_0}(x-x_0)\right)$$

est une sous-solution pour le problème de minimisation $\mathcal{M}_{x_0}$, ce qui se vérifie immédiatement.

La famille des w_{x_0} est une famille croissante de fonctions monotones, uniformément majorée sur $[0, +\infty[$; elle converge donc presque partout. Pour assurer la convergence uniforme, il suffit de vérifier que $w'_{x_0}(x)$ est uniformément bornée sur $[0, +\infty[$. Or:

$$\forall x \geqq 0, \quad 0 \geqq w'_{x_0}(x) \geqq w'_{x_0}(0) = \frac{\gamma^*(x_0)}{x_0}$$

et d'après le lemme 2.4,

$$\frac{\gamma^*(x_0)}{x_0} \geqq -\frac{1}{\sqrt{3}}\left(1 + \frac{1}{4}x_0^{-3/2}\right) \geqq C \quad \text{pour } x_0 \text{ assez grand.}$$

La famille w_{x_0} converge donc uniformément vers une fonction continue w.

3. Il nous reste à prouver que la limite de w_{x_0} est $V(1/\sqrt{3}, -\frac{1}{3}\sqrt{\frac{5}{2}}\,;\,\cdot)$. Rappelons d'abord que, d'après le théorème 1.1, v est donnée explicitément par

$$V\left(\frac{1}{\sqrt{3}}, -\frac{1}{3}\sqrt{\frac{5}{2}}\,;\,y\right) = 2\sqrt{2}\,\frac{e^{y+X}}{1+e^{(y+X)}}$$

où X est déterminé par

$$e^X = \sqrt{5} + \sqrt{6},$$

d'où

$$\forall y \geqq 0, \quad V\left(\frac{1}{\sqrt{3}}, -\frac{1}{3}\sqrt{\frac{5}{2}}\,;\,y\right) < \lambda e^{-y},$$

et puisque v est un majorant de w_{x_0}

$$\forall x_0 > 0, \forall y \geqq 0, \quad w_{x_0}(y) \leqq \lambda y e^{-y} \leqq C.$$

Nous avons ainsi tous les éléments pour passer à la limite dans l'équation

$$-w'' + \left(1 + \frac{y}{x_0^{3/2}}\right)w - w^3 = 0, \qquad y \geqq 0.$$

La limite w de w_{x_0} vérifie donc

$$\begin{aligned}
&-w'' + w - w^3 = 0,\\
&w(0) = \frac{1}{\sqrt{3}}, \qquad w(y) \to 0 \text{ quand } y \to +\infty,\\
&w'(0) < 0,
\end{aligned}$$

et w est bien égale à $V(1/\sqrt{3}, \frac{1}{3}\sqrt{\frac{5}{2}}\,;\,\cdot)$. En particulier la fonction $x_0 \to \gamma^*(x_0)/x_0$ tend vers $-\frac{1}{3}\sqrt{\frac{5}{2}}$ lorsque x_0 tend vers $+\infty$.

La majoration ii) du lemme provient de la majoration établie plus haut:

$$\forall x_0>0, \forall y \geqq 0, \quad w_{x_0}(y) \leqq \lambda e^{-y}.$$

Ces propriétés asymptotiques étant établies, nous poursuivons notre étude en prolongeant les fonctions $v^*(x_0;\ \cdot)$ sur $[0,x_0]$.

2.1.4. Description des solutions de $\mathscr{D}_R^+$. Nous définissons un problème de Cauchy sur $[0,x_0]$ (en renversant le sens des x), par:

$$-v''+xv-v^3=0, \qquad 0 \leqq x \leqq x_0, \tag{2.34}$$

$$\begin{gathered} v(x_0)=\sqrt{\frac{x_0}{3}}, \\ v'(x_0)=\gamma^*(x_0). \end{gathered} \tag{2.35}$$

L'existence locale d'une solution à ce problème provient simplement des théorèmes usuels d'équations différentielles. L'existence globale sera assurée par des estimations à priori.

LEMME 2.7. *On a pour une solution v de* (2.34), (2.35) *les estimations à priori*:

$$\|v\|_{L^\infty([0,x_0])} \leqq \varphi(x_0) \tag{2.36}$$

$$\|v'\|_{L^\infty([0,x_0])} \leqq \psi(x_0) \tag{2.37}$$

où $\varphi(x_0)$ et $\psi(x_0)$ sont données par:

$$\varphi^2(x_0)=x_0+\left(x_0^2+2\left(\gamma^{*2}(x_0)-\frac{5x_0^2}{18}\right)\right)^{1/2}, \tag{2.38}$$

$$\psi^2(x_0)=2x_0\varphi^2(x_0)+\gamma^{*2}(x_0)-\frac{5x_0^2}{18}. \tag{2.39}$$

Les estimations ne sont pas uniformes en x_0, car $\varphi(x_0)$ et $\psi(x_0)$ tendent vers l'infini avec x_0.

Démonstration. Les estimations seront établies en deux étapes. Nous montrerons d'abord que la suite des extrema locaux est croissante, puis nous majorerons le dernier extremum local, c'est-à dire celui qui est atteint le plus près de x_0.

Rappelons d'abord l'estimation d'énergie:

$$\frac{d}{dx}\left(v'^2-xv^2+\frac{v^4}{2}\right)=-v^2 \quad \text{sur } [0,x_0]. \tag{2.40}$$

La fonction

$$x \to v'^2-xv^2+\frac{v^4}{2}$$

est décroissante. Puisque v est le prolongement de $v^*(x_0,\cdot)$ elle est égale en x_0 à $(\gamma^*(x_0))^2-5x_0^2/18$ qui est positif, et donc

$$\forall x \in [0,x_0], \quad v'^2-xv^2+\frac{v^4}{2} \geqq \gamma^{*2}(x_0)-\frac{5x_0^2}{18}. \tag{2.41}$$

En particulier, en un point où la dérivée s'annule, nous obtenons:

$$\forall x \in [0,x_0], \quad v'(x)=0 \Rightarrow v^2 \geqq 2x. \tag{2.42}$$

Montrons le résultat suivant: *si deux extrema locaux successifs sont atteints en* x_1 *et* x_2, $0 \leqq x_1 < x_2 < x_0$, *alors* $|v(x_1)| < |v(x_2)|$.

Nous raisonnerons par l'absurde: soient x_1 et x_2 deux points tels que

$$0 < x_1 < x_2 < x_0, \qquad |v(x_1)| \geqq |v(x_2)|.$$

Nous sommes par exemple dans le cas de la figure 16.

Par commodité, nous noterons

$$v_1 = v(x_1), \qquad v_2 = v(x_2).$$

Nous intégrons l'égalité d'énergie (2.40) entre x_1 et x_2. Puisque v' s'annule en x_1 et x_2, nous avons:

$$-x_1v_1^2 + \frac{v_1^4}{2} - \left(-x_2v_2^2 + \frac{v_2^4}{2}\right) = \int_{x_1}^{x_2} v^2(x)\,dx.$$

Par hypothèse, nous pouvons majorer v sur $[x_1,x_2]$ par v_1^2. Après simplification, nous obtenons:

$$\frac{v_1^2+v_2^2}{2} \leqq x_2.$$

Mais d'autre part, d'après (2.42),

$$\frac{v_1^2+v_2^2}{2} \geqq v_2^2 \geqq 2x_2,$$

ce qui apporte une contradiction.

Le cas où un extremum local est atteint en zéro se traite de même.

Il nous reste à établir les estimations sur l'extremum le plus proche de x_0. Comme pour la démonstration précédente, deux situations peuvent se présenter: l'extremum est atteint, soit en 0, soit en $x_1 > 0$ (voir figure 17).

Nous traitons le premier cas, le deuxième se résout de même. Nous intégrons l'égalité d'énergie entre 0 et x_0:

$$v'^2 + \frac{v_0^4}{2} - \left(\gamma^{*2}(x_0) - \frac{5x_0^2}{18}\right) = \int_0^{x_0} v^2(x)\,dx.$$

Nous majorons v^2 sur $[0,x_0]$ par v_0^2 et nous obtenons

$$\frac{v_0^4}{2} - x_0v_0^2 - \left(\gamma^{*2}(x_0) - \frac{5x_0^2}{18}\right) \leqq 0,$$

ce qui n'est réalisé que si

$$v_0^2 \leqq x_0 + \left(x_0^2 + \left(\gamma^{*2}(x_0) - \frac{5x_0^2}{18}\right)\right)^{1/2}$$

ce qui constitue l'estimation (2.36).

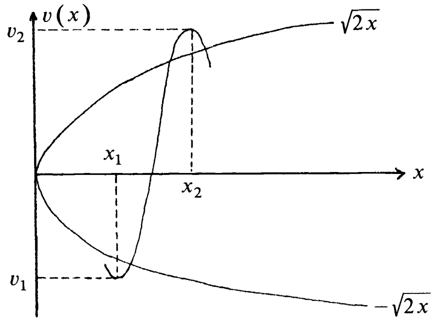
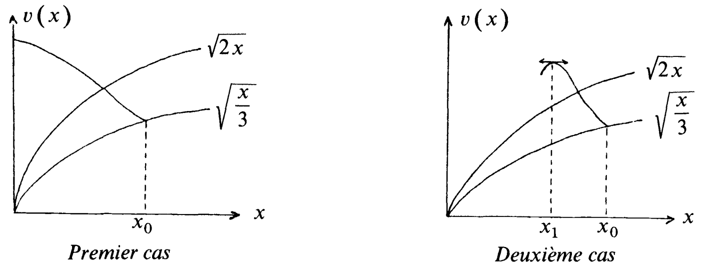


FIG. 16



Premier cas

Deuxième cas

FIG. 17

Pour établir l'estimation (2.37), il suffit de remarquer que

$$\forall x \in [0, x_0], \quad 0 \leqq v'^2 - xv^2 + \frac{v^4}{2} \leqq v_0'^2 + \frac{v_0^4}{2}$$

et

$$v_0'^2 + \frac{v_0^4}{2} \leqq \gamma^{*2}(x_0) - \frac{5x_0^2}{18} + x_0\varphi^2(x_0).$$

Nous avons ainsi:

$$\forall x \in [0, x_0], \quad v'^2 \leqq 2x_0\varphi^2(x_0) + \gamma^{*2}(x_0) - \frac{5x_0^2}{18} = \psi^2(x_0).$$

Nous avons ainsi établi les estimations L^∞ sur u et u'.

Le lemme 2.7 nous permet donc de prolonger $v^*(x_0; \cdot)$ sur toute la droite $\mathbb{R}_+$. En particulier, nous pouvons définir sa valeur en zéro ainsi que celle de sa dérivée, et donc une application de $\mathbb{R}_+$ dans $\mathbb{R}^2$ par

$$(2.43) \qquad \sigma(x_0) = \left(v^*(x_0; 0), \frac{dv^*}{dx}(x_0; 0) \right).$$

Nous noterons $\mathcal{S}$ l'image de $\mathbb{R}_+$ par σ:

$$(2.44) \qquad \mathcal{S} = \left\{ (v_0, v_0') \in \mathbb{R}^2, \exists x_0 \in \mathbb{R}_+, (v_0, v_0') = \sigma(x_0) \right\}.$$

Les solutions de $\mathcal{Q}_R^+$ sont les solutions de l'équation différentielle dont les valeurs initiales appartiennent à $\mathcal{S}$.

THÉORÈME 2.2. Caractérisation des solutions de $\mathcal{Q}_R^+$. *L'ensemble des solutions de* $\mathcal{Q}_R^+$ *est l'ensemble des* $v^*(x_0; \cdot)$ *et* $-v^*(x_0; \cdot)$ *lorsque* x_0 *décrit* $\mathbb{R}_+$.

Il ne nous reste plus qu'à relier les caractérisations des solutions des problèmes $\mathscr{Q}_c^-$ et $\mathscr{Q}_R^+$.

2.1.5. Le système d'équations non linéaires vérifié par v_0 et v_0'. Comme pour le problème $\mathscr{P}_{\alpha,p}$, nous pouvons établir une équivalence entre les problèmes réel et complexe sur $\mathbb{R}_+$:

LEMME 2.8. Equivalence entre $\mathscr{Q}_c^+$ et $\mathscr{Q}_R^+$. *La fonction u à valeurs complexes est solution de $\mathscr{Q}_c^+$ si et seulement si il existe un nombre complexe z_0 et une fonction réelle v tels que*

$$(2.45)\qquad u=\frac{z_0}{\sqrt{p}}v, \qquad |z_0|=1, \qquad v \textit{ solution de } \mathscr{Q}_R^+.$$

La démonstration est pratiquement la même que dans le lemme 1.2. Ce qui tient lieu ici d'énergie est la fonction

$$x\to|u'|^2-x|u|^2+\frac{p}{2}|u|^4$$

qui est décroissante, ce qui permet d'affirmer que u' tend vers zéro lorsque x tend vers $+\infty$.

Notons $E_{\alpha,p}$ la famille d'ellipses définie par:

$$(2.46)\qquad E_{\alpha,p}=\left\{(X,Y)\in\mathbb{R}^2, \left(Y-\frac{k}{2}X\right)^2+\frac{3}{4}k^2X^2-\alpha^2p=0\right\}$$

où k est le réel défini dans le lemme 2.1.

Nous pouvons maintenant décrire explicitement les solutions de $\mathscr{Q}_{\alpha,p}$.

PROPOSITION 2.1. Caractérisation des solutions de $\mathscr{Q}_{\alpha,p}$. *Le problème $\mathscr{Q}_{\alpha,p}$ admet une solution u si et seulement si il existe (v_0,v_0') appartenant à $\mathscr{S}\cap E_{\alpha,p}$. Elle s'écrit alors*:

$$(2.47)\qquad u(x)=\begin{cases}\dfrac{1}{\alpha p}\left(v_0'+ke^{-2i\pi/3}v_0\right)v^*(x_0;\ \cdot), & x\geqq 0,\\[2ex] \dfrac{1}{\alpha p}v_0\left(v_0'+ke^{-2i\pi/3}v_0\right)\dfrac{\mathrm{Ai}\left(xe^{2i\pi/3}\right)}{\mathrm{Ai}(0)}, & x\leqq 0,\end{cases}$$

où x_0 est tel que

$$(2.48)\qquad v_0=v^*(x_0;0), \qquad v_0'=\frac{d}{dx}v^*(x_0;0).$$

Démonstration. Soit u une solution de $\mathscr{Q}_{\alpha,p}$. D'après le lemme 2.8, u peut s'exprimer sur $\mathbb{R}_+$ à partir des solutions de $\mathscr{Q}_R^+$, et donc

$$(2.49)\qquad \exists x_0\in\mathbb{R}_+, \exists z_0\in\mathbb{C}, \quad u=\frac{z_0}{\sqrt{p}}v^*(x_0;\ \cdot), \qquad |z_0|=1.$$

Nous utilisons maintenant la caractérisation des solutions de $\mathscr{Q}_c^-$ donnée par le lemme 2.1, et les conditions de transmission

$$\begin{aligned}&u'(0_-)+ke^{2i\pi/3}u(0_-)=0,\\&u(0_-)=u(0_+),\\&u'(0_-)=u'(0_+)-\alpha.\end{aligned}$$

Ce système se traduit par

$$\frac{z_0}{\sqrt{p}}\left(\frac{dv^*}{dx}(x_0;\,0)+ke^{2i\pi/3}v^*(x_0;\,0)\right)=\alpha. \tag{2.50}$$

Notons

$$v_0=v^*(x_0;\,0),\qquad v_0'=\frac{dv^*}{dx}(x_0;\,0).$$

La relation (2.50) se réécrit sous la forme

$$\left|v_0'+ke^{2i\pi/3}v_0\right|^2=\alpha^2p,$$

$$z_0=\frac{1}{\alpha p}\left(v_0'+ke^{-2i\pi/3}v_0\right).$$

(v_0,v_0') appartient donc à $\mathscr{S}\cap E_{\alpha,p}$, et u est donnée par l'expression (2.47).

Réciproquement si (v_0,v_0') appartient à $\mathscr{S}\cap E_{\alpha,p}$, il leur correspond par $\mathscr{s}^{-1}$ un réel x_0 positif tel que $(v_0,v_0')=\mathscr{s}(x_0)$, et la fonction u définie par (2.47) est solution de $\mathscr{Q}_{\alpha,p}$.

Nous sommes donc ramenés à étudier le nombre de solutions du problème

$$(v_0,v_0')\in\mathscr{S}\cap E_{\alpha,p} \tag{2.50}$$

en fonction du paramètre α^2p. C'est un système non linéaire de deux relations à deux inconnues; l'une, $E_{\alpha,p}$, est connue exactement, et nous approcherons $\mathscr{S}$ numériquement.

2.2. Etude numérique du système non linéaire reliant v_0 et v_0'. Nous commençons par étudier l'approximation numérique de $v^*(x_0;\,\cdot)$. Un pas h de discrétisation étant donné, nous approchons $v^*(x_0;\,\cdot)$ par $v_h^*(x_0;\,\cdot)$ définie de la manière suivante:

Nous discrétisons l'équation différentielle $-v''+xv-v^3=0$ sur $[0,\,+\infty[$ par un schéma aux différences finies centré.

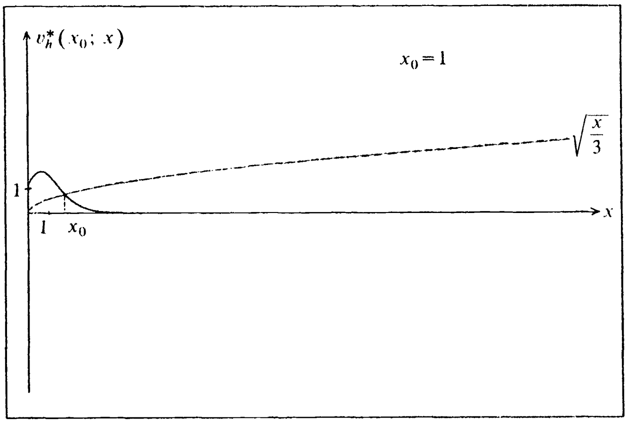
$$\begin{aligned}&-\frac{v_{i+1}-2v_i+v_{i-1}}{h}+x_iv_i-v_i^3=0,\\ &x_i=x_0+\varepsilon ih,\\ &v_i\sim v^*(x_0;\,x_i);\end{aligned} \tag{2.51}$$

ε prend la valeur $+1$ ou -1 selon qu'on se place sur $[0,x_0]$ ou $[x_0,\,+\infty]$.

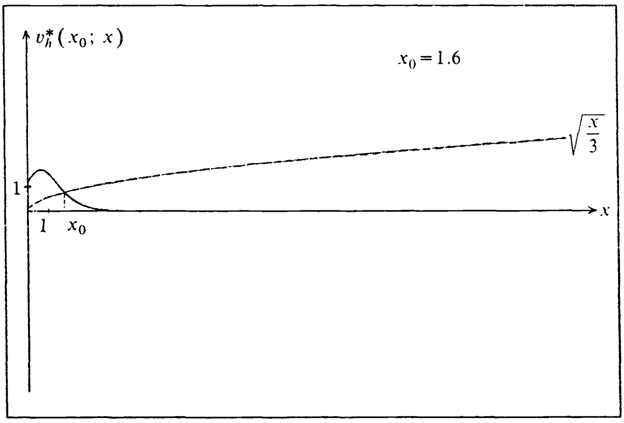
1. D'après le lemme 2.5, nous avons un encadrement de $\gamma^*(x_0)$ et de $v^*(x_0;\,\cdot)$ sur $[x_0,\,+\infty[$. Ces propriétés nous permettent de calculer une approximation $\gamma_h^*(x_0)$ de $\gamma^*(x_0)$ avec toute la précision souhaitée, et $v_h^*(x_0;\,\cdot)$ sur $[x_0,\,x_0+A[$ à l'aide du schéma (2.51).

2. Nous calculons $v_h^*(x_0;\,\cdot)$ sur $[0,x_0]$ comme solution de l'équation (2.51), associée aux valeurs de Cauchy $(v^*(x_0;\,x_0),\,\gamma_h^*(x_0))$.

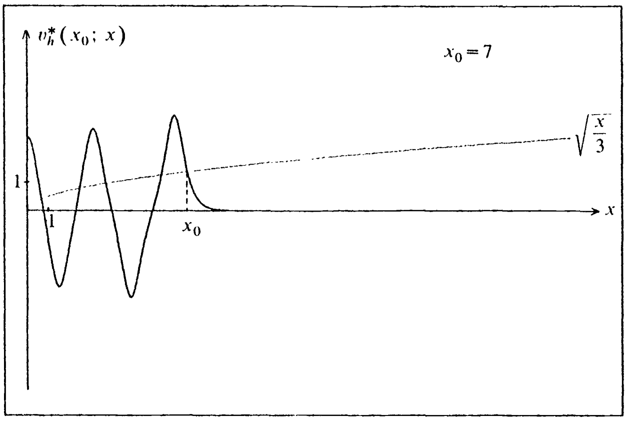
3. Nous traçons les courbes représentant $v_h^*(x_0;\,\cdot)$ pour différentes valeurs du paramètre x_0 (voir figure 18).



$x_0=1$. $v_h^*(x_0;\,\cdot)$ *est strictement décroissante sur* $[0,+\infty[$.

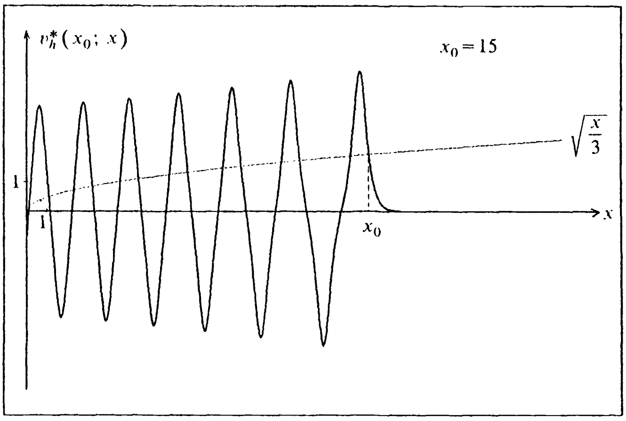


$x_0=1.6$. *Naissance d'une oscillation.*



$x_0=7$. $v_h^*(x_0;\,\cdot)$ *a* 2 *oscillations.*

FIG. 18. *Représentation de la solution approchée de* $\mathscr{Q}_R^+$ *pour différentes valeurs de* x_0, *pour un pas de discrétisation* $h=\frac{1}{100}$.



$x_0 = 15$. *Le nombre d'oscillations croît avec* x_0.

FIG. 18 (*continued*).

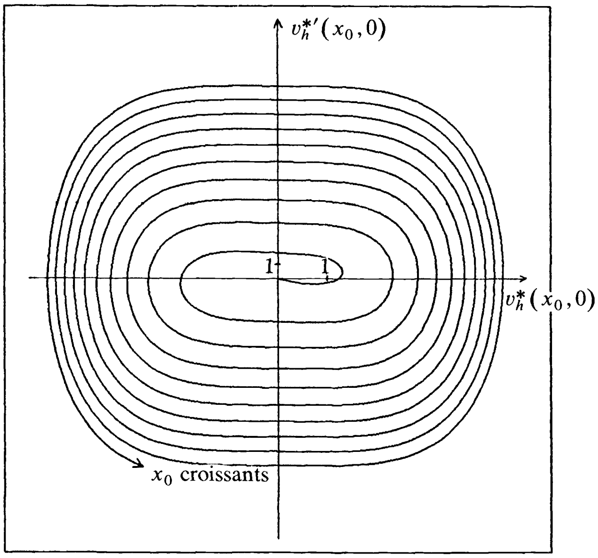


FIG. 19. *Représentation de l'approximation* $\mathcal{S}_h$ *de* $\mathcal{S}$.

Tant que x_0 reste petit, v_h^* est décroissante sur $[0, x_0]$. Pour une valeur de x_0 proche de 1, il apparaît une oscillation. Puis, lorsque x_0 grandit, le nombre d'oscillations de $v_h^*(x_0; \cdot)$ sur $[0, x_0]$ grandit. Pour $x \geqq x_0$, $v_h^*(x_0; \cdot)$ est décroissante et tend vers zéro de plus en plus rapidement lorsque x_0 croit. Tout ceci correspond bien aux résultats démontrés précédemment.

Nous représentons ci-dessous l'image par l'application σ_h du segment $[0, 20]$.

$$\mathcal{S}_h = \left\{ \left(v_h^*(x_0; 0), \frac{dv_h^*}{dx}(x_0; 0) \right); \; x_0 \in [0, 20] \right\};$$

$\sigma_h(x_0)$ a été calculé pour des valeurs de x_0 variant de 0 à 20 par pas de $\frac{1}{100}$ (voir figure 19).

La fonction $|s_h(x_0)|$ croît: $\mathscr{S}_h$ est une spirale qui se déroule;celà est du au fait que lorsque x_0 est grand, $v_h^*(x_0;\ \cdot)$ oscille de plus en plus sur $[0, x_0]$.

Nous n'avons pas d'estimation d'erreur sur $\mathscr{S}_h$, mais nous avons fait le calcul numérique pour différentes valeurs du pas de discrétisation: $h = \frac{1}{100}$, $h = \frac{1}{500}$, $h = \frac{1}{1000}$: les courbes sont strictement superposables: $\mathscr{S}_h$ semble donc constituer, pour x_0 variant entre 0 *et* 20, *une bonne approximation de $\mathscr{S}$.*

Par ailleurs nous n'avons pas pu établir le comportement asymptotique de $\mathscr{S}(x_0)$: la nécessité de prendre un pas de plus en plus petit augmente notablement le volume des calculs et en limite le fiabilité: nous ne savons donc pas s'il existe un cycle limite ou si, au contraire, $|\mathscr{s}_h(x_0)|$ tend vers l'infini avec x_0.

Revenons maintenant au nombre de solutions du problème $\mathscr{Q}_{\alpha,p}$. Il est donné, rappelons le, par le nombre d'éléments (v_0, v_0') dans l'intersection de $\mathscr{S}$ et $E_{\alpha,p}$. Dans la mesure où $\mathscr{S}_h$ est une approximation de $\mathscr{S}$, il est donné par le nombre d'éléments dans $\mathscr{S}_h \cap E_{\alpha,p}$. Nous avons représenté ci-dessous $\mathscr{S}_h$ et les ellipses $E_{\alpha,p}$ pour des valeurs du paramètre $\alpha^2 p$ croissant entre 1 et 5 (voir figure 20).

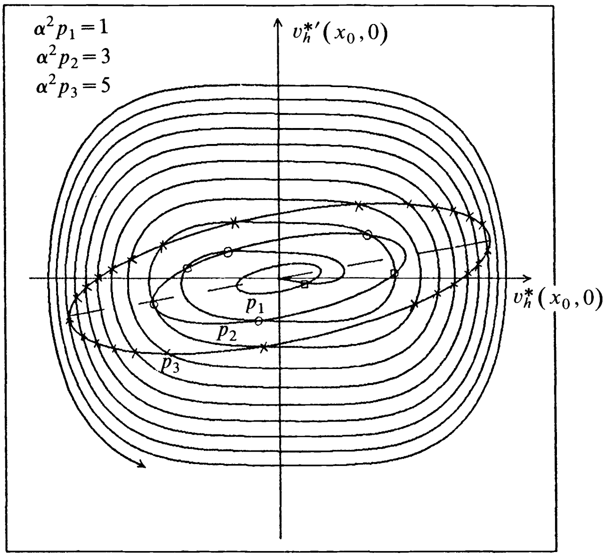


FIG. 20. *Représentation de $\mathscr{S}_h \cap E_{\alpha,p}$ pour $\alpha^2 p = 1, 3, 5$.*

Lorsque $\alpha^2 p$ varie, il y a toujours au moins un élément dans $\mathscr{S}_h \cap E_{\alpha,p}$. Plus $\alpha^2 p$ est grand, plus il y a d'éléments dans l'intersection. Par contre nous ne savons pas s'il existe une valeur limite de $\alpha^2 p$ au-dessus de laquelle $\mathscr{S}_h \cap E_{\alpha,p}$ contient une infinité d'éléments.

Rappelons que nous avons pris toutes les précautions pour que $\mathscr{S}_h$ soit une "bonne" approximation de $\mathscr{S}$. Dans cette mesure, il semble que:

Le problème $\mathscr{Q}_{\alpha,p}$ admet toujours au moins une solution. Lorsque $p\alpha^2$ grandit, le nombre des solutions croit comme le montre la figure 14. *L'ensemble des solutions peut être paramétré par $\{x_i\}_{i=1,I}$ et plus x_i est grand, plus la solution associée oscille.*

BIBLIOGRAPHIE

[1] M. ABRAMOWITZ ET I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] J. C. ADAM, A. GOURDIN-SERVENIERE ET G. LAVAL, *Efficiency of resonant absorption of electromagnetic waves in an inhomogeneous plasma*, Rapport du Centre de Physique Théorique de l'Ecole Polytechnique, 1981.

[3] V. Arnold, *Equations différentielles ordinaires*, Ed. MIR, Moscou, 1974.

[4] H. Berestycki et P. L. Lions, *Non linear scalar fields equations I*, Université Paris VI, Rapport interne N° 80 020, 1980.

[5] T. Cazenave, *Stabilité et instabilité des états stationnaires dans les équations de Schrödinger non linéaires*, Université Paris VI, Rapport interne N° 82008, 1982.

[6] J. Ginibre et G. Velo, *On a class of non linear Schrödinger equations*, I *et* II, J. Funct. Anal., 32 (1979), pp. 1–71.

[7] R. T. Glassey, *On the blowing-up of solutions to the Cauchy problem for non linear Schrödinger equations*, J. Math. Phys., 18 (1977), pp. 1794–1797.

[8] G. J. Morales et Y. C. Lee, *Generation of density cavities and localized electric fields in a non uniform plasma*, Phys. Fluids, 20 (1977), pp. 1135–1146.

[9] V. E. Zakharov et A. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in non linear media*, Soviet Phys. JETP, 34 (January 1972).

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

SHOCK LAYERS IN PERTURBED SYSTEMS RELATED TO STEADY CONSERVATION LAWS*

F. A. HOWES†

Abstract. Singular perturbation techniques are used to determine the locations of shock layers in steady solutions of a hyperbolic system to which viscosity terms have been added.

Key words. shock layer, jump condition, singular perturbation analysis

AMS(MOS) subject classifications. Primary 34B15, 34D15, 34E15

1. Introduction.

As a prelude to a study of the existence, asymptotic behavior and stability of solutions of initial-boundary value problems for the weakly coupled system

$$\text{(IBVP)}\qquad \underset{\sim}{u}_t+\mathrm{diag}\{f_1(\underset{\sim}{u}),\cdots,f_n(\underset{\sim}{u})\}\underset{\sim}{u}_x=\varepsilon\underset{\sim}{u}_{xx},$$

we consider the associated steady boundary value problem

$$\text{(BVP)}\qquad \begin{aligned}&\varepsilon\underset{\sim}{u}_{xx}=\mathrm{diag}\{f_1(\underset{\sim}{u}),\cdots,f_n(\underset{\sim}{u})\}\underset{\sim}{u}_x, \qquad a<x<b,\\ &\underset{\sim}{u}(a,\varepsilon)=\underset{\sim}{A}, \qquad \underset{\sim}{u}(b,\varepsilon)=\underset{\sim}{B},\end{aligned}$$

as the positive perturbation parameter ε tends to zero. The steady problem is of course much simpler than the time-dependent one, and so it is a natural starting point for the investigation of the structure of solutions of (IBVP), some of which evolve as $t\to\infty$ into solutions of (BVP). Despite its relative simplicity, though, (BVP) has solutions with interesting features, including boundary layers and interior (shock) layers. Shock layers connect different constant equilibrium states, and in a multicomponent system like (BVP), the location of such a layer in one component usually depends on the occurrence or nonoccurrence of interior layers in the other components. Thus the theory for (BVP) is inherently more complicated than the well-known theory for its scalar counterparts $(S_1)\,\varepsilon u_{xx}=f(u)u_x$ or $(S_2)\,\varepsilon u_{xx}=f(x,u)u_x+g(x,u)$.

Before turning to our treatment of (BVP), let us review briefly some of the previous work on (BVP) and (IBVP). The scalar theory for (S_1) and (S_2) is now fairly well in hand; see, for example, [16, Chap. 5], [9, Chap. 2] and [4]. Using some of these results, O'Malley [17], [18], O'Donnell [13], [15] and the author [6], [7] have studied the Dirichlet problem for the more general steady system $\varepsilon\underset{\sim}{u}_{xx}=F(x,\underset{\sim}{u})\underset{\sim}{u}_x+\underset{\sim}{g}(x,\underset{\sim}{u})$, where either $F(x,\underset{\sim}{u})=\mathrm{diag}\{f_1(x,\underset{\sim}{u}),\cdots,f_n(x,\underset{\sim}{u})\}$ or F is an arbitrary $(n\times n)$-matrix-valued function. In the present treatment of (BVP) we make extensive use of many of O'Donnell's ideas, and we generalize his results on the occurrence of shock layers by allowing shocks in different components at the same point. Turning to the time-dependent problem (IBVP), we note that the scalar version has been studied by a number of authors; cf. for example [9, Chap. 4], [20, Part I], [3] and [8] and the references contained therein. Less is known about the general problem (IBVP), but some representative results can be found in [10], [19, Part III], [11] and [6].

* Received by the editors December 7, 1984, and in revised form June 3, 1985. This work was supported by the National Science Foundation under grant DMS-8319783.

† Department of Mathematics, University of California at Davis, Davis, California 95616.

2. The componentwise theory. The work of O'Donnell [14], [15] (cf. also [1, Chap. 7]) is concerned, in part, with the Dirichlet problem on (a,b)

$$\text{(DP)}\qquad \begin{gathered} \varepsilon u_i''=f_i(x,u_1,\cdots,u_n)u_i'+g_i(x,u_1,\cdots,u_n),\\ u_i(a,\varepsilon)=A_i,\qquad u_i(b,\varepsilon)=B_i \end{gathered}$$

where $'=d/dx$ and $i=1,\cdots,n$. Under the principal assumption that the reduced system $f_i(x,\underset{\sim}{u})u_i'+g_i(x,\underset{\sim}{u})=0$ has solutions $u_i=U_i(x)$ that satisfy this system in the strong sense that

$$f_i(x,u_1,\cdots,U_i(x),\cdots,u_n)U_i'(x)+g_i(x,u_1,\cdots,U_i(x),\cdots,u_n)=0 \tag{2.1}$$

for all values of u_j $(j\neq i)$ in some domain of interest, he is able to prove the existence of solutions of (DP) as $\varepsilon\to 0$ which are close to certain reduced solutions $\underset{\sim}{U}$ in most of $[a,b]$. In neighborhoods of the boundaries and/or interior points, however, such reduced solutions often must be supplemented with layer terms that allow either a boundary condition to be satisfied (boundary layer term) or a smooth transition from one reduced solution to another to take place (shock layer term).

To be more precise, let us look first for solutions of (DP) that exhibit boundary layer behavior at $x=a$; analogous results for boundary layer behavior at $x=b$ then follow by making the change of variable $x\to b+a-x$. We begin with the assumption that the reduced problem

$$f_i(x,\underset{\sim}{u})u_i'+g_i(x,\underset{\sim}{u})=0,\qquad u_i(b)=B_i,$$

has a solution $\underset{\sim}{U}=(U_1(x),\cdots,U_n(x))$ of class $C^{(2)}([a,b])$, and we define the regions for $i=1,\cdots,n$

$$\mathscr{D}_i=\{u_i\colon |u_i-U_i(x)|\leqq d_i(x)\},$$

where each d_i is a smooth positive function such that $d_i(x)\equiv|A_i-U_i(a)|+\delta$ for x in $[a,a+\delta/2]$ and $d_i(x)\equiv\delta$ for x in $[a+\delta,b]$, with δ a small positive constant. Each of the regions $\mathscr{D}_i$ is thus a δ-tube around $U_i(x)$ that widens near $x=a$ to include a boundary layer of size $|A_i-U_i(a)|+\delta$, since, in general, $U_i(a)\neq A_i$. It is within the region $\mathscr{D}=\prod_{i=1}^n\mathscr{D}_i$ that we look for a solution of (DP) by imposing further conditions on $f_i(x,\underset{\sim}{u})$ in $[a,b]\times\mathscr{D}$. The second assumption is that $\underset{\sim}{U}$ additionally satisfies the reduced system in the stronger sense that (2.1) obtains for all u_j $(j\neq i)$ in $\mathscr{D}_j$. This assumption allows us to decouple the system, and thereby apply the extensive scalar theory to the problem (DP) componentwise. Finally we require $\underset{\sim}{U}$ to be attractive in the sense that there exist positive constants k_i such that for x in $[a,b]$

$$f_i(x,u_1,\cdots,U_i(x),\cdots,u_n)\leqq -k_i<0$$

for all u_j $(j\neq i)$ in $\mathscr{D}_j$. In particular, we must have $f_i(x,\underset{\sim}{U}(x))\leqq -k_i<0$ in $[a,b]$, and so these inequalities imply that $\underset{\sim}{U}$ is necessarily asymptotically stable in the linear approximation. To see this, note that the perturbation $w_i=u_i-U_i$ satisfies approximately the equation $\varepsilon w_i''=f_i(x,\underset{\sim}{U}+\underset{\sim}{w})w_i'$ and the terminal condition $w_i(b,\varepsilon)=0$, since U_i is a solution of the reduced problem; hence, $w_i(x,\varepsilon)=\mathcal{O}(|A_i-U_i(a)|\exp[-k_i(x-a)/\varepsilon])$ for x in $[a,b]$ as $\varepsilon\to 0$. Then under these assumptions it follows from a continuity argument that the problem (DP) has a solution $\underset{\sim}{u}=(u_1,\cdots,u_n)(x,\varepsilon)$ as $\varepsilon\to 0$ which is close to $(U_1(x),\cdots,U_n(x))$ in $[a+\delta,b]$, provided the "boundary layer jumps" $|A_i-U_i(a)|$ are sufficiently small. Estimates for the maximum allowable sizes of these jumps are given by O'Donnell [13] (cf. also [1, Chap. 7]) in the form of the integral conditions: if $U_i(a) \neq A_i$ then

$$(2.2) \qquad (U_i(a)-A_i)\int_{U_i(a)}^{\xi} f_i(a,A_1,\cdots,A_{i-1},s,A_{i+1},\cdots,A_n)\,ds>0$$

for all values of ξ between $U_i(a)$ and A_i, $\xi \neq U_i(a)$. Such conditions generalize the well-known integral condition of Coddington and Levinson [2] (cf. also [4]) for the scalar equation (S_2), and they are the starting point for our investigation in §4 of shock layers in solutions of (BVP). The usefulness of (2.2) is easily seen by examining a scalar problem like $\varepsilon u''=-uu'+u$, $u(0,\varepsilon)=A$, $u(1,\varepsilon)=B$; cf. [4], [9, Chap. 2]. Suppose that $B>1$. Then $U(x)=x+B-1$ is the solution of the reduced problem $uu'=u$, $u(1)=B$, and it is attractive in the sense that $f(U(x)) \leqq -(B-1)<0$ in $[0,1]$ for $f(u)=-u$. Now if $A>0$ then $f(u)<0$ for all values of u between A and $B-1$, and so we know that the solution of this problem has a boundary layer at $x=0$, that is, $(*)\lim_{\varepsilon\to 0}u(x,\varepsilon)=x+B-1$ in $[\delta,1]$. On the other hand, if $A\leqq 0$ then this inequality on f does not obtain for all such values of u, and we might be tempted to conclude that $(*)$ does not obtain. However, we see that $(*)$ *is* valid provided $A>-(B-1)$, since condition (2.2) is satisfied with this restriction on $A\leqq 0$. Thus it is the "integrated" effect of f in the boundary layer, as measured by (2.2), that determines if there is a boundary layer relative to the attractive reduced solution U. The n conditions in (2.2) are an extension of this basic fact to the system (DP).

Let us illustrate O'Donnell's approach by considering next solutions of (DP) in which there is a shock layer in just the kth component. Our basic assumption is that the reduced system has two $C^{(2)}$-solutions $\underset{\sim}{u}_L=(U_1(x),\cdots,U_{k-1}(x),\ U_L(x),\ U_{k+1}(x),\cdots,U_n(x))$ and $\underset{\sim}{u}_R=(U_1(x),\cdots,U_{k-1}(x),\ U_R(x),\ U_{k+1}(x),\cdots,U_n(x))$ which exist in $[a,b]$ and satisfy $U_i(a)=A_i$ and $U_i(b)=B_i$ for $i\neq k$, $U_L<U_R$, $U_L(a)=A_k$ and $U_R(b)=B_k$. In addition, we assume that $\underset{\sim}{u}_L$ and $\underset{\sim}{u}_R$ satisfy the reduced system in the stronger sense that for $i\neq k$ (2.1) obtains in (a,b) for all values of u_j $(j\neq i)$ in $\tilde{\mathscr{D}}_j$. (Here $\tilde{\mathscr{D}}_i=\{u_i\colon |u_i-U_i(x)|\leqq\delta\}$ for $i\neq k$ and $\tilde{\mathscr{D}}_k=\{u_k\colon U_L(x)-\delta\leqq u_k\leqq U_R(x)+\delta\}$.) For $i=k$ we assume that in (a,b)

$$f_k(x,u_1,\cdots,v,\cdots,u_n)v'+g_k(x,u_1,\cdots,v,\cdots,u_n)=0$$

for $v=U_L$ or U_R and for all values of u_j $(j\neq k)$ in $\tilde{\mathscr{D}}_j$. We can guarantee that these solutions are attractive by assuming also that, for $i\neq k$,

$$|f_i(x,u_1,\cdots,U_i(x),\cdots,u_n)|>0,$$

and that there exists a positive constant K for which

$$f_k(x,u_1,\cdots,U_L(x),\cdots,u_n)\geqq K>0, \qquad f_k(x,u_1,\cdots,U_R(x),\cdots,u_n)\leqq -K<0$$

in the regions defined above. Then the theory [13], [6] implies that (DP) has a solution $\underset{\sim}{u}=\underset{\sim}{u}(x,\varepsilon)$ as $\varepsilon\to 0$ such that for $i\neq k$

$$\lim_{\varepsilon\to 0}u_i(x,\varepsilon)=U_i(x) \quad \text{in } [a,b]$$

and

$$\lim_{\varepsilon\to 0}u_k(x,\varepsilon)=\begin{cases} U_L(x) & \text{in } [a,x_0-\delta],\\ U_R(x) & \text{in } [x_0+\delta,b],\end{cases}$$

provided the "shock strength" $|U_L(x_0)-U_R(x_0)|$ is sufficiently small. The location x_0 of the shock layer is determined as a solution $x=x_0$ of the functional equation

$$J[x]=\int_{U_L(x)}^{U_R(x)} f_k(x,U_1(x),\cdots,U_{k-1}(x),s,U_{k+1}(x),\cdots,U_n(x))\,ds=0$$

which also satisfies $J'[x_0]\neq 0$. This result can be extended in an obvious way to cases in which there are shock layers in different components of the solution at distinct points. The example $\varepsilon u''=-uu'+u$, $u(0,\varepsilon)=A$, $u(1,\varepsilon)=B$, again provides us with a good illustration of this result; cf. [4], [9, Chap. 2]. We saw above that if $B>1$ then the function $U_R(x)=x+B-1$ is the solution of the reduced problem $uu'=u$, $u(1)=B$, which satisfies $f(U_R(x))<0$ in $[0,1]$ for $f(u)=-u$. Similarly, if $A<-1$ then the function $U_L(x)=x+A$ is the solution of the left-hand reduced problem $uu'=u$, $u(0)=A$, and by virtue of the restriction on A, $f(U_L(x))>0$ in $[0,1]$. Now if $A<-(B-1)$ then the integral condition (2.2) does not obtain, and if $B>-(A+1)$ then the integral condition corresponding to (2.2) at $x=1$ also does not obtain. Thus if $A<-1$, $B>1$ and $|A+B|<1$ then boundary layer behavior relative to U_L or U_R is impossible; however, for these values of A and B,

$$(\dagger)\qquad J[x]=\int_{x+A}^{x+B-1}(-s)\,ds=0 \quad \text{at } x=x_0=(1-B-A)/2$$

in $(0,1)$ and $J'[x_0]\neq 0$. The solution of the example is therefore a shock layer joining $U_L(x)$ and $U_R(x)$ at the point x_0. This should come as no surprise since the relation $(\dagger)$, that is, $U_L^2(x_0)=U_R^2(x_0)$ or $U_L(x_0)+U_R(x_0)=0$, is nothing more than the Rankine-Hugoniot condition for a stationary shock wave at x_0; cf. [10, §2], [20, Chap. 2].

O'Donnell's shock layer theory does not apply to an autonomous problem like (BVP) since attractive reduced solutions are constants, and the functional J is thereby independent of x. In addition, his theory for (DP) is unable to describe what happens when there is a shock layer in two or more components at the *same* point, because this is no longer a scalar phenomenon. It turns out that we can now provide a shock layer theory for (BVP) which complements O'Donnell's boundary layer theory and which covers the occurrence of a shock layer in several components at the same point. Before doing so we review briefly the scalar theory for (S_1).

3. Shock layer theory for (S_1). The occurrence of shock layers in solutions of the scalar boundary value problem

$$(\mathrm{S}_1)\qquad \begin{aligned} &\varepsilon u''=f(u)u', \qquad 0<x<1,\\ &u(0,\varepsilon)=A, \quad u(1,\varepsilon)=B, \quad A<B, \end{aligned}$$

has been considered in [5] and [12]. (In this and the following sections we replace, without loss of generality, the finite interval $[a,b]$ with $[0,1]$, so as to simplify some of our formulas.) Let us recall briefly the results in [5].

Suppose that the function f is continuous for all u in $[A,B]$ and that the states $U_L\equiv A$ and $U_R\equiv B$ are such that $f(A)>0$ and $f(B)<0$. Then if $F(A)<F(B)$ $[F(A)>F(B)]$, for F an antiderivative of f, there is a boundary layer at $x=1$ $[x=0]$, that is, the solution $u=u(x,\varepsilon)$ of (S_1) satisfies $\lim_{\varepsilon\to 0}u(x,\varepsilon)=A$, $0\leqq x\leqq 1-\delta$ $[\lim_{\varepsilon\to 0}u(x,\varepsilon)=B$, $\delta\leqq x\leqq 1]$ $(0<\delta<1)$. This follows by virtue of the corresponding integral conditions (2.2), expressed now in terms of F. Consequently, if $F(A)=F(B)$ (the Rankine-Hugoniot relation) boundary layer behavior is impossible, and there must be a shock layer in (0, 1) joining A and B; indeed, the solution (S_1) satisfies

$$\lim_{\varepsilon\to 0} u(x,\varepsilon)=\begin{cases} A, & 0\leqq x\leqq x_0-\delta, \\ B, & x_0+\delta\leqq x\leqq 1, \end{cases} \tag{3.1}$$

where $x_0=f(B)/(f(B)-f(A))$ (in (0,1)) is the location of the shock layer. To see this, note that such a solution satisfies $u'(x)>0$ in $[0,1]$, since $u'(x)=(\text{const})\exp[\int^x f(u(s))\,ds]$. Thus (S_1) can be rewritten as $\varepsilon u''/u'=f(u)$, that is, $\varepsilon[\ln u']'=f(u)$. Integrating this equation from $x=0$ to $x=x_0$, and from $x=x_0$ to $x=1$, we obtain the two equations

$$\varepsilon\left[\ln u'(x_0)-\ln u'(0)\right]=\int_0^{x_0} f(u(s))\,ds, \tag{3.2}$$

$$\varepsilon\left[\ln u'(1)-\ln u'(x_0)\right]=\int_{x_0}^{1} f(u(s))\,ds. \tag{3.3}$$

Since $F(A)=F(B)$ it follows by integrating both sides of (S_1) from $x=0$ to $x=1$ and using the boundary conditions that $u'(0)=u'(1)$. Thus, adding (3.2) and (3.3) together, we have that

$$\int_0^{x_0} f(u(s))\,ds+\int_{x_0}^{1} f(u(s))\,ds=0. \tag{3.4}$$

By virtue of (3.1) $u\to A$ on $(0,x_0)$ and $u\to B$ on $(x_0,1)$ as $\varepsilon\to 0$, and so by the continuity of f and the Dominated Convergence Theorem, it follows from (3.4) that $f(A)x_0+f(B)(1-x_0)=0$, that is, $x_0=f(B)/(f(B)-f(A))$.

Simple arguments such as these allow us to study the occurrence of shock layers in the more general problem (BVP), to which we now turn.

4. Shock layer theory for (BVP). Let us consider finally the occurrence of shock layers in solutions of the problem

$$\begin{aligned} &\varepsilon u_i''=f_i(u_1,\cdots,u_n)u_i', && 0<x<1, \\ &u_i(0,\varepsilon)=A_i,\ u_i(1,\varepsilon)=B_i, && A_i<B_i, \end{aligned} \tag{BVP}$$

where the functions f_i are continuous for all $(u_1,\cdots,u_n)$ in $\Pi_{i=1}^n[A_i,B_i]$. Then (BVP) has a solution $\underset{\sim}{u}=(u_1,\cdots,u_n)$ of class $C^{(2)}([0,1])$ for each $\varepsilon>0$ that satisfies $A_i\leqq u_i(x,\varepsilon)\leqq B_i$ and $u_i'(x,\varepsilon)>0$ in $[0,1]$ for $i=1,\cdots,n$, since $\Pi_{i=1}^n[A_i,B_i]$ is an invariant region (cf. [19, Chap. 14]) and $u_x'=(\text{const})\exp[\int^x f_i(u_1(s),\cdots,u_n(s))\,ds]$.

In order to discuss our results in the simplest setting possible, we examine first the two-dimensional system

$$\begin{aligned} &\varepsilon u_1''=f_1(u_1,u_2)u_1', && u_1(0,\varepsilon)=A_1, && u_1(1,\varepsilon)=B_1, \\ &\varepsilon u_2''=f_2(u_1,u_2)u_2', && u_2(0,\varepsilon)=A_2, && u_2(1,\varepsilon)=B_2. \end{aligned} \tag{P}$$

The theory of §2 tells us, for instance, that if B_1, B_2 are such that

$$\begin{aligned} &f_1(B_1,\mu)<0 && \text{for all } \mu \text{ in } [A_2,B_2], \\ &f_2(\lambda,B_2)<0 && \text{for all } \lambda \text{ in } [A_1,B_1], \end{aligned} \tag{4.1}$$

then there is a solution of (P) as $\varepsilon \to 0$ with boundary layers (in both components) at $x=0$, provided

$$\begin{aligned} &\int_{B_1}^{\xi} f_1(s,A_2)\,ds>0 \quad \text{for } \xi \text{ in } [A_1,B_1), \\ &\int_{B_2}^{\eta} f_2(A_1,s)\,ds>0 \quad \text{for } \eta \text{ in } [A_2,B_2). \end{aligned} \tag{4.2}$$

In other words, we have that

$$\lim_{\varepsilon\to 0}(u_1(x,\varepsilon),u_2(x,\varepsilon))=(B_1,B_2) \quad \text{in } [\delta,1].$$

Similarly, if A_1, A_2 are such that

$$f_1(A_1,\mu)>0 \quad \text{and} \quad f_2(\lambda,A_2)>0 \tag{4.3}$$

for the ranges of λ and μ in (4.1), and

$$\begin{aligned} &\int_{A_1}^{\xi} f_1(s,B_2)\,ds>0 \quad \text{for } \xi \text{ in } (A_1,B_1], \\ &\int_{A_2}^{\eta} f_2(B_1,s)\,ds>0 \quad \text{for } \eta \text{ in } (A_2,B_2], \end{aligned} \tag{4.4}$$

then there is a solution of (P) as $\varepsilon\to 0$ with boundary layers (in both components) at $x=1$. In other words, we have that

$$\lim_{\varepsilon\to 0}(u_1(x,\varepsilon),u_2(x,\varepsilon))=(A_1,A_2) \quad \text{in } [0,1-\delta].$$

Finally under the appropriate combination of the integral inequalities there are solutions of (P) as $\varepsilon\to 0$ with a boundary layer at $x=0$ in one component and a boundary layer at $x=1$ in the other.

Suppose now that we look for a solution of (P) whose ith component has only a shock layer at x_i in (0,1), that is,

$$\lim_{\varepsilon\to 0} u_i(x,\varepsilon)=\begin{cases} A_i, & 0\leqq x\leqq x-\delta, \\ B_i, & x_i+\delta\leqq x\leqq 1. \end{cases} \tag{4.5}$$

Our basic result is contained in the following

THEOREM 1. *Assume that the reduced solutions* $\underset{\sim}{U}_L=(A_1,A_2)$ *and* $\underset{\sim}{U}_R=(B_1,B_2)$ *are attractive in the sense that the inequalities* (4.1) *and* (4.3) *obtain for the stated values of* λ *and* μ. *Assume also that*

$$F_1(B_1,A_2)\geqq F_1(A_1,A_2)$$

and

$$F_1(B_1,B_2)\leqq F_1(A_1,B_2),$$

where $F_1(\cdot,A_2$ *or* $B_2)$ *is an antiderivative of* $f_1(\cdot,A_2$ *or* $B_2)$ *and that*

$$F_2(A_1,B_2)\geqq F_2(A_1,A_2)$$

and

$$F_2(B_1,B_2)\leqq F_2(B_1,A_2),$$

where $F_2(A_1$ *or* $B_1,\cdot)$ *is an antiderivative of* $f_2(A_1$ *or* $B_1,\cdot)$*. Then for these boundary values there is a solution* (u_1,u_2) *of* (P) *as* $\varepsilon\to 0$ *whose ith component satisfies the limiting relation* (4.5).

Proof. For this choice of boundary values we know that there is a solution $\underset{\sim}{u}$ of (P) as a consequence of the fact that the rectangle $[A_1,B_1]\times[A_2,B_2]$ is invariant with respect to (P); cf. [19, Chap. 14]. However boundary layer behavior is impossible, since none of the boundary layer inequalities in (4.2) and (4.4) obtains by virtue of the inequalities involving F_1 and F_2. Consequently each component u_i of $\underset{\sim}{u}$ has a shock layer at x_i in $(0,1)$.

It remains for us to determine the locations x_i. We begin by noting that $\varepsilon\{\ln u_i'(1)-\ln u_i'(0)\}\to 0$ as $\varepsilon\to 0$, since $\underset{\sim}{U}_L$ *and* $\underset{\sim}{U}_R$ are constant vectors that satisfy the system (P) in $(0,1)$ and the boundary conditions at $x=0$ and $x=1$, respectively. Suppose, for definiteness, that $x_1\leqq x_2$. Then by arguing as in §3 we see that for $i=1,2$

$$\begin{aligned}(4.6)\qquad 0&\sim\int_0^1 f_i(u_1(s),u_2(s))\,ds\\ &\sim\int_0^{x_1} f_i(A_1,A_2)\,ds+\int_{x_1}^{x_2} f_i(B_1,A_2)\,ds+\int_{x_2}^1 f_i(B_1,B_2)\,dds\\ &=f_i(A_1,A_2)x_1+f_i(B_1,A_2)(x_2-x_1)+f_i(B_1,B_2)(1-x_2).\end{aligned}$$

Suppose we consider first the case when $x_1=x_2=x_0$, that is, both components have a shock layer at the same point. Then (4.6) implies that for $i=1,2$

$$[-f_i(A_1,A_2)+f_i(B_1,B_2)]x_0=f_i(B_1,B_2);$$

whence,

$$\begin{aligned}x_0&=f_1(B_1,B_2)/(f_1(B_1,B_2)-f_1(A_1,A_2))\\ &=f_2(B_1,B_2)/(f_2(B_1,B_2)-f_2(A_1,A_2)),\end{aligned}$$

provided that

$$f_1(A_1,A_2)f_2(B_1,B_2)=f_1(B_1,B_2)f_2(A_1,A_2).$$

The point x_0 lies in $(0,1)$ by virtue of the inequalities (4.1) and (4.3). Thus for such boundary values the problem (P) has a solution as $\varepsilon\to 0$ satisfying the limiting relation (4.5) with $x_1=x_2=x_0$.

Suppose that we consider next the case when there is a shock layer in u_1 at x_1 and a shock layer in u_2 at x_2 for $x_1<x_2$. The points x_1, x_2 are solutions of the linear system (cf. (4.6))

$$(4.7)\qquad [f_i(B_1,A_2)-f_i(A_1,A_2)]x_1+[f_i(B_1,B_2)-f_i(B_1,A_2)]x_2=f_i(B_1,B_2).$$

It follows that if

$$(4.8)\qquad f_1(A_1,A_2)f_2(B_1,B_2)<f_1(B_1,B_2)f_2(A_1,A_2)$$

the determinant of the coefficient matrix of (4.7) is positive by virtue of the inequalities (4.1) and (4.3). Thus (4.7) can be solved uniquely for x_1, x_2, and these values satisfy $0<x_1<x_2<1$. Indeed we have

$$x_1=\Delta^{-1}\det\begin{pmatrix}f_1(B_1,B_2) & f_1(B_1,B_2)-f_1(B_1,A_2)\\ f_2(B_1,B_2) & f_2(B_1,B_2)-f_2(B_1,A_2)\end{pmatrix}$$

and

$$x_2=\Delta^{-1}\det\begin{pmatrix} f_1(B_1,A_2)-f_1(A_1,A_2) & f_1(B_1,B_2) \\ f_2(B_1,A_2)-f_2(A_1,A_2) & f_2(B_1,B_2) \end{pmatrix}$$

for

$$\Delta=[f_1(B_1,A_2)-f_1(A_1,A_2)][f_2(B_1,B_2)-f_2(B_1,A_2)] \\ -[f_1(B_1,B_2)-f_1(B_1,A_2)][f_2(B_1,A_2)-f_2(A_1,A_2)]>0.$$

The remaining case when $x_2<x_1$ can be handled without difficulty, as well as the cases $A_i>B_i$, $i=1,2$, or $A_1>B_1$, $A_2<B_2$ or $A_1<B_1$, $A_2>B_2$ in (P). We turn finally to the general problem $(1\leqq i\leqq n)$

$$\text{(BVP)} \qquad \begin{gathered} \varepsilon u_i''=f_i(u_1,\cdots,u_n)u_i', \qquad 0<x<1, \\ u_i(0,\varepsilon)=A_i, \quad u_i(1,\varepsilon)=B_i, \quad A_i<B_i. \end{gathered}$$

The basic result for (BVP) is contained in the following theorem. It follows by noting that each of the corresponding boundary layer inequalities (cf. (2.2)) is violated and then proceeding as in the proof of Theorem 1.

THEOREM 2. *Assume that* $\underset{\sim}{U}_L=\underset{\sim}{A}=(A_1,\cdots,A_n)$ *and* $\underset{\sim}{U}_R=\underset{\sim}{B}=(B_1,\cdots,B_n)$ *are attractive in the sense that (cf.* §2) *for* $i=1,\cdots,n$

$$f_i(u_1,\cdots,A_i,\cdots,u_n)>0 \quad \textit{and} \quad f_i(u_1,\cdots,B_i,\cdots,u_n)<0$$

for all values of u_j *in* $[A_j,B_j]$ $(j\neq i)$. *Assume also that for* $i=1,\cdots,n$

$$F_i(A_1,\cdots,A_{i-1},A_i,A_{i+1},\cdots,A_n)\leqq F_i(A_1,\cdots,A_{i-1},B_i,A_{i+1},\cdots,A_n)$$

and

$$F_i(B_1,\cdots,B_{i-1},A_i,B_{i+1},\cdots,B_n)\geqq F_i(B_1,\cdots,B_{i-1},B_i,B_{i+1},\cdots,B_n),$$

where $F_i(A_1$ *or* $B_1,\cdots,A_{i-1}$ *or* $B_{i-1},\cdot,A_{i+1}$ *or* $B_{i+1},\cdots,A_n$ *or* $B_n)$ *is an antiderivative of* $f_i(A_1$ *or* $B_1,\cdots,A_{i-1}$ *or* $B_{i-1},\cdot,A_{i+1}$ *or* $B_{i+1},\cdots,A_n$ *or* $B_n)$. *Then there is a solution of* (BVP) *as* $\varepsilon\to 0$ *having a shock layer in each component.*

The simplest situation occurs when each shock layer is located at the same point x_0 in $(0,1)$. By proceeding as before, we see that for $i=1,\cdots,n$

$$0\sim\int_0^1 f_i(\underset{\sim}{u}(s))\,ds\sim\int_0^{x_0} f_i(\underset{\sim}{A})\,ds+\int_{x_0}^1 f_i(\underset{\sim}{B})\,ds,$$

and so

$$x_0=f_i(\underset{\sim}{B})/(f_i(\underset{\sim}{B})-f_i(\underset{\sim}{A})),$$

provided

$$\text{(4.9)} \qquad f_i(\underset{\sim}{B})/(f_i(\underset{\sim}{B})-f_i(\underset{\sim}{A}))=f_j(\underset{\sim}{B})/(f_j(\underset{\sim}{B})-f_j(\underset{\sim}{A}))$$

for all $i,j=1,\cdots,n$. The consistency conditions in (4.9) reduce to $\binom{n}{2}=n(n-1)/2$ equations. Thus it is usually the case that the components of a solution of (BVP) do not all have a shock layer at the same point, but rather different components have shock layers at different points. The locations x_i of the shocks are found as before by solving a linear system of n equations. Unfortunately the prohibitive number of algebraic manipulations prevents us from providing explicit conditions such as (4.8) for the general problem (BVP), although the analysis is straightforward and could be performed numerically.

5. Examples. In this final section we illustrate some of the theory developed above.

Example 1. Consider first the problem

$$\begin{aligned} &\varepsilon u_i''=-u_ig_i(u_1,\cdots,u_{i-1},u_{i+1},\cdots,u_n)u_i', \qquad 0<x<1,\\ &u_i(0,\varepsilon)=A_i,\quad u_i(1,\varepsilon)=B_i,\quad A_i<0,\quad B_i>0, \end{aligned}\tag{E1}$$

where $g_i>0$ for all values of u_j in $[A_j,B_j]$ $(j\neq i)$. We anticipate that each component of the solution behaves very much like the solution of the scalar problem $\varepsilon y''=-yy'$, $y(0,\varepsilon)=A$, $y(1,\varepsilon)=B$ (cf. §3 or [16, Chap. 1] where the exact solution is given). In particular, if $A=-B<0$ then $f(A)>0$, $f(B)<0$ and $F(A)=F(B)$, for $f(y)=-y$ and $F(y)=-y^2/2$. Consequently there is a solution $y=y(x,\varepsilon)$ as $\varepsilon\to 0$ with a shock layer at $x_0=f(B)/(f(B)-f(A))=1/2$ joining $U_L=A(x<x_0)$ and $U_R=B$ $(x>x_0)$. For the problem (E1) we see that

$$f_i(u_1,\cdots,A_i,\cdots,u_n)>0 \quad\text{and}\quad f_i(u_1,\cdots,B_i,\cdots,u_n)<0\tag{5.1}$$

for all values of u_j in $[A_j,B_j]$, since $g_i>0$ there. If we assume that $A_i=-B_i<0$ for $i=1,\cdots,n$, then

$$\begin{aligned} F_i(A_1,\cdots,A_i,\cdots,A_n)&=F_i(A_1,\cdots,B_i,\cdots,A_n),\\ F_i(B_1,\cdots,A_i,\cdots,B_n)&=F_i(B_1,\cdots,B_i,\cdots,B_n), \end{aligned}\tag{5.2}$$

since $F_i(u_1,\cdots,u_{i-1},\gamma,u_{i+1},\cdots,u_n)=-(\gamma^2/2)g_i(u_1,\cdots,u_{i-1},u_{i+1},\cdots,u_n)$. Thus we know that (E1) has a solution with a shock layer in each component. It only remains to locate the position x_i of the ith shock. Suppose first that

$$g_i(A_1,\cdots,A_n)g_j(B_1,\cdots,B_n)=g_i(B_1,\cdots,B_n)g_j(A_1,\cdots,A_n)\tag{5.3}$$

for $i\neq j$. Then it follows that $(x_0=)$ $f_i(\underset{\sim}{B})/(f_i(\underset{\sim}{B})-f_i(\underset{\sim}{A}))=f_j(\underset{\sim}{B})/(f_j(\underset{\sim}{B})-f_j(\underset{\sim}{A}))$ for $i\neq j$, and so there is a shock layer in each component at $x=x_0$ in $(0,1)$ joining A_i $(x<x_0)$ and B_i $(x>x_0)$. However if the $\binom{n}{2}$ equalities in (5.3) do not obtain, then different components of the solution of (E1) have shock layers at different points in $(0,1)$. Rather than discuss this situation in its full generality, we turn to a much simpler special case.

Example 2. Consider then the two-dimensional problem on $(0,1)$

$$\begin{aligned} &\varepsilon u_1''=-u_1(1-u_2)u_1',\quad u_1(0,\varepsilon)=A_1,\quad u_1(1,\varepsilon)=B_1,\\ &\varepsilon u_2''=-u_2(1+u_1)u_2',\quad u_2(0,\varepsilon)=A_2,\quad u_2(1,\varepsilon)=B_2, \end{aligned}\tag{E2}$$

where $A_i=-B_i<0$ and $|A_i|<1$ for $i=1,2$. Under these restrictions on the boundary values the relations (5.1) and (5.2) certainly hold, but $g_1(\underset{\sim}{A})g_2(\underset{\sim}{B})=(1-A_2)(1+B_2)>(1-B_2)(1+A_1)=g_1(\underset{\sim}{B})g_2(\underset{\sim}{A})$, which implies that $f_1(\underset{\sim}{A})f_2(\underset{\sim}{B})<f_2(\underset{\sim}{A})f_1(\underset{\sim}{B})$. Consequently there is a solution (u_1,u_2) of (E2) as $\varepsilon\to 0$ such that u_1 $[u_2]$ has a shock layer at x_1 $[x_2]$ in $(0,1)$ with $x_1<x_2$. Proceeding as in §4 we find that $x_1=2B_1B_2(B_1+1)/\Delta$ and $x_2=2B_1B_2(2B_1+B_2+1)/\Delta$, for $\Delta=4B_1B_2(B_1+B_2+1)$, that is,

$$x_1=(B_1+1)/[2(B_1+B_2+1)]$$

and

$$x_2=(2B_1+B_2+1)/[2(B_1+B_2+1)].$$

Acknowledgments. The author thanks the National Science Foundation for its financial support and the typist, Ida Zalac, for her secretarial support.

REFERENCES

[1] K. W. CHANG AND F. A. HOWES, *Nonlinear Singular Perturbation Phenomena: Theory and Applications*, Springer-Verlag, New York, 1984.

[2] E. A. CODDINGTON AND N. LEVINSON, *A boundary value problem for a nonlinear differential equation with a small parameter*, Proc. Amer. Math. Soc., 3 (1952), pp. 75–81.

[3] R. J. DIPERNA, *Singularities and oscillations in solutions to conservation laws*, Physica, 12D (1984), pp. 363–368.

[4] F. A. HOWES, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Memoirs Amer. Math. Soc., 203 (1978).

[5] ——, *An analytical treatment of the formation of one-dimensional steady shock waves in uniform and diverging ducts*, J. Comp. Appl., Math., 10 (1984), pp. 195–201.

[6] ——, *Asymptotic structures in nonlinear dissipative and dispersive systems*, Physica, 12D (1984), pp. 382–390.

[7] ——, *Boundary layer behavior in perturbed second-order systems*, J. Math. Anal. Appl., 104 (1984), pp. 467–476.

[8] ——, *Multi-dimensional reaction-convection-diffusion equations*, Proc. Conference on Differential Equations, Dundee, 1984, Lecture Notes in Mathematics, Springer-Verlag, New York, in press.

[9] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[10] P. D. LAX, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS Regional Conference in Applied Mathematics 11, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

[11] T. P. LIU, *Quasilinear hyperbolic systems*, Commun. Math. Phys., 68 (1979), pp. 141–172.

[12] J. LORENZ, *Nonlinear boundary value problems with turning points and properties of difference schemes*, Lecture Notes in Mathematics 942, Springer-Verlag, New York, 1982, pp. 150–169.

[13] M. A. O'DONNELL, *Boundary and interior layer behavior in singularly perturbed nonlinear systems*, Ph.D. dissertation, Univ. California, Davis, 1983.

[14] ——, *Boundary and corner layer behavior in singularly perturbed semilinear systems of boundary value problems*, this Journal, 15 (1984), pp. 317–332.

[15] ——, *Turning point behavior in singularly perturbed nonlinear systems*, Nonlinear Analysis, in press.

[16] R. E. O'MALLEY, JR., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[17] ——, *On multiple solutions of singularly perturbed systems in the conditionally stable case*, in Singular Perturbations and Asymptotics, R. E. Meyer and S. V. Parter, eds., Academic Press, New York, 1980, pp. 87–108.

[18] ——, *Shock and transition layers for singularly perturbed second order vector systems*, SIAM J. Appl. Math., 43 (1983), pp. 935–943.

[19] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[20] G. B. WHITHAM, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

010

SINGULAR PERTURBATIONS FOR NONLINEAR HYPERBOLIC-PARABOLIC PROBLEMS*

AHMED BENAOUDA[†] AND MONIQUE MADAUNE TORT[†]

Abstract. A boundary value problem for a semilinear hyperbolic equation with a small parameter ε and its reduced problem of parabolic type are considered. It is proved that the solution u_ε of the initial problem is approximated for $\varepsilon \to 0$ by the solution u of the reduced problem in Sobolev spaces. Estimates of the difference-norm are given in function of ε and are improved by using boundary layer correctors.

Key words. singular perturbations, semilinear hyperbolic equations, boundary value problems

AMS(MOS) subject classifications. Primary 35B25, 35L20

Introduction. We consider "hyperbolic-parabolic" singular perturbation problems for semi-linear equations of the type:

$$L_2^{(\varepsilon)}u_\varepsilon + L_1 u_\varepsilon + G u_\varepsilon = f, \qquad (t,x) \in]0,T[\times\Omega \tag{0.1}$$

where T is a real, $T > 0$, Ω is a bounded open set in $\mathbf{R}^n$, $\varepsilon > 0$,

$$L_2^{(\varepsilon)}u = \varepsilon\frac{\partial^2 u}{\partial t^2} - \Delta u, \tag{0.2}$$

$$L_1 u = a(t,x)\frac{\partial u}{\partial t} + \sum_{k=1}^{n} b_k(t,x)\frac{\partial u}{\partial x_k}, \tag{0.3}$$

G: $\mathbf{R} \to \mathbf{R}$ is a continuous function which satisfies a monotonicity condition and a growth limitation at ∞ (the precise hypotheses are given in §1). Two examples of such functions G are:

(i) G is a Lipschitzian function,

(ii) $Gu = |u|^\rho u$, $\rho > 0$.

The questions of singular perturbations of "hyperbolic-parabolic" type do not seem to have been much studied. However we may mention the works of Zlamal [20] (1960), Kisynski [4] (1963), Nazarkulova–Pankov [17] (1971), Lenjuk–Fedoruk [6] (1972), Muradov [15] (1978), Muradov–Gasanov [16] (1974), concerning linear problems to which we may add the study of nonlinear problems done in Muradov [14] (1974) and N. A. Lar'kin [5] (1980). In [14] Muradov considers an equation like (0.1) with $\Omega \subset \mathbf{R}$ where $\partial^2 u/\partial x^2$ and $\partial u/\partial x$ may have discontinuous coefficients, and in [5] N. A. Lar'kin studies a hyperbolic regularization of a one-sided problem for the Burgers equation in Sobolev spaces when $\Omega \subset \mathbf{R}$. Moreover two recent papers point out the interest of such problems which model oscillator phenomena in a highly viscous medium similarly to Prandtl's model for ordinary differential equation (Hsiao–Weinacht [2] (1979), [3] (1983)).

* Received by the editors February 5, 1985, and in revised form September 1, 1985.

[†] Université de Pau, Faculté des Sciences Exactes, Département de Mathématiques, Pau, France 64000.

Hence the mathematical model:

$$P_\varepsilon \quad \begin{cases} \varepsilon^2 \dfrac{\partial^2 u_\varepsilon}{\partial t^2} - \dfrac{\partial^2 u_\varepsilon}{\partial x^2} + \dfrac{\partial u_\varepsilon}{\partial t} + G u_\varepsilon = 0, & (t,x) \in]0, +\infty[\times \mathbf{R}, \\ u_\varepsilon(0,x) = u_0(x), & \varepsilon \dfrac{\partial u}{\partial t}(0,x) = u_1(x). \end{cases}$$

The study of this problem is given in [3] when G belongs to $C^r(r \geq 8)$ and satisfies $G'(x) > 0$, $\forall x \in \mathbf{R}$. An asymptotic representation of u_ε is given with an explicit calculation of the boundary layer correctors which are computed to all order when the reduced problem has a bounded classical solution on $[0, T] \times \mathbf{R}$.

Here, we study the variational point of view under weak hypotheses when x belongs to an open set of $\mathbf{R}^n$, in order to obtain the convergence in Sovolev spaces of u_ε to u, the solution of a problem relative to the equation:

$$L_2^{(0)} u + L_1 u + G u = f.$$

This paper is an extension of a part of the work [1].

1. Notation–hypotheses. Ω is a bounded open set in $\mathbf{R}^n$, of class $\eta^{(1),1}$ (Nečas [18]), Γ is the boundary of Ω, T is a given real, $T > 0$. We set $Q =]0, T[\times \Omega$, $\Sigma = [0, T] \times \Gamma$. For each real p, $2 \leq p \leq +\infty$, we note $|\ |_p$ (resp. $\|\ \|_p$) the usual norm of $L^p(\Omega)$ (resp. $W^{1,p}(\Omega)$). We represent the inner product in $L^2(\Omega)$ by $(\cdot,\cdot)$; we keep the same notation for the duality between $H^{-1}(\Omega)$, $H_0^1(\Omega)$ and $L^p(\Omega)$, $L^{p'}(\Omega)(1/p + 1/p' = 1)$. $(u, v) \mapsto \alpha(u, v)$ is the bilinear form defined by

$$\alpha(u,v) = \int_\Omega \vec{\nabla} u, \vec{\nabla} v \, dx.$$

The derivatives of u in the sense of vector-valued distributions on $]0, T[$ are represented by u', u'', $\cdots$.

We recall the hyperbolic problem that we consider:

$$P_\varepsilon \quad \begin{cases} L_2^{(\varepsilon)} u_\varepsilon + L_1 u_\varepsilon + G u_\varepsilon = f, & (t,x) \in Q, \\ u_\varepsilon(0,x) = u_0(x), \quad u_\varepsilon'(0,x) = u_1(x), & x \in \Omega, \\ u_\varepsilon|_\Sigma = 0. \end{cases}$$

We are going to study the behavior of u_ε when $\varepsilon \to 0_+$ ($L_2^{(\varepsilon)}$ and L_1 are respectively defined by (0.2) and (0.3)). This work is undertaken when the following hypothesis is satisfied:

$H_{1.1}$ (i) $u_0 \in H_0^1(\Omega)$, $u_1 \in L^2(\Omega)$, $f \in L^2(Q)$,

(ii) $a \in W^{1,\infty}(Q) \cap C^0(\bar{Q})$ and $\inf_{\bar{Q}} a = \delta > 0$, $b_k \in W^{1,\infty}(Q)$, $1 \leq k \leq n$,

(iii) G is a continuous function from $\mathbf{R}$ to $\mathbf{R}$ such that $Gv = Fv + \theta v$ where θ is a real constant and F a function from $\mathbf{R}$ to $\mathbf{R}$ satisfying: $F0 = 0$ and $\exists p \in \mathbf{R}$, $p \geq 2$, $\exists \beta > 0$, $\exists \gamma > 0$ such that:

$$\beta |u - v|^p \leq (Fu - Fv)(u - v) \leq \gamma \left(|u|^{p-2} + |v|^{p-2} \right) |u - v|^2, \tag{1.1}$$

(iv) n and p satisfy the inequality:

$$n \leq 2 + \frac{2}{p-2} \qquad \left(n \in \mathbf{N}^* \text{ if } p = 2 \right).$$

The condition $\mathrm{H}_{1.1}$(iv) implies the algebraic and topological inclusion:

$$(1.2)\qquad H_0^1(\Omega)\hookrightarrow L^q(\Omega)\quad \forall q\in[2,2(p-1)].$$

A comment on hypothesis $\mathrm{H}_{1.1}$(iii) *and examples of functions G satisfying* $\mathrm{H}_{1.1}$(iii): In fact (1.1) implies that the nonlinear part of G, denoted by F, satisfies the condition $F'(z)\geq 0$, $\forall z\in\mathbf{R}$ (instead of $F'(z)>0$, $\forall z\in\mathbf{R}$ as in [3]). So here F' may be equal to zero in a subset of $\mathbf{R}$ and G is not necessary a monotone function as θ may be a nonpositive real.

(1) A first example of such a function is given by every Lipschitzian function G: $\mathbf{R}\to\mathbf{R}$ such that $G0=0$ (for example $Gv=\sin v$). Indeed if we denote by l a Lipschitz constant of G, we may choose $\theta=-l$ and $p=2$.

(2) Another classical example is given by $G(u)=|u|^\rho u$, $\rho\geq 0$. Indeed, $\mathrm{H}_{1.1}$(iii) is verified for $\theta=0$, $F=G$, and $p=\rho+2$; then the existence of a constant β is ensured (see Lions [8, p. 200]).

A result on existence and uniqueness. We associate with the problem P_ε the variational problem:

$$\mathcal{T}_\varepsilon\begin{cases}(1.3)\quad \varepsilon(u_\varepsilon'',v)+\alpha(u_\varepsilon,v)+(L_1u_\varepsilon,v)+(Gu_\varepsilon,v)=(f,v)\\ \qquad\qquad \forall v\in H_0^1(\Omega),\ \text{a.e. } s\in\,]0,T[,\\ (1.4)\quad u_\varepsilon\in L^\infty(0,T;H_0^1(\Omega)),\qquad u_\varepsilon'\in L^\infty(0,T;L^2(\Omega)),\\ (1.5)\quad u_\varepsilon(0,x)=u_0(x),\qquad u_\varepsilon'(0,x)=u_1(x)\end{cases}$$

for which we have the following property:

THEOREM 1.1. *Under the hypothesis* $\mathrm{H}_{1.1}$, *for each* $\varepsilon>0$, *the problem* $\mathcal{T}_\varepsilon$ *has a unique solution.*

The proof of this theorem is founded on Galerkin's method and is similar to the one given in Lions [7, Thm. 1.1] for the case $Gu=|u|^\rho u$ or in Saut [19] for the case of a Lipschitzian function G. We remark that the conditions (1.4) imply thanks to the results of Lions–Magenes [9] that u_ε is continuous from $[0,T]$ to $L^2(\Omega)$ and that u_ε' is continuous from $[0,T]$ to $H^{-1}(\Omega)$, so (1.5) has a sense.

A result of regularity. We introduce the hypothesis:

$\mathrm{H}_{1.2}$: $\mathrm{H}_{1.1}$ with $u_0\in H^2(\Omega)\cap H_0^1(\Omega)$, $u_1\in H_0^1(\Omega)$, $f'\in L^2(Q)$.

Then we have

THEOREM 1.2. *Under the hypothesis* $\mathrm{H}_{1.2}$, *for each* $\varepsilon>0$, *the solution of the problem* $\mathcal{T}_\varepsilon$ *is such that*:

$$u_\varepsilon\in L^\infty(0,T;H^2(\Omega)\cap H_0^1(\Omega)),\quad u_\varepsilon'\in L^\infty(0,T;H_0^1(\Omega)),\quad u_\varepsilon''\in L^\infty(0,T;L^2(\Omega)).$$

The proof of this theorem is analogous to a proof given in Lions [7, Thm. 1.3] or in Saut [19]. The only difficulty is to justify the differentiation of the nonlinear term. Now, the hypothesis $\mathrm{H}_{1.1}$(iii) implies that F is a local Lipschitzian function. Therefore $F'x$ exists almost everywhere on $\mathbf{R}$ (see for example Marcus–Mizel [13]) and verifies thanks to (1.1)

$$0\leq F'x\leq 2\gamma|x|^{p-2}\quad \text{a.e. } x\in\mathbf{R}.$$

Moreover thanks to the results of [13, Thm. 2.1 and property 1.5] we may assert that: for every function $u\in L^\infty(0,T;W_0^{1,2(p-1)}(\Omega))$ such that $u'\in L^\infty(0,T;L^{2(p-1)}(\Omega))$, we have $Fu\in L^\infty(0,T;H_0^1(\Omega))$, $(Fu)'\in L^\infty(0,T;L^2(\Omega))$ and the equalities:

$$\frac{\partial}{\partial x_k}(Fu)=(F'u)\frac{\partial u}{\partial x_k},\quad 1\leq k\leq n,\qquad (Fu)'=(F'u)u'.$$

Scheme of this paper. In the second paragraph we establish a priori estimates on u_ε and its derivatives $\partial u_\varepsilon/\partial x_k$ $(1\le k\le n)$, u'_ε by methods similar to those done in Madaune and Genet [12] and Madaune [10]. These estimates are sufficient to establish the convergence of u_ε when $\varepsilon\to 0_+$, by compactness arguments in the third paragraph; some results of strong convergence and estimates of $(u_\varepsilon-u)$ are also given in this paragraph. Lastly some remarks are gathered in the fourth paragraph in particular when the operator L_1 is nonlinear, Ω being a segment of $\mathbf{R}$.

2. A priori estimates.

THEOREM 2.1. *Under the hypothesis* $\mathrm{H}_{1.1}$, *the solution* u_ε *of the problem* $\mathscr{T}_\varepsilon$ *is such that*:

$$\|u_\varepsilon\|_{L^\infty(0,T;H_0^1(\Omega))}+\sqrt{\varepsilon}\,|u'_\varepsilon|_{L^\infty(0,T;L^2(\Omega))}+|u'_\varepsilon|_{L^2(Q)}\le C.K_1 \tag{2.1}$$

where

$$K_1^2=|f|^2_{L^2(Q)}+\|u_0\|_2^2+|u_0|_p^p+\varepsilon|u_1|_2^2 \tag{2.2}$$

and C is a positive constant independent of f, u_0, u_1 *and* ε.

Proof. We first establish the estimates (2.1) under the hypothesis $\mathrm{H}_{1.2}$; then we use a method of regularization to prove the theorem.

(1) *With assumption* $\mathrm{H}_{1.2}$. We consider the equality (1.3):

$$\varepsilon(u''_\varepsilon,v)+\alpha(u_\varepsilon,v)+(au'_\varepsilon,v)+\left(\sum_{k=1}^{n}b_k\frac{\partial u_\varepsilon}{\partial x_k},v\right)+(Gu_\varepsilon,v)=(f,v)$$

$$\forall v\in H_0^1(\Omega)\quad \text{a.e. } s\in]0,T[.$$

We put $v=2u'_\varepsilon(s)$; that is possible thanks to Theorem 1.2. We obtain

$$\frac{d}{dt}\left[\varepsilon|u'_\varepsilon|_2^2+\alpha(u_\varepsilon,u_\varepsilon)\right]+2|\sqrt{a}\,u'_\varepsilon|_2^2+2(Fu_\varepsilon,u'_\varepsilon) \tag{2.3}$$

$$=2(f,u'_\varepsilon)-2\left(\sum_{k=1}^{n}b_k\frac{\partial u_\varepsilon}{\partial x_k},u'_\varepsilon\right)-2\theta(u_\varepsilon,u'_\varepsilon).$$

Thanks to Hölder's inequality, the second member of (2.3) may be bounded by:

$$\delta|u'_\varepsilon|_2^2+\frac{3}{\delta}|f|_2^2+\frac{3}{\delta}\left(\sum_{k=1}^{n}|b_k|_{L^\infty(Q)}\right)^2\alpha(u_\varepsilon,u_\varepsilon)+\frac{3}{\delta}\theta^2|u_\varepsilon|_2^2.$$

Then, we integrate (2.3) from 0 to t. We denote by $\tilde{F}$ the primitive of F which vanishes at 0; we obtain:

$$\varepsilon|u'_\varepsilon(t)|_2^2+\alpha(u_\varepsilon(t),u_\varepsilon(t))+\delta\int_0^t|u'_\varepsilon(s)|_2^2ds+2\int_\Omega\tilde{F}u_\varepsilon(t)\,dx$$

$$\le\varepsilon|u_1|_2^2+\alpha(u_0,u_0)+2\int_\Omega\tilde{F}u_0\,dx+\frac{3}{\delta}|f|^2_{L^2(Q)}+M\int_0^t\|u_\varepsilon(s)\|_2^2ds$$

where

$$M=\frac{3}{\delta}\max\left(\theta^2,\left(\sum_{k=1}^{n}|b_k|_{L^\infty(Q)}\right)^2\right).$$

As the inequality (1.1) of hypothesis $\mathrm{H}_{1.1}$(iii) implies:

$$\forall v \in \mathbf{R}, \qquad \frac{\beta}{p}|v|^p \leq \tilde{F}v \leq \frac{\gamma}{p}|v|^p, \tag{2.4}$$

it is easy to deduce the estimates (2.1) from Poincare's inequality and Gronwall's lemma.

(2) *With assumption* $\mathrm{H}_{1.1}$. Let $W = L^2(Q) \times H_0^1(\Omega) \times L^2(\Omega)$ provided with the product topology and let $(f, u_0, u_1) \in W$. There is a sequence $(f_\mu; u_{0,\mu}; u_{1,\mu})$ which satisfies $H_{1,2}$ and converges to (f, u_0, u_1) in W. The solution $u_{\varepsilon,\mu}$ of the problem $\mathscr{T}_{\varepsilon,\mu}$ associated with $(f_\mu; u_{0,\mu}; u_{1,\mu})$ satisfies the estimates:

$$|u_{\varepsilon,\mu}|_{L^\infty(0,T;\, H_0^1(\Omega))} + \sqrt{\varepsilon}\,|u'_{\varepsilon,\mu}|_{L^\infty(0,T;\, L^2(\Omega))} + |u'_{\varepsilon,\mu}|_{L^2(Q)} \leq CK_{1,\mu} \tag{2.5}$$

where $K_{1,\mu} = |f_\mu|^2_{L^2(Q)} + \|u_{0,\mu}\|_2^2 + |u_{0,\mu}|_p^p + \varepsilon |u_{1,\mu}|_2^2$ and C is a constant independent of μ, f, u_0, u_1 and ε. Taking into account the convergence properties of the sequence $(f_\mu; u_{0,\mu}; u_{1,\mu})$, $K_{1,\mu}$ may be bounded independently of μ. Then, we can extract from the sequence $\{u_{\varepsilon,\mu}\}_\mu$ a subsequence still written $\{u_{\varepsilon,\mu}\}$ such that:

$$u_{\varepsilon,\mu} \rightharpoonup v_\varepsilon \quad \text{in } L^\infty(0,T; H_0^1(\Omega)) \text{ weak star}, \tag{2.6}$$

$$u'_{\varepsilon,\mu} \rightharpoonup v'_\varepsilon \quad \text{in } L^\infty(0,T; L^2(\Omega)) \text{ weak star}. \tag{2.7}$$

We deduce from (2.6) and (2.7) that $u_{\varepsilon,\mu}$ converges weakly in $H^1(Q)$ to v_ε. Consequently, we may choose the subsequence $\{u_{\varepsilon,\mu}\}$ such that $u_{\varepsilon,\mu}$ converges to v_ε in $L^2(Q)$ and almost everywhere in Q. Then, thanks to the continuity of G and to the condition (1.1), we may assert that $Gu_{\varepsilon,\mu}$ converges weakly to Gv_ε in $L^2(Q)$. Moreover $\varepsilon u''_{\varepsilon,\mu} = f_\mu + \Delta u_{\varepsilon,\mu} - L_1 u_{\varepsilon,\mu} - Gu_{\varepsilon,\mu}$ is bounded in $L^2(0,T; H^{-1}(\Omega))$ thanks to (2.5); so $\varepsilon u''_{\varepsilon,\mu}$ converges weakly in $L^2(0,T; H^{-1}(\Omega))$ to $\varepsilon v''_\varepsilon$. Hence, we can take the limit on μ in the equation satisfied by $u_{\varepsilon,\mu}$ and in the initial conditions. We deduce $v_\varepsilon = u_\varepsilon$ which gives us thanks to (2.5) the estimates of the theorem.

THEOREM 2.2. *Under hypothesis* $\mathrm{H}_{1.2}$, *the solution* u_ε *of the problem* $\mathscr{T}_\varepsilon$ *is such that*:

$$\varepsilon |u''_\varepsilon|_{L^\infty(0,T;\, L^2(\Omega))} + \sqrt{\varepsilon}\,|u''_\varepsilon|_{L^2(Q)} + \sqrt{\varepsilon}\,\|u'_\varepsilon\|_{L^\infty(0,T;\, H_0^1(\Omega))} \leq C \tag{2.8}$$

where C is a constant independent of ε.

Proof. To obtain the estimates (2.8), the idea is to differentiate the equality (1.3) with respect to t and then, to put $v = u''_\varepsilon(s)$. But we are not allowed to do so because u_ε is not smooth enough, even under hypothesis $\mathrm{H}_{1.2}$ (see Theorem 1.2). Therefore, we use a method of difference quotients.

Let h be a positive real, which will tend to 0 and let $s \in]0, T-h[$. We consider the equality (1.3) at times s and $s+h$ and we substract these two equalities. Then if for each function w: $Q \to \mathbf{R}$ we denote by $\tau_h w(s)$ the difference quotient $(1/h)(w(s+h) - w(s))$, we obtain:

(2.9)

$$\varepsilon(\tau_h u''_\varepsilon(s), v) + \alpha(\tau_h u_\varepsilon(s), v) + (a(s+h)\tau_h u'_\varepsilon(s), v) + \left(\sum_{k=1}^{n} b_k(s+h)\left[\tau_h \frac{\partial}{\partial x_k} u_\varepsilon(s)\right], v\right)$$

$$= (\tau_h f(s), v) - ([\tau_h a(s)]\, u'_\varepsilon(s), v) - \left(\sum_{k=1}^{n} [\tau_h b_k(s)] \frac{\partial u_\varepsilon}{\partial x_k}(s), v\right) - (\tau_h G u_\varepsilon(s), v)$$

$$\forall v \in H_0^1(\Omega), \text{ a.e. } s \in]0, T-h[.$$

Thanks to the results of Theorem 1.2, we may put $v=2\tau_h u_\varepsilon'(s)$ in (2.9); we have:

(2.10)

$$\frac{d}{ds}\left[\varepsilon|\tau_h u_\varepsilon'(s)|_2^2+a(\tau_h u_\varepsilon(s),\tau_h u_\varepsilon(s))\right]+2\delta|\tau_h u_\varepsilon'(s)|_2^2$$

$$\leq 2(\tau_h f(s),\tau_h u_\varepsilon'(s))-2([\tau_h a(s)]u_\varepsilon'(s),\tau_h u_\varepsilon'(s))-2\left(\sum_{k=1}^{n}[\tau_h b_k(s)]\frac{\partial u_\varepsilon}{\partial x_k}(s),\tau_h u_\varepsilon'(s)\right)$$

$$-2\left(\sum_{k=1}^{n}b_k(s+h)\left[\tau_h\frac{\partial}{\partial x_k}u_\varepsilon(s)\right],\tau_h u_\varepsilon'(s)\right)-2(\tau_h G u_\varepsilon(s),\tau_h u_\varepsilon'(s))\quad \text{a.e. } s\in\,]0,T-h[.$$

First, we bound the second member of (2.10) as follows (where μ denotes a constant that we will choose later):

$$|2(\tau_h f(s),\tau_h u_\varepsilon'(s))|\leq\mu|\tau_h f(s)|_2^2+\frac{1}{\mu}|\tau_h u_\varepsilon'(s)|_2^2,$$

$$|2([\tau_h a(s)]u_\varepsilon'(s),\tau_h u_\varepsilon'(s))|\leq\mu|a'|_{L^\infty(Q)}^2|u_\varepsilon'(s)|_2^2+\frac{1}{\mu}|\tau_h u_\varepsilon'(s)|_2^2,$$

$$\left|2\left(\sum_{k=1}^{n}[\tau_h b_k(s)]\frac{\partial u_\varepsilon}{\partial x_k}(s),\tau_h u_\varepsilon'(s)\right)\right|\leq\mu\left(\sum_{k=1}^{n}|b_k'|_{L^\infty(Q)}\right)^2\|u_\varepsilon(s)\|_2^2+\frac{1}{\mu}|\tau_h u_\varepsilon'(s)|_2^2,$$

$$\left|2\left(\sum_{k=1}^{n}b_k(s+h)\left[\tau_h\frac{\partial}{\partial x_k}u_\varepsilon(s)\right],\tau_h u_\varepsilon'(s)\right)\right|$$

$$\leq\mu\left(\sum_{k=1}^{n}|b_k|_{L^\infty(Q)}\right)^2\|\tau_h u_\varepsilon(s)\|_2^2+\frac{1}{\mu}|\tau_h u_\varepsilon'(s)|_2^2,$$

and finally, thanks to inequality (1.1), the nonlinear term is bounded by

$$|2(\tau_h G u_\varepsilon(s),\tau_h u_\varepsilon'(s))|\leq 2\gamma\int_\Omega\left(|u_\varepsilon(s+h)|^{p-2}+|u_\varepsilon(s)|^{p-2}\right)|\tau_h u_\varepsilon(s)||\tau_h u_\varepsilon'(s)|\,dx$$

$$+2|\theta||(\tau_h u_\varepsilon(s),\tau_h u_\varepsilon'(s))|;$$

therefore by using Hölder's inequality (with $(p-2)/(2(p-1))+1/(2(p-1))+\frac{1}{2}=1$), Theorem 2.1 and (1.2) we have:

$$|2(\tau_h G u_\varepsilon(s),\tau_h u_\varepsilon'(s))|\leq k_1\|\tau_h u_\varepsilon(s)\|_2^2+\frac{1}{\mu}|\tau_h u_\varepsilon'(s)|_2^2,$$

where k_1 is a positive constant independent of ε and h. Then, we integrate (2.10) from 0 to $t\in\,]0,T-h[$ and we set $\mu=\frac{5}{\delta}$. We obtain, thanks to the estimates (2.1) and hypothesis $H_{1.1}$(ii):

$$\text{(2.11)}\quad \varepsilon|\tau_h u_\varepsilon'(t)|_2^2+\|\tau_h u_\varepsilon(t)\|_2^2+\delta\int_0^t|\tau_h u_\varepsilon'(s)|_2^2ds\leq c_1(h)+c_2+c_3\int_0^t\|\tau_h u_\varepsilon(s)\|_2^2ds$$

where c_2, c_3 are two positive constants independent of ε and h, $c_1(h)$ is equal to:

$$c_1(h)=\frac{5}{\delta}\int_0^t|\tau_h f(s)|_2^2ds+\varepsilon|\tau_h u_\varepsilon'(0)|_2^2+a(\tau_h u_\varepsilon(0),\tau_h u_\varepsilon(0)).$$

It follows from hypothesis $\mathrm{H}_{1.2}$, Theorem 1.2 and equality (0.1) that $\varepsilon c_1(h)$ is bounded independently of h and ε. So, we deduce from (2.11) by Gronwall's lemma, the estimates:

$$\varepsilon|\tau_h u_\varepsilon'|_{L^\infty(0,T;L^2(\Omega))}+\sqrt{\varepsilon}\,|\tau_h u_\varepsilon'|_{L^2(Q)}+\sqrt{\varepsilon}\,\|\tau_h u_\varepsilon\|_{L^\infty(0,T;H_0^1(\Omega))}\leq C,$$

where $C>0$, is independent of h and ε. Therefore, the properties of the difference quotient allow us to take the limit when $h\to 0_+$. So we obtain the estimates (2.8). The theorem is proved.

3. Convergence.

3.1. Convergence under hypothesis $\mathbf{H_{1.1}}$. We assume that hypothesis $\mathrm{H}_{1.1}$ holds. The estimates obtained in the Theorem 2.1 imply the existence of a subsequence of $\{u_\varepsilon\}_{\varepsilon>0}$, again written $\{u_\varepsilon\}$ such that:

$$u_\varepsilon \rightharpoonup u \quad \text{in } L^\infty(0,T;H_0^1(\Omega)) \text{ weak-star,} \tag{3.1}$$

$$u_\varepsilon' \rightharpoonup u' \quad \text{weakly in } L^2(Q). \tag{3.2}$$

Thanks to the compactness of the injection from $H^1(Q)$ to $L^2(Q)$, we have:

$$u_\varepsilon \to u \quad \text{in } L^2(Q) \text{ and a.e. on } Q. \tag{3.3}$$

Then, it results from the continuity of G and from (1.1) that

$$Gu_\varepsilon \rightharpoonup Gu \quad \text{weakly in } L^2(Q). \tag{3.4}$$

Finally, as $\varepsilon u_\varepsilon''$ is bounded in $L^2(0,T;H^{-1}(\Omega))$ thanks to (1.3) and as $\varepsilon u_\varepsilon''$ tends to 0 in the sense of the distributions, we have:

$$\varepsilon u_\varepsilon'' \rightharpoonup 0 \quad \text{weakly in } L^2(0,T;H^{-1}(\Omega)). \tag{3.5}$$

Therefore, we may take the limit in the equality (1.3) as $\varepsilon\to 0_+$. We find that the limit function u verifies:

$$\begin{aligned}&\alpha(u,v)+(L_1u,v)+(Gu,v)=(f,v)\quad \forall v\in H_0^1(\Omega),\text{ a.e. } s\in]0,T[,\\ &u\in L^\infty(0,T;H_0^1(\Omega));\qquad u'\in L^2(Q).\end{aligned}$$

Moreover, it results from (3.1) and (3.2) that u_ε and u are continuous from $[0,T]$ to $L^2(\Omega)$ and then:

$$u_\varepsilon(t)\rightharpoonup u(t) \quad \text{weakly in } L^2(\Omega) \quad \forall t\in[0,T].$$

Therefore the limit function u satisfies the initial condition:

$$u(0,x)=u_0(x) \quad \text{a.e. } x\in\Omega.$$

Then, the limit function u is solution of the parabolic problem

$$\mathscr{T}\begin{cases}\alpha(u,v)+(L_1u,v)+(Gu,v)=(f,v)\quad \forall v\in H_0^1(\Omega),\text{ a.e. } s\in]0,T[,\\ u\in L^\infty(0,T;H_0^1(\Omega)),\qquad u'\in L^2(Q),\\ u(0,x)=u_0(x)\quad \text{a.e. } x\in\Omega.\end{cases} \tag{3.6}$$

Now, such a problem has a unique solution, under hypothesis $\mathrm{H}_{1.1}$. In consequence, we have proved

THEOREM 3.1. *When $\varepsilon \to 0_+$, the solution u_ε of the problem $\mathscr{T}_\varepsilon$ converges to the solution u of the problem $\mathscr{T}$ in the following spaces*:

(i) $u_\varepsilon \rightharpoonup u$ *in* $L^\infty(0,T;H_0^1(\Omega))$ *weak-star and weakly in* $H^1(Q)$,
$Gu_\varepsilon \rightharpoonup Gu$ *weakly in* $L^2(Q)$;

(ii) $u_\varepsilon \to u$ *in* $L^2(Q)$.

Now, we are going to establish an estimate for $u_\varepsilon - u$ in $L^\infty(0,T;L^2(\Omega))$ and also in $L^2(0,T;H_0^1(\Omega))$.

THEOREM 3.2. *The solution u_ε of the problem $\mathscr{T}_\varepsilon$ and the solution u of the problem $\mathscr{T}$ satisfy*:

$$|u_\varepsilon - u|_{L^\infty(0,T;L^2(\Omega))} + \|u_\varepsilon - u\|_{L^2(0,T;H_0^1(\Omega))} \leq C\sqrt{\varepsilon},$$

$$|u_\varepsilon - u|_{L^p(Q)} \leq C(\varepsilon)^{1/p}$$

where $C > 0$, is independent of ε.

Proof. For each $v \in H_0^1(\Omega)$ and for a.e. $s \in]0,T[$, we have, by substracting (1.3) and (3.6):

$$\varepsilon(u_\varepsilon'', v) + \alpha(w_\varepsilon, v) + (aw_\varepsilon', v) + \left(\sum_{k=1}^n b_k \frac{\partial w_\varepsilon}{\partial x_k}, v\right) + (Gu_\varepsilon - Gu, v) = 0 \tag{3.7}$$

where $w_\varepsilon = u_\varepsilon - u$. We can set $v = w_\varepsilon(s)$ in (3.7). We obtain:

$$\varepsilon(u_\varepsilon'', w_\varepsilon) + \alpha(w_\varepsilon, w_\varepsilon) + (aw_\varepsilon', w_\varepsilon) + \left(\sum_{k=1}^n b_k \frac{\partial w_\varepsilon}{\partial x_k}, w_\varepsilon\right) + (Gu_\varepsilon - Gu, w_\varepsilon) = 0.$$

We integrate from 0 to t and we use the condition (1.1). Then we have:

$$2\int_0^t \|w_\varepsilon(s)\|_2^2 ds + \delta|w_\varepsilon(t)|_2^2 + 2\beta\int_0^t |w_\varepsilon(s)|_p^p ds + 2(\theta-1)\int_0^t |w_\varepsilon(s)|_2^2 ds$$

$$\leq -2\varepsilon(u_\varepsilon'(t), w_\varepsilon(t)) + 2\varepsilon\int_0^t (u_\varepsilon', w_\varepsilon') ds + \left|a' + \sum_{k=1}^n \frac{\partial b_k}{\partial x_k}\right|_{L^\infty(Q)} \int_0^t |w_\varepsilon(s)|_2^2 ds.$$

Thanks to the estimates (2.1) and Hölder's inequality, we obtain:

$$2\int_0^t \|w_\varepsilon(s)\|_2^2 ds + \frac{\delta}{2}|w_\varepsilon(t)|_2^2 + 2\beta\int_0^t |w_\varepsilon(s)|_p^p ds \leq k_1\varepsilon + k_2\int_0^t |w_\varepsilon(s)|_2^2 ds$$

where k_1 and k_2 are two positive constants independent of ε. Then, the theorem results from Gronwall's lemma.

3.2. Convergence under hypothesis $\mathbf{H_{1,2}}$. The additional estimates obtained in Theorem 2.2 allow us to complement Theorem 3.2 by the results of

THEOREM 3.3. *For each $\varepsilon > 0$, $u_\varepsilon - u$ where u_ε (resp. u) is the solution of the problem $\mathscr{T}_\varepsilon$ (resp. $\mathscr{T}$) satisfies the additional estimates*:

$$\|u_\varepsilon - u\|_{L^\infty(0,T;H_0^1(\Omega))} + |u_\varepsilon' - u'|_{L^2(Q)} \leq C\sqrt{\varepsilon}$$

where $C > 0$ is independent of ε.

Proof. It results from (3.7) that $w_\varepsilon = u_\varepsilon - u$ satisfies:

$$-\Delta w_\varepsilon + aw_\varepsilon' + \sum_{k=1}^n b_k \frac{\partial w_\varepsilon}{\partial x_k} + Gu_\varepsilon - Gu = -\varepsilon u_\varepsilon'' \quad \text{in } L^2(\Omega), \text{ a.e. } s \in]0,T[. \tag{3.8}$$

We can take the inner product in $L^2(\Omega)$ of the two members of (3.8) with $2w_\varepsilon'(s)$. We do so and we integrate from 0 to t. It follows that:

$$(3.9)\qquad \int_0^t -\left(2\Delta w_\varepsilon, w_\varepsilon'\right)ds + 2\int_0^t \left|\sqrt{a}\,w_\varepsilon'\right|_2^2 ds$$

$$= -2\int_0^t \left[\varepsilon\left(u_\varepsilon'', w_\varepsilon'\right) + \left(\sum_{k=1}^n b_k \frac{\partial w_\varepsilon}{\partial x_k}, w_\varepsilon'\right) + \left(Gu_\varepsilon - Gu, w_\varepsilon'\right)\right]ds.$$

As $w_\varepsilon \in L^\infty(0,T;H_0^1(\Omega)) \cap L^2(0,T;H^2(\Omega))$ and $w_\varepsilon' \in L^2(Q)$, we may write:

$$(3.10)\qquad \int_0^t -\left(2\Delta w_\varepsilon, w_\varepsilon'\right)ds = a\left(w_\varepsilon(t), w_\varepsilon(t)\right).$$

(This result is established for instance in Madaune [11] by a technique of regularization.) Next, in order to bound the three terms of the second member of (3.9), we first use Theorem 2.2, then the properties of the coefficients b_k, and finally for the nonlinear term, the condition (1.1), Hölder's inequality (with $(p-2)/(2(p-1)) + 1/(2(p-1)) + \frac{1}{2} = 1$) and (1.2). We obtain:

$$2\left|\int_0^t \left[\left(\varepsilon u_\varepsilon'', w_\varepsilon'\right) + \left(\sum_{k=1}^n b_k \frac{\partial w_\varepsilon}{\partial x_k}, w_\varepsilon'\right) + \left(Gu_\varepsilon - Gu, w_\varepsilon'\right)\right]ds\right|$$

$$\leq k_1\varepsilon + k_2\int_0^t \|w_\varepsilon\|_2^2 ds + \delta\int_0^t |w_\varepsilon'|_2^2 ds$$

where k_1 and k_2 are two positive constants, independent of ε. Then, we deduce from (3.9) and (3.10), the inequality:

$$\|w_\varepsilon(t)\|_2^2 + \delta\int_0^t |w_\varepsilon'(s)|_2^2 ds \leq c_1\varepsilon + c_2\int_0^t \|w_\varepsilon(s)\|_2^2 ds$$

with c_1 and c_2, two constants independent of ε. Gronwall's lemma gives us the estimate of $(u_\varepsilon - u)$ in $L^\infty(0,T;H_0^1(\Omega))$ and the estimate of $(u_\varepsilon' - u')$ in $L^2(Q)$; the proof is thus achieved.

4. Some remarks. In this section, we collect some remarks. In particular, we introduce the notion of boundary layer corrector which will allow us to improve the estimates on $u_\varepsilon - u$. To do so, we need some regularity properties of u that we can obtain under some additional assumptions on f and u_0. Finally, we end this section by giving some results when the operator L_1 is nonlinear.

4.1. Regularity of *u*. We have the first property:

PROPERTY 4.1. *Under hypothesis* $\mathrm{H}_{1.2}$, *the solution* u *of the problem* $\mathcal{T}$ *is such that*: $u' \in L^2(0,T;H_0^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$; $u'' \in L^2(0,T;H^{-1}(\Omega))$.

Proof. To establish the property for u', one may consider a sequence $\{u_m\}$ of approximated solutions of $\mathcal{T}$, constructed by Galerkin's method. It is easy to show, by classical techniques, that the sequence $\{u_m'\}$ is bounded independently of m in $L^2(0,T;H_0^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$. So by taking the limit on m one obtains the property for u'. Next to have: $u'' \in L^2(0,T;H^{-1}(\Omega))$, it is enough to differentiate the equality (3.6) with respect to t.

Now we assume the hypothesis

$\mathrm{H}_{1.3}$: $\mathrm{H}_{1.2}$ with $u_0 \in H^3(\Omega)$, $f(0) \in H^1(\Omega)$.

PROPERTY 4.2. *Under hypothesis* $\mathrm{H}_{1.3}$, *the solution* u *of the problem* $\mathcal{T}$ *is such that*:

$$u' \in L^\infty\left(0,T;H_0^1(\Omega)\right), \qquad u'' \in L^2(Q).$$

Proof. Once again, we consider the sequence $\{u_m\}$ of approximated solutions of $\mathscr{T}$. We show that u'_m (resp. u''_m) is bounded independently of m in $L^\infty(0,T;H_0^1(\Omega))$ (resp. in $L^2(Q)$). (The technique we use to do this means formally that we differentiate the equation (3.6) and we choose $v=u''$.) The additional regularity imposed on u_0 and $f(0)$ allows us to bound $u'_m(0)$ in $H^1(\Omega)$, independently of m. Next, we take the limit on m.

Consequence. Under the assumptions of Property 4.2, $u'\in C([0,T];L^2(\Omega))$, so $u'(0,\cdot)$ exists and

$$u'(0,\cdot)=f(0)+\Delta u_0-\sum_{k=1}^{n} b_k(0,\cdot)\frac{\partial u_0}{\partial x_k}-Gu_0. \tag{4.1}$$

4.2. Utilization of boundary layer correctors. We assume the hypothesis $\mathrm{H}_{1.2}$.

4.2.1. Definition of the boundary layer corrector θ_ε. Let be $g_\varepsilon\in L^2(Q)$ such that $|g_\varepsilon|_{L^2(Q)}$ is bounded independently of ε. Then, we consider the problem

$$\Pi_\varepsilon\begin{cases}\varepsilon(\theta''_\varepsilon,v)+\alpha(\theta_\varepsilon,v)+(L_1\theta_\varepsilon,v)+(G(\theta_\varepsilon+u)-Gu,v)=\varepsilon(g_\varepsilon-u'',v)\\ \qquad\qquad\qquad\qquad\forall v\in H_0^1(\Omega),\ \text{a.e. } s\in\,]0,T[,\\ \theta_\varepsilon\in L^\infty(0,T;H_0^1(\Omega)),\qquad \theta'_\varepsilon\in L^\infty(0,T;L^2(\Omega)),\\ \theta_\varepsilon(0,x)=0,\quad \theta'_\varepsilon(0,x)=u_1(x)-u'(0,x)\quad\text{in } H^{-1}(\Omega)\end{cases} \tag{4.2}$$

(we know, from the Property 4.1, that $u''\in L^2(0,T;H^{-1}(\Omega))$). We may assert that the problem Π_ε has a unique solution since $\theta_\varepsilon+u$ satisfies the problem:

$$\begin{cases}\varepsilon((\theta_\varepsilon+u)'',v)+\alpha(\theta_\varepsilon+u,v)+(L_1(\theta_\varepsilon+u),v)+(G(\theta_\varepsilon+u),v)\\ \qquad\qquad=(f+\epsilon g_\varepsilon,v)\quad\forall v\in H_0^1(\Omega),\ \text{a.e. } s\in\,]0,T[,\\ \theta_\varepsilon+u\in L^\infty(0,T;H_0^1(\Omega)),\qquad (\theta_\varepsilon+u)'\in L^\infty(0,T;L^2(\Omega)),\\ (\theta_\varepsilon+u)(0,x)=u_0(x),\qquad (\theta_\varepsilon+u)'(0,x)=u_1(x)\quad\text{a.e. } x\in\Omega\end{cases}$$

which has a solution and only one, thanks to the Theorem 1.1.

PROPERTY 4.3. *Under hypothesis* $\mathrm{H}_{1.2}$, *we have*:

$$\|u_\varepsilon-(\theta_\varepsilon+u)\|_{L^\infty(0,T;H_0^1(\Omega))}+|u'_\varepsilon-(\theta_\varepsilon+u)'|_{L^2(Q)}\le C\varepsilon,$$
$$|u'_\varepsilon-(\theta_\varepsilon+u)'|_{L^\infty(0,T;L^2(\Omega))}\le C\sqrt{\varepsilon},$$

where C is a positive constant independent of ε.

Proof. Let be $z_\varepsilon=u_\varepsilon-(\theta_\varepsilon+u)$; z_ε satisfies

$$\begin{aligned}&\varepsilon z''_\varepsilon-\Delta z_\varepsilon+L_1z_\varepsilon+Gu_\varepsilon-G(\theta_\varepsilon+u)=-\varepsilon g_\varepsilon\quad\text{in } L^2(\Omega)\quad\text{a.e. } s\in\,]0,T[,\\ &z_\varepsilon(0,x)=z'_\varepsilon(0,x)=0.\end{aligned} \tag{4.3}$$

By using a method of regularization as in the proof of the Theorem 2.1 we may suppose that $g'_\varepsilon\in L^2(Q)$. Then, thanks to the Theorem 1.2, we can take the inner product in $L^2(\Omega)$ of the two members of (4.3) with $2z'_\varepsilon(s)$. We do so and we integrate from 0 to t.

By calculations analogous to the ones done in the proof of the Theorem 3.3, we obtain the inequality:

$$\varepsilon|z_\varepsilon'|_2^2+\|z_\varepsilon\|_2^2+\delta\int_0^t|z_\varepsilon'(s)|_2^2ds\le c_1\varepsilon^2+c_2\int_0^t\|z_\varepsilon(s)\|_2^2ds$$

where c_1 and c_2 are two positives constant independent of ε and g_ε. The utilization of Gronwall's lemma gives us the estimates of the Property 4.3.

4.2.2. Example of a boundary layer corrector. We introduce

$$\theta_\varepsilon=\varepsilon\left[u'(0,x)-u_1(x)\right]e^{-a(x,0)(t/\varepsilon)} \tag{4.4}$$

and we shall prove that θ_ε is a boundary layer corrector. Thus, the element θ_ε which is a boundary layer corrector for the case of the linear equation $L_2^{(\varepsilon)}u_\varepsilon+L_1u_\varepsilon=f$ remains a corrector for the case of nonlinearities of the type introduced here. We need some additional properties on f and u_0.

$H_{1.4}$: $H_{1.2}$ with $u_0\in H^4(\Omega)$, $u_1\in H^2(\Omega)$, $f(0)\in H^2(\Omega)$, $Gu_0\in H^2(\Omega)$.

THEOREM 4.4. *Under hypothesis* $H_{1.4}$, θ_ε *defined by* (4.4) *is a boundary layer corrector that is* θ_ε *satisfies the Property* 4.3.

Proof. We first remark that the assumption $H_{1.4}$ implies, thanks to (4.1), that: $u'(0,\cdot)-u_1\in H^2(\Omega)$. We assume, in order to simplify the calculations, that $a\equiv 1$. The result holds in the general case; only the calculations are more complicated. It is easy to prove that θ_ε satisfies the equality (4.2) with

$$g_\varepsilon=u''-\left(\Delta u'(0,\cdot)-\Delta u_1\right)e^{-t/\varepsilon}$$
$$+\sum_{k=1}^n b_k\left(\frac{\partial}{\partial x_k}u'(0,\cdot)-\frac{\partial}{\partial x_k}u_1\right)e^{-t/\varepsilon}+\frac{1}{\varepsilon}\left(G(\theta_\varepsilon+u)-Gu\right)$$

and that $|g_\varepsilon|_{L^2(Q)}$ is bounded independently of ε. So, we can apply results of Property 4.3 to θ_ε defined by (4.4).

4.2.3. Consequence. The corrector θ_ε defined by (4.4) is such that:

$$\|\theta_\varepsilon\|_{L^\infty(0,T;H_0^1(\Omega))}\le C\varepsilon,\qquad\left|\sqrt{t}\,\theta_\varepsilon'\right|_{L^2(Q)}\le C\varepsilon,$$

where C is a positive constant independent of ε. Therefore we obtain

THEOREM 4.5. *Under hypothesis* $H_{1.4}$, *the solution* u_ε *of the problem* $\mathscr{T}_\varepsilon$ *and the solution* u *of the problem* $\mathscr{T}$ *are such that*:

(i) $\|u_\varepsilon-u\|_{L^\infty(0,T;H_0^1(\Omega))}\le C.\varepsilon$; $|\sqrt{t}(u_\varepsilon'-u')|_{L^2(Q)}\le C\varepsilon$,
where C is a positive constant independent of ε,

(ii) *for each* $\tau\in]0,T[$, $|u_\varepsilon'-u'|_{L^\infty(\tau,T;L^2(\Omega))}\le C_1\sqrt{\varepsilon}$,
where C_1 *is a positive constant independent of* ε, *but dependent on* τ.

4.3. A remark when L_1 is nonlinear and $\Omega\subset\mathbf{R}$. When Ω is an interval of $\mathbf{R}$, the results of the Theorems 1.1 and 1.2 remain valid when the operator L_1 is nonlinear:

$$L_1u=a(x,t,u)\frac{\partial u}{\partial t}+b(x,t,u)\frac{\partial u}{\partial x}$$

upon condition of assuming instead of $H_{1.1}$(ii), the condition

$$H_{1.1}'(\mathrm{ii}):\ a,b\in C^1(\overline{Q}\times\mathbf{R}),\ a(x,t,u)\ge\delta>0\text{ on }\overline{Q}\times\mathbf{R},\ b\in L^\infty(Q\times\mathbf{R}).$$

To prove this, it is sufficient to work again with Galerkin's method used when L_1 is a linear operator. Next, it is easy to verify that the solution u_ε of the problem $\mathscr{T}_\varepsilon$ satisfies the estimates of the Theorem 2.1 and the convergence results of the Theorem 3.1. Now, if we assume that the coefficient a is independent of u, we may prove that u_ε satisfies the estimates of the Theorem 2.2 and the convergence properties of the Theorems 3.2 and 3.3. Of course, modifications must be brought on the treatment of the term $b(x,t,u)\partial u/\partial x$; we use in particular the algebraic and topological inclusion: $H^1(\Omega) \hookrightarrow L^\infty(\Omega)$.

REFERENCES

[1] A. BENAOUDA, *Thèse de 3ème cycle*, Pau, 1983.

[2] G. C. HSIAO AND R. J. WEINACHT, *A singularly perturbed Cauchy problem*, J. Math. Anal. Appl., 71 (1979), pp. 242–250.

[3] ______, *Singular perturbations for a semilinear hyperbolic equation*, this Journal, 14 (1983), pp. 1168–1179.

[4] J. KISYNSKI, *Sur les èquations hyperboliques avec petit paramètre, problème de Cauchy abstrait*, Colloq. Math., 10 (1963), pp. 331–343.

[5] N. A. LAR'KIN, *Hyperbolic regularization of the Burgers' Equation*, Differential Equations, 16 (1980), pp. 77–79.

[6] M. P. LENJUK AND V. V. FEDORUK, *The generalised heat equation*, Dopovidi Akad. Nauk. Ukrain. RSR Ser. A (1972), pp. 1075–1078, 1149. (In Russian)

[7] J. L. LIONS, *Quelques méthodes de rèsolution de problèmes aux limites non linèaires*, Dunod, Paris, 1969.

[8] ______, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Mathematics 323, Springer-Verlag, Berlin, New York, 1973.

[9] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.

[10] M. MADAUNE, *Perturbations singulières pour une classe d'èquations hyperboliques quasi-linèaires*, Ann. Fac. Sci. Toulouse, 1 (2) (1979), pp. 137–163.

[11] ______, *Un thèorème d'unicitè pour des inèquations variationnelles paraboliques dègènèrèes*, Comm. Partial Differential Equations, 7 (1982), pp. 433–468.

[12] M. MADAUNE AND J. GENET, *Singular perturbations for a class of nonlinear hyperbolic hyperbolic problems*, J. Math. Anal. Appl., 64 (1978), pp. 1–24.

[13] M. MARCUS AND V. J. MIZEL, *Absolute continuity on tracks and mappings of Sobolev spaces*, Arch. Rat. Mech. Anal., 45 (4) (1972), pp. 294–320.

[14] R. I. MURADOV, *A mixed problem for nonlinear hyperbolic equation with a small parameter*, Izv. Akad. Nauk Azerkaidžan SSR, Ser. Fiz-Tehn. Mat. Nauk, 2 (1974), pp. 87–92.

[15] ______, *The mixed problem for a singular hyperbolic equation with a small parameter*, Azerbaidžan Gos. Univ. Učen-Zap, 6 (1978), pp. 3–13.

[16] R. I. MURADOV AND M. G. GASANOV, *The Cauchy problem for a hyperbolic equation with a small parameter*, Azerbaidžan Gos. Univ. Ućem-Zap., 2–3, (1974), pp. 43–50.

[17] B. NAZARKULOVA AND P. S. PANKOV, *An application of asymptotic methods in the theory of telegrapher's equations for large velocities–Studies in integro-differential eq.*, in Kirghizia n°8, Izdat. "Ilim", Frunze, (1971), pp. 285–292.

[18] J. NEČAS, *Les mèthodes directes en thèorie des èquations elliptiques*, Masson, Paris, 1967.

[19] J. C. SAUT, *Thèse de 3ème cycle*, Paris-Orsay, 1972.

[20] M. ZLAMAL, *Problème mixte pour une èquation hyperbolique avec petit paramètre*, Czechoslovak Math. J., 10 (1960), pp. 83–122.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

011

THE ONE-DIMENSIONAL NONLINEAR HEAT EQUATION WITH ABSORPTION: REGULARITY OF SOLUTIONS AND INTERFACES*

MIGUEL A. HERRERO[†] AND JUAN L. VÁZQUEZ[‡]

Abstract. We consider the equation $u_t=(u^m)_{xx}-\lambda u^n$ with $m>1$, $\lambda>0$, $n\geqq m$ as a model for heat diffusion with absorption. Hence we assume that $u\geqq 0$ for $x\in\mathbb{R}$, $t\geqq 0$. We study the regularity of the solution to the Cauchy problem for this degenerate parabolic equation. When the initial datum $u_0(x)$ is positive only in a part of the space $\mathbb{R}$, we also study the regularity of the free boundaries that appear. The asymptotic behavior of solutions and free boundaries is also discussed.

Key words. nonlinear diffusion with absorption, regularity, interfaces or free boundaries, waiting time, asymptotic behavior

AMS(MOS) subject classifications. Primary 35Q20, 35K55, 35K65

Introduction. In this paper we study the regularity and propagation properties of the solutions to the following Cauchy problem:

$$u_t=(u^m)_{xx}-\lambda u^n \quad \text{when } (x,t)\in S=\mathbb{R}\times(0,\infty), \tag{0.1}$$

$$u(x,0)=u_0(x) \quad \text{for } x\in\mathbb{R}, \tag{0.2}$$

where λ, m and n are positive constants and u_0 is a nonnegative, continuous and bounded real function.

Equations like (0.1) are used as mathematical models in a number of problems, in particular in describing thermal propagation with absorption; then u stands for the temperature, $(u^m)_{xx}$ is the diffusion term and $-\lambda u^n$ represents the absorption of heat by the medium. We see that in general the thermal conductivity mu^{m-1} and the absorption coefficient λu^{n-1} are temperature-dependent. We refer to Zeldovich and Raizer [ZR] for a detailed account of nonlinear heat conduction problems.

In this paper we consider the case of "slow diffusion" $m>1$ and "weak absorption" $n\geqq m$. The first restriction implies that initial data with compact support propagate with finite velocity, so that two *interfaces* or *free boundaries* arise, $x=\zeta_1(t)$ and $x=\zeta_2(t)$, that bound the support of $u(\cdot,t)$ for every $t>0$; see Oleinik, Kalashnikov and Chzou [OKC] for the case $\lambda=0$ and Kalashnikov [K2] for positive λ. We study here the regularity of the solution u and the free boundaries $\zeta_i(t)$. We obtain the following results.

I) *Regularity.*

i) For a.e. $(x,t)\in S$ we have:

$$|(u^{m-1})_x(x,t)|^2\leqq c\|u_0\|_\infty^{m-1}t^{-1} \tag{0.3}$$

* Received by the editors August 6, 1985; accepted for publication August 8, 1985.

[†] Departamento de Ecuaciones Funcionales, Facultad de Matemáticas, Universidad Complutense, 28040-Madrid, Spain. This author was partially supported by SFPI, Ministerio de Educación y Ciencia, Spain, to visit the Mathematical Institute, University of Oxford.

[‡] Divisiôn de Matemâticas, Universidad Autonoma de Madrid, 28049-Madrid, Spain. This author was partially supported by a Fulbright award.

for a positive constant $c=c(m)$ ($\|f\|_p$, $1\leqq p\leqq +\infty$, denotes the L^p-norm of a function f). This is proved in Theorem 1 under the more general assumptions $m>1$, $n>0$, $m+n\geqq 2$.

The fact that $(u^{m-1})_x$ is bounded for $t\geqq\tau>0$ was already proved by Kalashnikov [K3]. From this it follows that u is Hölder continuous in S with respect to x with exponent $\alpha=\min(1,1/m-1)$, and consequently (cf. [G]) with respect to t with exponent $\alpha/2$. It is also proved in [K3] that the result cannot be true in general for $m+n<2$.

ii) u^{m-1} is a semiconvex function of x for every $t>0$ and we have

$$(u^{m-1})_{xx}\geqq -k/t \tag{0.4}$$

with $k=k(n,m)>0$ (Theorem 2). This result is due to Aronson and Bénilan [AB] in the nonabsorption case $\lambda=0$, and has played a prominent role in the theory of that "porous medium" equation. From (0.4) a sharper form of (0.3) follows, see (2.14).

II) *Interfaces*. We use the above results to discuss the regularity of the solution u and the free boundaries. We prove that the $\zeta_i(t)$ are C^1, strictly increasing functions after a possibly positive time t_i^* (Theorem 5) whose occurrence we characterize in terms of u_0 (Theorem 4). The equation satisfied on the free boundary is proved to be the same as in the porous medium case:

$$-\zeta_i'(t)=\lim v_x(x,t) \quad \text{as } x\to\zeta_i(t), \quad x\in(\zeta_1(t),\zeta_2(t)), \tag{0.5}$$

where $v=(m/m-1)u^{m-1}$ (Theorem 3). Note that by (0.3) $|v_x|$ is locally bounded for $t>0$. It now follows from (0.5) that whenever $\zeta_i'(t)\neq 0$ (i.e., for $t>t_i^*$) the function $v_x(x,t)$ cannot be continuous at $(\zeta(t),t)$: we thus obtain the *optimal regularity* for solutions with an interface. This is exactly the situation in the porous medium case (cf. Aronson [A1], Knerr [Kn1]). Of course, if $u>0$ in S then s is smooth everywhere.

It is to be noted that in our study of the free boundary we do not require, as it was customary in the literature ([K2], [Kn1], [Ke1], $\cdots$) that u_0 have compact support, but only that

$$d=\sup\{x\in\mathbb{R}\colon\ u_0(x)>0\}<+\infty. \tag{0.6}$$

This is the natural condition on u_0 in order to have a (right-hand) interface $x=\zeta(t)$, defined by $\zeta(t)=\sup\{x\in\mathbb{R}\colon\ u(x,t)>0\}$, $t\geqq 0$.

III) *Asymptotic behavior*. We use our previous results to describe the behavior as $t\to\infty$ of 1) the solution u of (0.1), (0.2) for general u_0, 2) its interface $\zeta(t)$ if u_0 satisfies (0.6).

Under the hypothesis that u_0 has compact support $I=[a,b]$ (and is positive in (a,b)) Kersner [Ke1] studies the asymptotic behavior for $m>1$, $n\geqq 1$ and distinguishes three regions: $1<n<m$, $m<n<m+2$ and $n>m+2$, with three limit cases: $n=1$, $n=m$ and $n=m+2$. He gives rates for u and ζ as $t\to\infty$ in the different cases. A parallel study is done by Knerr [Kn2]. The set of estimates if then improved by Herrero [H] and completed by Bertsch, Kersner and Peletier [BKP1].

We give simple proofs of some of the difficult estimates above (under our more general conditions) by applying the results in i) and ii). Since we use (0.5) to convert estimates on v_x into estimates on $\zeta'(t)$ this quantity is also controlled. But our main result is to point out a basic difference between the cases $m\leqq n<m+2$ and $n>m+2$ that happens precisely because we consider general initial data: in the first region $m\leqq n<m+2$ the asymptotic behavior of u (and that of ζ if u_0 satisfies (0.6)) is essentially *unique* and its rates are given by the *absorption* (Theorem 8). On the

contrary, if $n>m+2$ solutions corresponding to $u_0\in L^1(\mathbb{R})$ have minimal rates given by the *diffusion* term (Theorem 6; they agree with the ones for the porous medium equation, see Vàzquez [V1]), whereas when $u_0(x)\geqq c>0$ for all $x\ll 0$ and some constant c, the rates are maximal and given by the *absorption* as in the previous case (Theorem 7).

Finally the paper is divided into several sections, according to the following plan:

1. Preliminaries.
2. Regularity of solutions.
3. Proof of Theorem 1.
4. Proof of Theorem 2.
5. Regularity of interfaces.
6. Asymptotic behavior.

1. Preliminaries. It is known that for $\lambda=0$ the problem (0.1), (0.2) does not have classical solutions for initial data u_0 that vanish in some interval, even if they are smooth; this happens because of the degeneracy of the equation at $u=0$. A similar situation occurs for $\lambda>0$ in view of particular explicit solutions and in general because of the regularity results that we shall prove. Therefore a concept of *generalized* solution is needed. A function $u(x,t)$ continuous, nonnegative and bounded in $\bar{S}=\mathbb{R}\times[0,\infty)$ is said to be a generalized solution of (0.1) (0.2) if $u(x,0)=u_0(x)$ and the following equality holds for every function $\phi(x,t)\in C^{2,1}(\bar{S})$ having compact support in $\bar{S}$:

$$(1.1)\qquad I(u,\phi)\equiv\iint_S(u^m\phi_{xx}+u\phi_t-\lambda u^n\phi)\,dx\,dt+\int_{\mathbb{R}}u_0(x)\phi(x,0)\,dx=0.$$

When a function $u(x,t)$ as above satisfies $I(u,\phi)\leqq$ (resp. $I(u,\phi)\geqq 0$) for all nonnegative test functions ϕ as before we say that u is a *supersolution* of (0.1), (0.2) (resp. a *subsolution*).

The existence, uniqueness and properties of generalized solutions have been studied by Kalashnikov [K2], Kersner [Ke2] and Knerr [Kn2]. (Incidentally, they use slightly different but equivalent definitions.) In the sequel we shall use a series of results that we summarize here. We refer to the pertinent literature for some of them and prove for the reader's convenience those for which we could find no proof.

THEOREM 0. *Let $m>1$, $n>0$, $m+n\geqq 2$ and let u_0 be a continuous, nonnegative and bounded real function. Then*:

i) *There exists a unique generalized solution $u(x,t)$ of* (0.1) (0.2).

ii) *u is smooth in the open set $\{(x,t)$: $u(x,t)>0\}$. In particular, if $u_0>0$ everywhere and $n\geqq 1$ then u is smooth in $\bar{S}$.*

iii) *If $\hat{u}$ is a supersolution and $\bar{u}$ a subsolution to* (0.1) *with initial data $\hat{u}_0$ and $\bar{u}_0$ respectively, and $\hat{u}_0\geqq\bar{u}_0$ then $\hat{u}\geqq\bar{u}$ in S.*

iv) *Let $u_j(x,t)$ $(j=1,2\cdots)$ and $u(x,t)$ be solutions to* (0.1) *with initial data $u_{0_j}(x)$, $u_0(x)$ respectively. If $u_{0_j}\to u_0$ uniformly on compacts as $j\to\infty$ then $u_j(x,t)\to u(x,t)$ uniformly on compacts of S as $j\to\infty$.*

Proof. Existence and uniqueness is obtained by Kalashnikov [K2] for the more general equation

$$(1.2)\qquad u_t=(\alpha(u))_{xx}-\psi(u)$$

under assumptions that in our case mean $n\geqq 1$ and u_0^m Lipschitz continuous. The range $0<n<\infty$ is dealt with by Kersner [Ke2], as a particular case of equation (1.2), by using a sequence of approximate problems,

$$(1.3)\qquad \begin{aligned} &u_t=(u^m)_{xx}-\lambda u^n+\lambda\varepsilon^n,\\ &u(x,0)=u_{0_\varepsilon}(x), \end{aligned}\qquad \varepsilon>0$$

where $u_{0_\varepsilon}(x)>u_0(x)$ and $u_{0_\varepsilon}(x)\downarrow u_0(x)$ as $\varepsilon\downarrow 0$. This allows him to avoid the degeneracy of (0.1). He assumes compact support of u_0. His proof of uniqueness (i), smoothness (ii) and comparison (iii) can be repeated in our case (for this last point see also Bertsch [Be]). But let us prove in detail the existence (i) and dependence (iv) parts.

As to the existence, we consider the approximate problems (1.3) with a sequence of smooth, positive functions $u_{0_\varepsilon}(x)\geqq\varepsilon>0$ that converge to u_0 uniformly on compacts and is bounded by a constant $M>0$ independent of ε. It is easy to see that (1.3) has, for each ε, a unique classical solution in S $u_\varepsilon(x,t)$ and $\varepsilon\leqq u_\varepsilon\leqq M$ in S (cf. [Ke2]). Then, using Kalashnikov's result (cf. [K3]), the uniform bound

$$(1.4)\qquad \left|\left(u_\varepsilon^{m-1}\right)_x\right|\leqq C(\tau)\quad \text{for } t\geqq\tau$$

follows and therefore the functions u_ε are Hölder continuous with respect to x uniformly in $\varepsilon>0$, $x\in\mathbb{R}$ and $t\geqq\tau>0$. Hence $\{u_\varepsilon\}$ is also uniformly Hölder continuous in t (see Kruzhkov [Kr], Gilding [G]). Thus, passing to a subsequence (that we label again u_ε), we find that u_ε converges to a continuous function $u(x,t)$ uniformly on compacts on S, and therefore for every $(x,t)\in S$. Since we have

$$(1.5)\qquad I(u_\varepsilon,\phi)=-\lambda\varepsilon^n\iint\phi(x,t)\,dx\,dt$$

in the limit $\varepsilon\downarrow 0$ we get $I(u,\phi)=0$.

We have yet to check that u is continuous down to $t=0$ and that $u(x,0)=u_0(x)$. The classical results for quasilinear parabolic equations, cf. [LSU], do not apply directly because of the degeneracy of the equation.

Consider first a point x_0 where $u_0(x_0)>0$. We construct an explicit lower barrier for u_0 at x_0 as follows: given $0<\delta<u_0(x_0)$ there is an interval $I=[x_0-a,x_0+a]$ where $u_0(x)\geqq u_0(x_0)-\delta\equiv b>0$. We now consider the function

$$(1.6)\qquad \tilde v(x,t)=B-\frac{B}{a^2}(x-x_0)^2-ct,$$

with $B=(m/(m-1))b^{m-1}$ and $c>0$, in the set

$$(1.7)\qquad \Omega=\left\{(x,t)\in S\colon\ B>\frac{B}{a^2}\cdot(x-x_0)+ct\right\}.$$

If $c>2B^2(m-1)/a^2+\mu B^\beta$ for some μ, $\beta\geqq 0$, $\tilde v$ satisfies in Ω

$$(1.8)\qquad L_1\tilde v\equiv\tilde v_t-(m-1)\tilde v\tilde v_{xx}-\tilde v_x^2+\mu\tilde v^\beta<0.$$

We now perform the change of variables $\tilde v=(m/m-1)\tilde u^{m-1}$. Then $\tilde u\in C^\infty(\Omega)$ and satisfies there the inequality

$$(1.9)\qquad L\tilde u\equiv\tilde u_t-(\tilde u^m)_{xx}+\lambda\tilde u^n<0,$$

where $\mu=\lambda m((m-1)/m)^\beta$ and $\beta=(m+n-2)/(m-1)$ (hence we assume $\lambda\geqq 0$, $n+m\geqq 2$). It is easy to see that for ε small the approximations u_ε above to the solution $u(x,t)$ of (0.1) with initial data $u_0(x)$ satisfy: i) $u_\varepsilon\in C^\infty(\Omega)$, ii) $Lu_\varepsilon>0$ in Ω, iii) $u_\varepsilon(x,t)\geqq\tilde u(x,t)$ in $\partial\Omega$. By the maximum principle it follows that $u_\varepsilon(x,t)\geqq\tilde u(x,t)$ in Ω. Letting $\varepsilon\downarrow 0$ we have $u(x,t)\geqq\tilde u(x,t)$ in Ω, hence as $(x,t)\in S\to(x_0,0)$ we get

$$\liminf u(x,t)\geqq\tilde u(x_0,0)=u_0(x)-\delta.$$

Letting $\delta \to 0$ the lower semicontinuity of u at $(x_0, 0)$ follows. The result is obviously true if $u_0(x_0)=0$ since $u \geqq 0$.

The upper semicontinuity at $t=0$ is now easy. For any $\delta > 0$ we can take a smooth function $\hat{u}_1(x)$ such that $\hat{u}_1(x) > \delta$ and $u_0(x) \leqq \hat{u}_1(x) \leqq u_0(x)+\delta$. Let $\hat{u}(x,t)$ be the corresponding solution. By the argument above $\hat{u}(x,t)$ and its approximations $\hat{u}_\varepsilon(x,t)$, ε small, will be uniformly positive in every strip $S_T = \mathbb{R} \times (0, T]$, T small, and the classical theory implies that $\hat{u}$ is continuous at $t=0$. Therefore as $(x,t) \in S \to (x_0, 0)$ we have

$$\limsup u(x,t) \leqq \lim \hat{u}(x,t) = u(x_0) \leqq u_0(x_0) + \delta.$$

To end the proof of the theorem we remark that the dependence result, part iv), can be proved by repeating these same arguments.

2. Regularity results. To motivate the results in this section, we disregard the absorption term in (0.1), thus obtaining the porous medium equation:

$$u_t = (u^m)_{xx}, \qquad m > 1, \tag{2.1}$$

that appears in a variety of situations (see for instance Peletier [P]), in particular as as model for the flow of an isentropic gas through a porous medium. In this case u represents the density of the gas and two other physical magnitudes play a role in describing the flow: the *pressure*, defined by

$$v = \frac{m}{m-1} u^{m-1} \tag{2.2}$$

and the *local velocity*, given by $-v_x$.

Equation (2.1) has been extensively studied and many of its properties are well known by now. In particular, with respect to the regularity Aronson [A1] proved that for every solution of (2.1), (0.2) the velocity is *bounded* in every strip $S_{\tau,\infty} = \mathbb{R} \times (\tau, \infty)$ where $\tau > 0$, and this result is best possible in the sense that v_x can be discontinuous. This gives a regularity threshold for solutions, since $u(x,t)$ is Hölder continuous in x with exponent $\alpha = \min(1, 1/(m-1))$, which is sharp as it can be tested against the explicit similarity solutions obtained by Barenblatt [B]

$$\hat{u}_p(x,t) = t^{-1/(1+m)} \left[p - \frac{m-1}{2m(m+1)} \left(x t^{-1/(1+m)} \right)^2 \right]_+^{1/(m-1)} \tag{2.3}$$

(valid also for $m < 1$) where $p > 0$ is arbitrary. We just note here that this threshold no longer appears when $m < 1$, in which case disturbances from rest propagate with infinite speed (cf. Aronson and Bénilan [AB] and Vâzquez [V2]) and nonnegative solutions are always positive and smooth everywhere.

A second important regularity result is also shown in [AB]; it states that the pressure is a *semiconvex* function with respect to the space variable. More precisely,

$$v_{xx} \geqq -\frac{1}{(m+1)t}, \tag{2.4}$$

where the inequality is to be understood in the sense of distributions in S. Noting from this and the equation $v_t = (m-1) v v_{xx} + v_x^2$ satisfied by v, it easily follows that

$$u_t \geqq -\frac{u}{(m+1)t} \quad \text{in } S.$$

On the other hand Kalashnikov [K3] proved that if we apply the previous definitions of pressure and velocity to equation (0.1) (as we shall do in the sequel), the velocity of solutions to (0.1), (0.2) is bounded in every strip of the form $S_{\tau,\infty}$ with $\tau>0$ if $m+n\geqq 2$, but not in general if $m+n<2$. Though the boundedness of v_x is already an important fact, for some applications we need to know the way in which this bound depends on time. This is included in our first result.

THEOREM 1. *Let $u(x,t)$ be the solution to* (0.1), (0.2) *with n, m as above and $n+m\geqq 2$, $n>0$. Then $v(x,t)$ is Lipschitz continuous in x for each $t>0$ and we have a.e.*:

$$(2.5)\qquad |v_x(x,t)|^2\leqq\frac{2}{m}\|v_0\|_\infty t^{-1}.$$

Moreover, for every t_1, t_2 with $0\leqq t_1<t_2$ one has

$$(2.6)\qquad \sup_{x\in R}|v_x(x,t_2)|<\sup_{x\in R}|v_x(x,t_1)|.$$

The proof of Theorem 1 is postponed to §3. To obtain from the theorem sharper bounds for $|v_x|$ as $t\to\infty$, we remark that there exist two a priori bounds for solutions of the problem under consideration. The first one is valid if $u_0(x)\in L^p(R)$ with $1\leqq p<+\infty$ and asserts that there exists a constant $c=c(m,p)>0$ such that for every solution of (0.1), (0.2) we have in S

$$(2.7)\qquad \dot{u}(x,t)\leqq c\|u_0\|_p^\alpha t^{-\theta},$$

where

$$(2.8)\qquad \alpha=\frac{2p}{2p+m-1},\qquad \theta=\frac{1}{2p+m-1}.$$

This has been proved for the case $\lambda=0$ by Véron [Ve]. Since solutions of (2.1) are supersolutions to (0.1) with the same initial data, the estimate holds for $\lambda>0$. Therefore, using (2.7) in the time interval $[0,t/2]$ and (2.5) in $[t/2,t]$ after displacing the origin of time from 0 to $t/2$, we get the following.

COROLLARY 1. *Let u, u_0,m,n be as above and let also $u_0\in L^p(\mathbb{R})$, $1\leqq p<+\infty$. Then we have in S*

$$(2.9)\qquad |v_x(x,t)|\leqq c_1\|u_0\|_p^\gamma\cdot t^{-\delta},$$

where $c_1=c_1(m,p)>0$, $\gamma=p(m-1)/(2p+m-1)$, $\delta=(p+m-1)/(2p+m-1)$.

The second bound occurs for $n>1$ and is a consequence of the existence of a solution of (0.1) of the form:

$$u^*(x,t)=(\lambda(n-1)t)^{-1/(n-1)}.$$

It then follows by comparison that for every solution of (0.1), (0.2) we have in S

$$(2.10)\qquad u(x,t)\leqq(\lambda(n-1)t)^{-1/(n-1)}.$$

Combining (2.10) and (2.5) as above yields the following.

COROLLARY 2. *Let u, u_0,m as in Theorem 1 and let $n>1$. Then for every $(x,t)\in S$ we have*

$$(2.11)\qquad |v_x(x,t)|^2\leqq c_2\cdot t^{-(m+n-2)/(n-1)}$$

with

$$c_2^2=\frac{2^\sigma}{m-1}\left(\frac{1}{\lambda(n-1)}\right)^{(m-1)/(n-1)},\qquad \sigma=2+\frac{m-1}{n-1}.$$

Our second result consists in establishing a semiconvexity property like (2.4) for (0.1), (0.2) under the restrictions $m>1$, $n\geqq m$.

THEOREM 2. *Let u, v be as above and assume that $n\geqq m$. Then*

$$v_{xx}\geqq-\frac{k}{t}\quad in\ D'(S)\tag{2.12}$$

where

$$k=\frac{1}{m+1}+\left(\frac{2^\sigma(m+n-2)}{m(m^2-1)}\right)^{1/2}.$$

The proof of Theorem 2 fails for $1<n<m$, though it can be adapted to cover the case $n\leqq 1$. When $n=1$ one is easily reduced to the nonabsorption equation (2.1) by means of the change in variables (see Martinson and Pavlov [MP]):

$$\begin{gathered}\tau=\frac{1}{\lambda(m-1)}\left(1-e^{-\lambda(m-1)t}\right),\\ u(x,t)=\hat{u}(x,\tau)e^{-\lambda t}.\end{gathered}\tag{2.13}$$

In proving Theorem 2, essential use is made of estimate (2.5). Conversely, (2.5) is a simple consequence of Theorem 2 and the boundedness of v. In fact a stronger result holds.

COROLLARY 3. *Under the assumptions of Theorem 2, we have for every $(x,t)\in S$:*

$$|v_x(x,t)|\leqq\left(2k\|v(x,t)\|_\infty t^{-1}\right)^{1/2},\tag{2.14}$$

where k is the constant in (2.12).

This result is a consequence of the following Calculus Lemma applied to $v(\cdot,t)$: "Let f be a bounded real function such that $f''(x)\geqq -d>0$. Then f is Lipschitz continuous and

$$|f'(x)|\leqq\left(2\|f\|_\infty d\right)^{1/2}$$

for every $x\in R$."

3. The proof of Theorem 1. We begin by deriving the basic result, that consists in applying Bernstein's method to the pressure v assuming that our solution is smooth. Let us write $S_T=\mathbb{R}\times(0,T]$ with $T>0$. We obtain the following.

LEMMA 3.1. *Let $v(x,t)$ be a smooth, positive and bounded solution of the equation*

$$v_t=(m-1)vv_{xx}+v_x^2-cv^\beta\tag{3.1}$$

in $\bar{S}_T=\mathbb{R}\times[0,T]$ and let c,β be nonnegative constants. Then we have:

$$v_x^2(x,t)\leqq\frac{2}{m}\|v(x,0)\|_\infty t^{-1}\tag{3.2}$$

for every $(x,t)\in S_T$. In addition, for $0\leqq t_1<t_2$

$$\sup_{x\in\mathbb{R}}|v_x(x,t_2)|<\sup_{x\in\mathbb{R}}|v_x(x,t_1)|.\tag{3.3}$$

Proof. We begin as in Aronson's proof for the case $\lambda=0$, [A1, p. 463] (see also [K3]). Let $N=\sup_{x\in\mathbb{R}}v(x,t)$. We choose a C^2 function ϕ: $[0,1]\to[0,N]$ strictly increasing, concave and such that $(\phi''/\phi')'\leqq 0$; further specifications will be given below. We also take a cut-off function $\zeta\in C(\overline{S}_T)\cap C^\infty(S_T)$ such that $0\leqq\zeta\leqq 1$ and $\zeta=0$ for $t=0$ or $|x|\geqq c$, c being some positive constant. Setting $v=\phi(w)$ we obtain from (3.1)

$$w_t=(m-1)\phi w_{xx}+\left[(m-1)\phi\frac{\phi''}{\phi'}+\phi'\right]w_x^2-c\frac{\phi^\beta}{\phi'}.$$

Now we differentiate with respect to x, multiply by $w_x\zeta^2$ and consider a point (x_0,t_0) of S_T where the function (x_0,t_0) of $S_T$where the function $z=\zeta^2w_x^2$ attains a maximum, so that we have $z_t\geqq 0$, $z_x=0$, $z_{xx}\leqq 0$ at (x_0,t_0) and $t_0>0$ (unless $z\equiv 0$, a case we may disregard). It then follows that at such a point:

(3.4)

$$\left[-m\phi''-(m-1)\phi\left(\frac{\phi''}{\phi'}\right)'\right]\zeta^2w_x^4\leqq\left[\zeta\zeta_t+2(m-1)\phi\zeta_x^2-(m-1)\phi\zeta\zeta_{xx}-c\left(\frac{\phi^\beta}{\phi'}\right)'\zeta^2\right]$$
$$\cdot w_x^2-\left[(m+1)\phi'+2(m-1)\phi\frac{\phi''}{\phi'}\right]\zeta\zeta_xw_x^3.$$

Now set

$$a_1=\max|\zeta_t|,\quad a_2=\max|\zeta_x|,\quad a_3=\max|\zeta_{xx}|,$$

and assume that there are positive constants b_i, $i=1,\cdots,4$ such that

$$0<Nb_1\leqq\phi'\leqq Nb_2,\quad \phi''\leqq -Nb_3,\quad |\phi''/\phi'|\leqq b_4. \tag{3.5}$$

Taking also into account that $(\phi^\beta/\phi')'\geqq 0$, we can disregard this term. We then get:

$$\zeta^2w_x^4\leqq c_1w_x^2+c_2\zeta|w_x|^3,$$

where

$$c_1=\frac{1}{Nmb_3}(a_1+N(m-1)a_3+2N(m-1)a_2^2),$$

$$c_2=\frac{a_2}{mb_3}((m+1)b_2+2(m-1)b_4).$$

Since for every $\delta>0$, $c_2\zeta|w_x|^3\leqq\delta\zeta^2w_x^4+c_2^2w_x^2/4\delta$, we get

$$(1-\delta)\zeta^2w_x^4\leqq\left(c_1-\frac{c_2^2}{4\delta}\right)w_x^2.$$

Therefore for every $(x,t)\in S_T$ we have

$$z(x,t)\leqq\max z\leqq\frac{1}{1-\delta}\left(c_1+\frac{c_2^2}{4\delta}\right). \tag{3.6}$$

Since this bound depends on ζ and ϕ, we now fix a point $(x_1,t_1)\in S_T$ and make a suitable selection of these functions. We begin with ζ: we choose

$$\zeta_n(x,t)=\frac{t}{t_1}\psi\left(\frac{x-x_1}{n}\right), \tag{3.7}$$

where $\psi \in C_0^\infty(R)$ satisfies $0 \leqq \psi \leqq 1$, $\psi(x)=1$ if $|x| \leqq 1$, $\psi(x)=0$ if $|x| \geqq 2$. Plugging ζ_n into (3.4) we obtain (3.6) with constants c_{1_n}, c_{2_n} depending on a_{1_n}, a_{3_n}. Now $a_{1_n}=1/t_1$ whereas a_{2_n}, $a_{3_n} \to 0$ as $n \to \infty$. Thus passing to the limit we obtain from (3.6)

$$w_x^2(x_1,t_1)=z(x_1,t_1) \leqq \frac{1}{Nmb_3t_1}$$

and noting that $v_x=\phi'(w)w_x$ and $x_1, t_1>0$ were arbitrary, we arrive at

$$v_x^2(x,t) \leqq \frac{Nb_2^2}{mb_3t} \quad \text{for every } (x,t) \in S_T. \tag{3.8}$$

Now we choose a suitable ϕ. For this purpose we consider the simplest form that satsfies $(\phi''/\phi')' \leqq 0$ as well as (3.5), namely

$$\phi(r)=Nr(a-br), \qquad a,b>0.$$

Then $\phi(1) \geqq N$ if $a \geqq b+1$, ϕ is strictly increasing in $[0,1]$ if $a \geqq 2b$ and we have $b_2=a$, $b_3=2b$. Thus (3.8) becomes

$$v_x^2(x,t)=\frac{Na^2}{2mbt}. \tag{3.9}$$

The second member above is to be minimized with respect to a,b subject to the conditions $a,b>0$, $a \geqq b+1$, $a \geqq 2b$. This happens for $a=2$, $b=1$ for which (3.9) gives (3.2).

To prove (3.3) we may assume that $t_1=0$ and $w \in C(\bar{S}_T)$ by displacing the origin of the times. We then repeat the previous argument taking now

$$\zeta_n(x,t)=\psi\left(\frac{x-x_1}{n}\right)$$

to conclude that at any interior maximum of w_x^2, $w_x=0$ so that the maximum is achieved at $t=0$.

Proof of Theorem 1. Let us consider first the case $n \geqq 1$. Then for every real x and $0 \leqq t_1 \leqq t_2$, $u(x,t_1)>0$ implies $u(x,t_2)>0$ (see [Kn2], [Ke2]). Therefore if we assume that $u_0(x)$ is positive everywhere, so is $u(x,t)$ in S and, as stated in §1, Theorem 0, $u \in C^\infty(S)$. Since the change of variables $v=(m/(m-1))u^{m-1}$ transforms (0.1) into (3.1) with

$$\beta=\frac{m+n-2}{m-1}, \qquad c=\lambda m\left(\frac{m-1}{m}\right)^\beta, \tag{3.10}$$

it then follows that estimates (2.5), (2.6) are a consequence of Lemma 3.1. When $u_0(x)$ is only nonnegative, we just approach it by strictly positive data $u_0(x)$ such that $u_{0j}(x) \to u_0(x)$ as $j \to \infty$ uniformly on compact subsets of $\mathbb{R}$ and use part iv) in Theorem 0 to conclude the result.

When $2-m \leqq n<1$ we no longer can deal with smooth positive solutions of (0.1), (0.2), since solutions may disappear after a finite time (see [K2]). We then start from problems (1.3) with, say, $\varepsilon=1/j$ for $j=1,2,\cdots$. Then the estimates hold true for each $u_j(x,t)$, since in applying Bernstein's technique to the equation for $v_\varepsilon=(m/(m-2))u_\varepsilon^{m-1}$ it happens, as in Lemma 3.1, that the perturbation which appears with respect to the case $\lambda=0$ has the right sign and may be dropped (see for instance [Ke2, pp. 1957–1958] for this detail). Therefore letting $j \to \infty$ (2.5), (2.6) follow.

Remark. Estimate (3.2) does not depend on λ, so that it even applies to the case $\lambda=0$. However the constant in (3.2) is not best possible; for instance when $\lambda=0$ one has

$$v_x^2(x,t)\leqq\frac{2}{m+1}\|v(x,0)\|_\infty t^{-1}, \tag{3.11}$$

(see Vâzquez [V2]), a result which is sharp in view of the explicit solutions (2.3).

4. Semiconvexity of the pressure. This section is devoted to the proof of Theorem 2. To this aim we shall proceed in three steps.

i) The core of the proof lies in the following formal argument. Let $v=(m/(m-1))u^{m-1}$ where u is a smooth positive solution of (0.1) and c, β are given in (3.10). As in Aronson and Bénilan [AB] we set $p=v_{xx}$ and differentiate twice in (3.1) with respect to x to get

$$\begin{aligned} p_t=(m-1)vp_{xx}+2mv_xp_x+(m+1)p^2-c\beta v^{\beta-1}p \\ -c\beta(\beta-1)v^{\beta-2}v_x^2. \end{aligned} \tag{4.1}$$

Let us consider now the differential operator:

$$L\theta=(m-1)v\theta_{xx}+2mv_x\theta_x-c\beta v^{\beta-1}\theta+(m+1)\theta^2-\theta_t. \tag{4.2}$$

It then follows from (4.1) and estimates (2.10), (2.11) that

$$Lp=c\beta(\beta-1)v^{\beta-1}v_x^2\leqq\frac{2^\sigma(m+n-2)}{m(m-1)t^2}.$$

Take now $\hat{p}(x,t)=-k/t$ where $k>0$ is some positive constant to selected later. One then has

$$p\geqq\hat{p} \quad \text{at } t=0 \tag{4.3}$$

and on the other hand

$$L\hat{p}=(m+1)\hat{p}^2-c\beta v^{\beta-1}\hat{p}-\hat{p}_t\geqq(m+1)\frac{k^2}{t^2}-\frac{k}{t^2}.$$

Thus we obtain $L\hat{p}\geqq Lp$ in S if

$$(m+1)k^2\geqq k+\frac{2^\sigma(m+n2)}{m(m-1)}. \tag{4.4}$$

A simple choice for k is, for instance,

$$k=\frac{1}{m+1}+\left(\frac{2^\sigma(m+n-2)}{m(m^2-1)}\right)^{1/2}. \tag{4.5}$$

By the maximum principle it then follows that $p\geqq\hat{p}=-k/t$ in S.

ii) We next show how to justify this argument in the case where v is smooth in S_T, $T>0$, and satisfies there

$$v\leqq\Lambda, \quad |v_x|\leqq\mu, \quad v_{xx}\geqq-\nu \tag{4.6}$$

where Λ, μ, ν are some positive constants. Fix $\tau>0$ such that $\tau<\min(k/\nu, T)$, where k is given in (4.5). Now write $S_{\tau,T}=\mathbb{R}\times(\tau,T\,|$ and define

$$p^*(x,t)=-\frac{k}{t-\tau'}$$

where $0<\tau'<\tau$. Then if L is the differential operator in (4.2) one has upon replacing t by $(t-\tau')$ in estimates (2.10), (2.11)

$$Lp^*\geqq Lp \quad \text{in } S_{\tau,T},$$

whereas, by the choice of τ and the fact that $\tau'<\tau$, one has $p^*(x,t)\leqq p(x,\tau)$. From the maximum principle, cf. [IKO, Thm. 8], it follows that $p^*\leqq p$ in $S_{\tau,T}$ (though the operators in [IKO] are linear, the proof applies unchanged). Then let $\tau\downarrow 0$ to conclude, since $T>0$ is arbitrary.

iii) It only remains to show that under our current assumptions on n,m and u_0, $v(x,t)$ can be approximated by smooth solutions of (3.1) $v_j(x,t)$ $(j=1,2,\cdots)$ satisfying (4.6) with bounds possibly depending on j. To this aim we select a sequence $v_0^j(x)$ as follows:

> For each fixed $j, v_0^j(x)\in C^\infty(\mathbb{R}); v_0^j(x)>1/j$ in $\mathbb{R}$, $v_0^j(x)$, $|v_{0_x}^j(x)|$ and $|v_{0_{xx}}^j(x)|$ are bounded in $\mathbb{R}$ and $v_0^j(x)\rightarrow v_0(x)$ uniformly on compact subsets of $\mathbb{R}$.

We consider the Cauchy problem corresponding to (3.1) (where c,β are given in (3.10)) with initial datum v_0^j. Since the generalized solution $v_j(x,t)$ is such that $v_j(x,t) >(j^{n-1}+\lambda(n-1)t)^{-(1/n-1)}$, the exact solution with initial data $1/j$, the equation is then uniformly parabolic in each strip S_T with $T>0$. Then by classical results [LSU, Chap. V, Thm. 8] $v_j(x,t)$ is smooth and satisfies (4.6). Hence $p_j=v_{jxx}$ satisfies (2.12) and letting $j\rightarrow\infty$ the same is true for p.

Remark. The constant k in (4.5), which is independent of λ, is clearly not the best possible. This is not surprising, in view of our remark in §3, since we use (3.2) to calculate it. When $\lambda=0$, the best constant is $1/(m+1)$, as shown in [AB]. For $\lambda>0$ and $n=m$ one also can check (2.12) against the explicit solutions obtained by Bertsch, Kersner and Peletier [BKP2] which are of the form:

$$(y(x,t))^{m-1}=\frac{1}{f(t)}\left\{\rho-\frac{ch(\alpha x)-1}{g(t)}\right\}_+$$

where $\rho>0$ is arbitrary, $\alpha=\sqrt{\lambda}((m-1)/m)$ and f,g are positive increasing functions which in particular satisfy

$$(fg)'=a(\rho g+1) \quad \text{with } a=\frac{\lambda(m^2-1)}{m}.$$

In the support of y we thus have for $\bar{v}=(m/(m-1))y^{m-1}$

$$\bar{v}_{xx}=-\frac{\lambda(m-1)}{m}\cdot\frac{ch(\alpha x)}{fg}\geqq-\frac{\lambda(m-1)}{m}\frac{\rho g+1}{fg}.$$

Now since $(fg)(t)\geqq a\int(\rho g+1)\,dt\geqq at$ and $\rho/f=\sup y^{m-1}$ one gets, using (2.10),

$$\bar{v}_{xx}\geqq-\frac{k_1}{t} \quad \text{where } k_1=\left(\frac{1}{m+1}+\frac{1}{m}\right),$$

whereas our estimate reads

$$\bar{v}_{xx} \geqq -\frac{k}{t} \quad \text{where } k = \left(\frac{1}{m+1} + \frac{4}{\sqrt{m(m+1)}} \right).$$

5. The free boundary. This section is devoted to the study of the propagation properties of the solutions of problem (0.1), (0.2) under the hypotheses $m>1$, $n \geqq m$, u_0 continuous, nonnegative and bounded, plus the support condition (0.6):

$$\sup\{x:\ u_0(x)>0\} = d < +\infty.$$

Under the stronger hypothesis that u_0 is positive in a finite interval, say $(-d,d)$ with $0<d<+\infty$, and zero otherwise, the set $P=P[u] \equiv \{(x,t) \in S:\ u(x,t)>0\}$ has been studied by several authors: Kalashnikov [K2], Kersner [Ke1], Knerr [Kn2], $\cdots$. They prove that there exist two functions $\zeta_1(t)$, $\zeta_2(t) \in C[0,\infty) \cap C^{0,1}(0,\infty)$ such that $\zeta_1(0) = (-1)^i d$ $(i=1,2)$ and $P(t) \equiv \{x \in \mathbb{R}:\ (x,t) \in P\}$ is the interval $(\zeta_1(t), \zeta_2(t))$. In addition $(-1)^i \zeta_i(t)$ is nondecreasing and $|\zeta_i(t)| \to \infty$ as $t \to \infty$ for $i=1,2$. Actually these properties are true for $n \geqq m$, but not for $n<m$ (cf. [K2], [Ke3]). The curves $x=\zeta_i(t)$ are called *interfaces* or *free boundaries*. The proofs of these results can be easily adapted to our situation; then we obtain the existence of a *right-hand interface* $x=\zeta(t)= \sup\{x:\ u(x,t)>0\}$ with the above properties.

In the nonabsorption case $\lambda=0$ much more is known about the free boundaries: equation satisfied on the interface, C^1 regularity of $\zeta(t)$ (cf. [A1], [Kn1], [CF]). It is possible that the interface remains stationary for t less than or equal to a certain time $t^*[0,\infty)$, called *waiting time*, after which it moves with positive velocity: $\zeta'(t)>0$ if $t>t^*$ [K1], [Kn1]. The existence of a waiting time has been characterized by one of the authors [V3], and the behavior of ζ near t^* is studied in [ACK], [LOT] and [ACV].

It is our purpose to use the regularity results of the preceding chapters to prove similar facts for the absorption case $\lambda>0$. Our first result is the following theorem.

THEOREM 3. *Let u be the solution of* (0.1), (0.2) *under the conditions above, and let $x=\zeta(t)$ be the right-hand free boundary. For every $t>0$ the right-hand derivative $D^+\zeta(t)$ exists and*

$$D^+\zeta(t) = -v_x(\zeta(t),t), \tag{5.1}$$

where $v_x(\zeta(t),t)$ is understood as $\lim v_x(x,t)$ *as $x \to \zeta(t)$, $x<\zeta(t)$.*

Proof. The line of argument parallels that of the nonabsorption case $\lambda=0$, and uses Theorem 2 followed by suitable comparisons.

i) Since $v \leqq C$ and $v_{xx} \geqq -C \cdot t^{-1}$ for some $C>0$, the function $x \mapsto v(x,t)+Cx^2/2t$ is convex in $\mathbb{R}$ for every $t>0$, and hence it has one-sided derivatives at every point. Therefore the limit $v_x(\zeta(t),t)$ appearing in (5.1) exists for each $t>0$.

ii) To prove (5.1) at $t=t_0>0$, we first consider the solution $\hat{u}(x,t)$ of the problem:

$$\begin{aligned} &\hat{u}_t = (\hat{u}^m)_{xx} \quad \text{for } x \in \mathbb{R}, \quad t>t_0, \\ &\hat{u}(x,t_0) = u(x,t_0) \quad \text{for } x \in \mathbb{R}. \end{aligned} \tag{5.2}$$

It is clear (see e.g. [Ke2, Thm. 3]) that $\hat{u}$ is a supersolution for our problem when $t \geqq t_0$, so that by Theorem 0, part iii) $\hat{u}(x,t) \geqq u(x,t)$ for such t. Hence $\hat{\zeta}(t) \geqq \zeta(t)$ if $t>t_0$ and $\hat{\zeta}(t_0) = \zeta(t_0)$, $\hat{\zeta}$ being the interface for $\hat{u}$. Now for problem (5.2) the result holds, $D^+\hat{\zeta}(t_0) = -v_x(\hat{\zeta}(t_0),t_0)$ and therefore

$$\limsup_{h \to 0} \frac{1}{h}(\zeta(t_0+h) - \zeta(t_0)) \leqq -v_x(\hat{\zeta}(t_0),t_0). \tag{5.3}$$

Next we consider the problem:

$$(5.4)\qquad \begin{cases} \bar{u}_t=(\bar{u}^m)_{xx}-\lambda_0\bar{u} & \text{if } x\in\mathbb{R},\ t>t_0,\\ \bar{u}(x,t)=u(x,t_0) & \text{if } x\in\mathbb{R}, \end{cases}$$

where $\lambda_0=((n-1)t_0)^{-1}$, so that by (2.10) $\lambda u^n \leqq \lambda_0 u$. Then $u(x,t)$ is a supersolution for (5.4) for $t\geqq t_0$. On the other hand (5.4) can be converted into a nonabsorption problem by means of (2.13) and it is not difficult to see that $D^+\bar{\zeta}(t_0)=-v_x(\bar{\zeta}(t_0),t_0)$ also holds. Therefore

$$(5.5)\qquad \liminf_{h\to 0}\frac{1}{h}(\zeta(t_0+h)-\zeta(t_0))\geqq D^+\bar{\zeta}(t_0)=-v_x(\zeta(t_0),t_0)$$

and the theorem is proved.

We now extend the characterization of the existence of a positive waiting-time to our absorption case.

THEOREM 4. *Let $u(x,t)$, $u_0(x)$, $\zeta(t)$, m as above and let $n\geqq 1$. Let $t^*=\sup\{t\geqq 0$: $\zeta(t)=d\}$. Then t^* is positive if and only if*

$$(5.6)\qquad \sup_{x<d}\left((d-x)^{-(m+1)/(m-1)}\int_x^d u_0(s)\,ds\right)<+\infty.$$

Proof. If (5.6) holds we consider the solution $\hat{u}$ of (5.2) with $t_0=0$. By the results of [V3] we then have a positive waiting-time $\hat{t}_i^*$ for $\hat{\zeta}$. Since $\hat{u}\geqq u$, we conclude that $t_i^*\geqq \hat{t}_i^*>0$.

Conversely, if u does not satisfy (5.6), the solution of (5.2) with $t_0=0$ has a zero waiting time. By the change of variables (2.13) the same is true for the solution $\hat{u}$ of (5.4) with $t_0=0$ and λ_0 arbitrary. But choosing λ_0 large enough $\bar{u}(x,t)<u(x,t)$ whence $t_i^*<\bar{t}_i^*=0$.

Our next result deals with the behavior of $\zeta(t)$ after a waiting time.

THEOREM 5. *Under the assumptions of Theorem* 3, $\zeta(t)\in C^1(t^*,\infty)$, $\zeta'(t)>0$ *for* $t>t^*$ *and the function*

$$(5.7)\qquad \zeta'(t)\cdot t^{\rho}$$

is nondecreasing when $t>t^*$ for a certain $\rho=\rho(m,n)>0$ (see (5.13)).

Remark. When $\lambda=0$ it has been proved in [V1] that (5.7) holds with $\rho_0=m/(m+1)$ which is sharp. In particular for the Barenblatt solutions (2.3) $\zeta'(t)t^{\rho_0}$ is constant.

Proof of Theorem 5. i) We first prove a weak version of (5.7). At any point $t_1>t^*$ with $D^+\zeta(t_1)>0$ we adapt a subsolution $\bar{u}(x,t)$ to (0.1) in the strip $\mathbb{R}\times(t_1,\infty)$ satisfying $\bar{u}(x,t_1)\leqq u(x,t_1)$ and such that $\bar{u}$ has a good contact with u at $x=\zeta(t_1)$; namely we require that the corresponding interface and pressure, $\bar{\zeta}(t)$ and $\bar{v}(x,t)$, satisfy

$$(5.8)\qquad \begin{aligned} &\bar{\zeta}(t_1)=\zeta(t_1),\\ &\bar{\zeta}'(t_1)=D^+\zeta(t_1) \qquad (\text{i.e. } \bar{v}_x(\zeta(t_1),t_1)=v_x(\zeta(t_1),t_1)),\\ &\bar{v}_{xx}(x,t_1)=-\frac{k}{t_1}, \qquad k \text{ as in } (2.12). \end{aligned}$$

To this aim we start from $\hat{u}(x,t)$ given as in (2.3) by

$$(5.9)\qquad \hat{u}(x,t)=t^{-1/(1+m)}\left[p-\frac{m-1}{2m(m+1)}(x\cdot t^{-1/(1+m)})^2\right]_+^{1/(m-1)}$$

where $\rho > 0$ is to be selected presently. Now for $\lambda_1 > 0$ fixed we consider the solution of $u_t = (u^m)_{xx} - \lambda_1 u$ obtained from (5.9) through transformation (2.13). This reads

$$\bar{u}_{p,\lambda_1}(x,t) = e^{-\lambda_1 t}\left(\frac{1}{\lambda_1(m-1)}\right)^{-1/(m+1)}\left(1-e^{-\lambda_1(m-1)t}\right)^{-1/(m+1)}$$

$$\cdot\left[p - \frac{m-1}{2m(m+1)}\left(\frac{1}{\lambda_1(m-1)}\right)^{-2/(m+1)}\left(1-e^{-\lambda_1(m-1)t}\right)^{-2/(m+1)}x^2\right]_+^{1/(m-1)}.$$

We now set

$$\bar{u}(x,t) = \bar{u}_{p,\lambda_1}(x-x_0, t-\tau) \equiv \bar{u}(x,t;\lambda_1,p,x_0,\tau) \tag{5.10}$$

with

$$\lambda_1 = ((n-1)t_1)^{-1}$$

so that by estimate (2.10) $\lambda_1 \bar{u} \geqq \lambda \bar{u}^n$ and $\bar{u}$ is a subsolution to (0.1). Next we determine x_0, τ, p from (5.8). This last can be done in a unique way. In particular τ is obtained from the third condition in (5.8) which reads

$$\exp\left(\frac{m-1}{n-1}\left(1-\frac{\tau}{t_1}\right)\right) - 1 = \frac{m-1}{k(m+1)(n-1)},$$

whereas x_0, p are determined from

$$\zeta(t_1) - x_0 = \frac{t_1}{k}\cdot D^+\zeta(t_1),$$

$$\frac{t_1}{k}\cdot D^+\zeta(t_1) = \left(\frac{2m(m+1)p}{m-1}\right)^{1/2}\left(\frac{(n-1)t_1}{m+1}\right)^{1/(m+1)}\left(1-\exp\left(\frac{1-m}{n-1}\left(1-\frac{\tau}{t_1}\right)\right)\right)^{1/(m+1)}.$$

It follows easily from (5.8) that $u(x,t_1) \geqq \bar{u}(x,t_1)$. By the choice of λ_1, $u(x,t) \geqq \bar{u}(x,t)$ for $t \geqq t_1$ and therefore $\zeta(t_1+h) \geqq \bar{\zeta}(t_1+h)$ for any $h > 0$. Hence

$$\zeta(t_1+h) - \zeta(t_1) - hD^+\zeta(t_1) \geqq \bar{\zeta}(t_1+h) - \bar{\zeta}(t_1) - h\bar{\zeta}'(t_1). \tag{5.11}$$

On the other hand it follows from (5.10) that

(5.12)

$$\bar{\zeta}(t) = x_0 + \left(\frac{2m(m+1)p}{m-1}\right)^{1/2}\left(\frac{(n-1)t_1}{m+1}\right)^{1/(m+1)}\left(1-\exp\left(\frac{(1-m)(t-\tau)}{(n-1)t_1}\right)\right)^{1/(m+1)}$$

so that $\bar{\zeta}(t)$ is a C^∞-function for $t > \tau$, and the second member in (5.11) can be written as $h^2/2\cdot\bar{\zeta}''(t_1) + O(h^3)$. Besides one easily checks on (5.12) that:

$$\bar{\zeta}''(t_1) = -\rho\frac{\bar{\zeta}'(t_1)}{t_1} \quad \text{with } \rho = \left(\frac{m-1}{n-1} + mk\right). \tag{5.13}$$

We now divide (5.11) by h^2 to obtain

$$F_h(t) = \frac{2}{h^2}\left[\zeta(t+h) - \zeta(t) - hD^+\zeta(t)\right] \geqq -\frac{\rho}{t}D^+\zeta(t) + O(h) \tag{5.14}$$

at any point t_1 where $D^+\zeta(t_1)>0$. It is obvious that (5.14) also holds at t_1 if $D^+\zeta(t_1)=0$ (with $O(h)\equiv 0$ in this case). On the other hand, it follows from (5.12) that $O(h)$ is uniform in t. Therefore letting $h\to 0$ we obtain

$$\zeta''(t) \geqq -\frac{\rho}{t}D^+\zeta(t), \qquad t>0, \tag{5.15}$$

where ζ'' is to be understood in the sense of distributions. Since $\zeta(t)$ is locally Lipschitz continuous in $(0,\infty)$, $D^+\zeta(t)=\zeta'(t)$ a.e. and (5.15) may be rewritten as

$$(\zeta'(t)\cdot t^\rho)' \geqq 0 \quad \text{in } \mathscr{D}'(0,\infty). \tag{5.16}$$

Assume now that $\rho<1$. Then (5.16) means that the function $\eta(\tau)=\zeta(t)$ with $\tau=(1-\rho)^{-1}t^{1-\rho}$ is convex for $0\leqq\tau<\infty$. Therefore the lateral limits $D^+\eta(\tau)$, $D^-\eta(\tau)$ exist for every $\tau>0$, they are nondecreasing functions of τ, and

$$D^+\eta(\tau)\geqq D^-\eta(\tau)\geqq D^+\eta(\tau-h) \quad \text{for } \tau>\tau-h>0.$$

Since $D^{\pm}\eta(\tau)=D^{\pm}\zeta(t)\cdot t^\rho$, we then conclude that

$$\begin{aligned} &D^+\zeta(t)\cdot t^\rho,\ D^-\zeta(t)\cdot t^\rho \quad \text{are nonnegative and nondecreasing,} \\ &D^+\zeta(t)\geqq D^-\zeta(t)\geqq D^+\zeta(t-h)\left(1-\frac{h}{t}\right)^\rho \quad \text{for every } t>t-h>t^*. \end{aligned} \tag{5.17}$$

The same result is obtained if $\rho\geqq 1$, but now we have to take $\tau=\log t$ if $\rho=1$ and $\tau=-(\rho-1)t^{-(\rho-1)}$ if $\rho>1$. Note that (5.17) implies that once the interface starts it never stops.

ii) We have yet to prove that $D^+\zeta(t)=D^-\zeta(t)$ for every $t>t^*$ to obtain that $\zeta\in C^1(t^*,\infty)$ and (5.7) holds. But once the estimates on v_x (2.5) and v_{xx} (2.12) have been established, this can be done by repeating, with some minor modifications, the arguments in [ACK, Thm. B], for the case $\lambda=0$. We just indicate here for completeness an important auxiliary tool that is used in the proof and has some independent interest.

LEMMA 5.1. *Let $v=(m/(m-1))u^{m-1}$, where $u=u(x,t)$ is the solution of* (0.1), (0.2) *under our current hypotheses, and assume that*

$$-v_x(\zeta(t_0),t_0)=\gamma$$

for some $t_0>0$, $\gamma>0$. Then in a neighborhood of $(\zeta(t_0),t_0)$ one has

$$v(x,t)=L_\gamma(x-\zeta(t_0),t-t_0)+o(|x-x_0|+|t-t_0|)$$

where

$$L_\gamma(x,t)=(\gamma^2 t-\gamma x)_+.$$

This result is in [ACK, Prop. 2.3] when $\lambda=0$. The case $\lambda>0$ offers no difficulties.

6. Asymptotic behavior. In this section we study the asymptotic behavior of the solution of (0.1), (0.2) with $1<m\leqq n$. We also study the behavior of the interface $\zeta(t)$ when u_0 satisfies (0.6).

We shall use the following notation: If $f(t)$, $g(t)$ are nonnegative functions defined for all large $t>0$ we write:

i) $f(t)=O(g(t))$ if there exists a constant $C>0$ such that $f(t)\leqq Cg(t)$ for all large enough t.

ii) $f(t)\sim g(t)$ if $f(t)=O(g(t))$ and $g(t)=O(f(t))$.

Our main contribution is the description of different behaviors if $n>m+2$.

6.1. The case $n>m+2$. It follows from (2.10) that $\|u(\cdot,t)\|_\infty=O(t^{-1/(n-1)})$; this bound is due to the effect of the absorption and holds for all initial data. A stronger bound $O(t^{-1/(m+1)})$ holds for all solutions with intial data $u_0\in L^1(\mathbb{R})$, cf (2.7), which coincides with the one obtained in the pure diffusion case $\lambda=0$. We show next that both rates can be exact.

THEOREM 6. *Let $u_0\in L^1(\mathbb{R})$. Then*

$$u(\cdot,t)\sim t^{-1/(m+1)} \tag{6.1}$$

uniformly on sets of the form $|x|\leqq ct^{1/(m+1)}$, with $c>0$ small enough. If (0.6) *holds, one then has*

$$\zeta(t)\sim t^{1/m+1}, \qquad \zeta'(t)=O(t^{-m/(m+1)}). \tag{6.2}$$

Proof. The estimates on u_0 and ζ when u_0 has compact support are due to Kersner [Ke1]. In fact he constructs subsolutions of the form

$$w(x,t)=\sigma(t+\tau)^{-1/(m-1)-\varepsilon}\left(\rho-x^2(t+\tau)^{\varepsilon(m-1)-2/(m+1)}\right)_+^{1/(m-1)}, \tag{6.3}$$

where $\varepsilon>0$ must be small but otherwise arbitrary, and θ,τ,ρ are positive constants that can be selected independently of ε. It is clear by comparison of u with w for suitably chosen θ,τ,ρ that the rates in (6.1), (6.2) serve as lower bounds for every solution.

To obtain the upper bounds we observe that (6.1) comes from (2.7) and that (6.2) is a consequence of estimate (2.9) for v_x with $p=1$ and formula (0.5); we obtain thus $\zeta'(t)=O(t^{-m/(m+1)})$ and by integration the corresponding one for $\zeta(t)$.

We now obtain solutions with maximal rates.

THEOREM 7. *Assume that $u_0(x)\geqq\delta>0$ for all $x<0$, $|x|$ large. Then*

$$\lim_{t\to\infty}\|u(\cdot,t)\|_\infty t^{1/(n-1)}=(\lambda(n-1))^{-1/(n-1)} \tag{6.4}$$

and $u\sim t^{-(1/(n-1))}$ uniformly on sets of the form $x<ct^{n-m/(2(n-1))}$, $c>0$ small. If (0.6) *holds one has*

$$\zeta(t)\sim t^{n-m/(2(n-1))}, \qquad \zeta'(t)=O\left(t^{-\frac{n+m-2}{2(n-1)}}\right). \tag{6.5}$$

Proof. To get the upper bounds we argue as before: using (2.10) and (0.5) we obtain the bound for ζ', and by integration the one for ζ. The corresponding case in (6.4) comes also from (2.10).

For the lower bounds we construct a new subsolution as follows: Let $z=z(x,t)$ be the solution of the porous medium equation $z_t=(z^m)_{xx}$ with initial datum $z_0(x)$ defined by $z_0(x)=a>0$ if $x<0$, $z_0(x)=0$ if $x>0$. It is shown in [V3] that for each a there exists a unique generalized solution of the form

$$z(x,t)=f(x,t^{-1}/2), \tag{6.6}$$

where f: $\mathbb{R}\to\mathbb{R}$ satisfies

$$\begin{aligned}&(f^m)''(\xi)+\frac{1}{2}\xi f'(\xi)=0, \qquad \xi\in\mathbb{R},\\ &f(-\infty)=a, \qquad f(+\infty)=0.\end{aligned} \tag{6.7}$$

Moreover there exists $\xi_0 = \xi_0(a,m) > 0$ such that $0 < f(\xi) < a$ for $x < \xi_0$ and $f(\xi) = 0$ for $x \geqq \xi_0$. Now we set

$$w(x,t) = \sigma(t+\tau)^{-1/(n-1)} f\left((x-x_0)(t+\tau)^{-n-m/(2(n-1))}\right) \tag{6.8}$$

for some constants σ, x_0, τ such that σ, $\tau > 0$. We now impose the condition

$$\begin{aligned} Lw &\equiv w_t - (w8Um)_{xx} + \lambda w^n \\ &= \sigma(t+\tau)^{-(n/(n-1))}\left[-\frac{n-m}{2(n-1)}\eta f'(\eta) - \frac{1}{n-1} f(\eta) \right. \\ &\qquad \left. - \sigma^{m-1}(f^m)''(\eta) + \lambda\sigma^{n-1} f^{n-1}(\eta) \right] \\ &\leqq 0 \end{aligned} \tag{6.9}$$

where $\eta = (x-x_0)(t+\tau)^{-((n-m)/(2(n-1))}$. Now (6.9) is satisfied if we choose

$$\sigma = \left(\frac{n-m}{n-1}\right)^{1/(m-1)}, \qquad a\sigma = ((n-1)\lambda)^{-1/(n-1)}.$$

In order that $w(x,0) \leqq z_0(x)$ we choose appropriately x_0 and τ. It then follows that

$$u(x,t) \geqq w(x,t), \qquad \zeta(t) \geqq \zeta_w(t) = x_0 + \xi_0(t+\tau)^{n-m/2(n-1)} \tag{6.10}$$

where $\zeta_w(t)$ is the interface corresponding to $w(x,t)$. Taking into account that $\|w(x,t)\|_\infty = \sigma a(t+\tau)^{-(1/(n-1))}$ the estimates follow.

Remarks. 1) The subsolution (6.8) can be constructed for $n > m$. Therefore Theorem 7 holds for this range of n's.

2) We have shown the two extreme cases. Results for intermediate classes of initial data are not difficult to obtain in many instances. To point out a simple fact if $u_0 \in L^p(\mathbb{R})$, $1 < p < \infty$, we have by (2.7)

$$u(x,t) = O(t^{-\theta}), \qquad \theta = (2p+m-1)^{-1}, \tag{6.11}$$

uniformly in x. Arguing as above it follows that

$$\zeta'(t) = O(t^{-(p+m-1/(2p+m-1))}), \qquad \zeta(t) = O(t^{+(p/(2p+m-1))}), \tag{6.12}$$

as $t \to \infty$.

3) In the nonabsorption case, $\lambda = 0$, a very detailed description of the asymptotic behavior is available, cf. [V3] and its references.

6.2. Case $m \leqq n < m+2$. In this case there is only one type of asymptotic behavior summarized in the following result that extends Theorem 7 (see Remark 1 above).

THEOREM 8. *Let $m < n < n+2$ (resp. $n = m$) and let u be a solution of* (0.1), (0.2). *Then as $t \to \infty$*

$$u(\cdot,t) \sim t^{-1/n-1} \tag{6.13}$$

uniformly on sets of the form $|x| \leqq ct^{n-m/(2(n-1))}$ (resp. $|x| \leqq c \log t$) for c small. Moreover if u_0 satisfies (0.6) *we have*

$$\zeta'(t) = O(t^{-((m+n-2)/(2(n-1)))}), \qquad \zeta(t) \sim t^{((n-m)/(2(n-1)))} \tag{6.14}$$
$$\times \left(\textit{resp. } \zeta'(t) = O(t^{-1}), \ \zeta(t) \sim \log t\right).$$

Proof. This result was essentially known. Thus in [Ke1], [BKP1], the lower estimates are obtained by means of suitable subsolutions. As to the upper bounds, (6.13) comes from (2.10) and the estimate $\zeta'(t)=O(t^{-((m+n-2)/(2(n-1)))})$ (resp. $\zeta'(t)=O(t^{-1})$) is obtained as in Theorem 6 above. By integration we obtain (6.14).

A related argument to obtain the upper bounds is explained in [H]. For another argument see [BKP1]. In both cases u_0 has compact support.

REFERENCES

[A1] D. G. ARONSON, *Regularity properties of flows through porous media*, SIAM J. Appl. Math., 17 (1969), pp. 461–467.

[A2] ______, *Regularity properties of flows through porous media: the interface*, Arch. Rat. Mech. Anal., 37 (1970), pp. 1–10.

[AB] D. G. ARONSON AND PH. BÉNILAN, *Régularité des solutions de l'équation des milieux poreux dans R^n*, C. R. Acad. Sci. Paris, 288 (1979), pp. 103–105.

[ACK] D. G. ARONSON, L. C. CAFFARELLI AND S. KAMIN, *How an initially stationary interface begins to move in porous medium flow*, this Journal, 14 (1983), pp. 639–658.

[ACV] D. G. ARONSON, L. A. CAFFARELLI AND J. L. VÁZQUEZ, *Interfaces with a corner point in one-dimensional porous medium flow*, Comm. Pure Applied Math., 38 (1985), pp. 375–404.

[B] G. I. BARENBLATT, *On some unsteady motions of a liquid or a gas in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 67–78 (Russian).

[Be] M. BERTSCH, *A class of degenerate diffusion equations with a singular nonlinear term*, Nonlinear Analysis, TMA 7 (1983), pp. 117–127.

[BKP1] M. BERTSCH, R. KERSNER AND L. A. PELETIER, *Sur le comportement de la frontière libre dans une équation en théorie de la filtration*, C. R. Acad. Sc. Paris, 295 (1982), pp. 63–66.

[BKP2] ______, *Positivity versus localization in degenerate diffusion equations*, to appear.

[CF] L. A. CAFFARELLI AND A. FRIEDMAN, *Regularity of the free boundary for the one-dimensional flow of a gas in a porous medium*, Amer. J. Math., 101 (1979), pp. 1193–1218.

[G] B. H. GILDING, *Hölder continuity of solutions of parabolic equations*, J. London Math. Soc., 13 (1976), pp. 103–106.

[H] M. A. HERRERO, *On the growth of the interfaces of a nonlinear degenerate parabolic equation*, in Contributions to Nonlinear P.D.E., C. Bardos et al., eds., Pitman Research Notes, no. 86 (1983), pp. 218–224.

[IKO] A. M. IL'IN, A. S. KALASHNIKOV AND O. A. OLEINIK, *Second order linear equations of parabolic type*, Russian Math. Surveys, 17 (1962), pp. 1–143.

[K1] A. S. KALASHNIKOV, *On the occurrence of singularities in the solutions of the non-steady seepage equation*, Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 440–444.

[K2] ______, *The propagation of disturbances in problems of nonlinar heat conduction with absorption*, Zh. Vychisl. Mat. i Mat. Fiz., 14 (4) (1974), pp. 891–905.

[K3] ______, *On the differential properties of generalized solutions of nonstationary filtration type*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 29 (1974), pp. 62–68.

[Ke1] R. KERSNER, *On the behavior when $t \to \infty$ of generalized solutions of a degenerate parabolic equation*, Acta Math. Acad. Sci. Hungaricae, 34 (1979), pp. 157–163 (Russian).

[Ke2] ______, *Degenerate parabolic equations with general nonlinearities*, Nonlinear Analysis TMA 4 (6) (1980), pp. 1043–1062.

[Ke3] ______, *The behavior of temperature fronts in media with nonlinear thermal conductivity under absorption*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 33 (5) (1978), pp. 44–51.

[Kn1] B. F. KNERR, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.

[Kn2] ______, *The behavior of the support of solutions of the equation of nonlinear heat conduction with absorption in one dimension*, Trans. Amer. Math. Soc., 249 (1979), pp. 409–424.

[Kr] S. N. KRUZHKOV, *Results concerning the nature of the continuity of solutions of parabolic equations and some of their applications*, Matematicheskie Zametki, 6 (1969), pp. 97–108.

[LOT] A. LACEY, J. R. OCKENDON AND A. TAYLER, *Waiting time solutions of a nonlinear diffusion equation*, SIAM J. Appl. Math., 42 (1982), pp. 1252–1264.

[LSU] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, AMS Translations, 1969.

[MP] L. K. Martinson and K. B. Pavlov, *Thermal localization in nonlinear heat conduction*, Zh. Vychisl. Mat. i Mat. Fiz., 12 (4) (1972), pp. 1048–1053.

[OKC] O. A. Oleĭnik and A. S. Kalashnikov and Chzou Y-L, *The Cauchy problem and boundary value problems for equations of the type of nonstationary filtration*, Izv. Akad. Nauk. SSSR. Ser. Mat., 22 (1958), pp. 667–704. (In Russian.)

[P] L. A. Peletier, *The porous medium equation*, in Application of Nonlinear Analysis in the Physical Sciences, H. Amman et al., eds., Pitman, New York (1981), pp. 229–241.

[V1] J. L. Vázquez, *Asymptotic behavior and propagation properties of the one-dimensional flow of gas in a porous medium*, Trans. Amer. Math. Soc., 277 (1983), pp. 507–527.

[V2] ______, *Behavior of the velocity of one-dimensional flow in porous media*, Trans. Amer. Math. Soc., 286 (1984), pp. 787–802.

[V3] ______, *The interface of one-dimensional flows in porous media*, Trans. Amer. Math. Soc., 285 (1984), pp. 717–737.

[Ve] L. Véron, *Effects régularisants de semigroupes non-linéaires dans les espaces de Banach*, Ann. Fac. Sci. Toulouse, 1 (1979), pp. 171–200.

[ZR] Y. B. Zeldovich and Y. P. Raizer, *Physics of shock waves and high-temperature hydrodynamic phenomena*, Academic Press, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

A VOLTERRA EQUATION WITH L^2-SOLUTIONS*

STIG-OLOF LONDEN†

Abstract. Consider the nonlinear Volterra equation $x'(t)+\int_0^t a(t-s)g(x(s))\,ds=f(t)$ where a is strongly positive and $f\in L^2(R^+)$. We formulate sufficient conditions for bounded solutions to be square integrable on R^+. The result generalizes earlier work by Staffans [Proc. Amer. Math. Soc., 78 (1980), pp. 213–217.]

Key words. Volterra equations, Fourier transforms

AMS(MOS) subject classifications. 45D05, 45G99, 45J05, 45M05

1. Introduction. In this note we consider the asymptotic size of bounded solutions of the (scalar, real) nonlinear Volterra equation

$$x'(t)+\int_0^t g(x(t-s))a(s)\,ds=f(t),\qquad t\in R^+\equiv[0,\infty),\quad x(0)=x_0, \tag{1.1}$$

and prove the following result:

THEOREM 1. *Let $x\in L^\infty(R^+)$ be a solution of* (1.1) *on R^+ satisfying*

$$\sup_{T>0}\int_0^T \varphi(t)(a*\varphi)(t)\,dt<\infty. \tag{1.2}$$

Also assume that

$$a \text{ is strongly positive definite}, \tag{1.3}$$

$$a'\in L^1(R^+), \tag{1.4}$$

$$g\in C(R),\qquad xg(x)>0,\quad x\neq 0,\qquad 0<\liminf_{|x|\downarrow 0}\frac{g(x)}{x}, \tag{1.5}$$

$$f\in L^2(R^+). \tag{1.6}$$

Then

$$x,x'\in L^2(R^+). \tag{1.7}$$

We define $x(t)$ to be a solution of (1.1) on R^+ if x is locally absolutely continuous and satisfies (1.1) almost everywhere on R^+. Above $\varphi(t)\overset{\text{def}}{=}g(x(t))$ and $*$ denotes convolution, thus $(a*\varphi)(t)=\int_0^t a(t-s)\varphi(s)\,ds$. By definition the kernel $a(t)$ satisfies (1.3) if and only if there exists $\varepsilon>0$ such that $a(t)-\varepsilon e^{-t}$ is positive definite. From (1.4) follows that $\hat a(\omega)=\lim_{\sigma\downarrow 0}\tilde a(\sigma+i\omega)=\lim_{\sigma\downarrow 0}\int_0^\infty a(t)e^{-\sigma t-i\omega t}dt$ is well defined for $\omega\neq 0$ and consequently (1.3) amounts to

$$\operatorname{Re}\hat a(\omega)\geqq \varepsilon[1+\omega^2]^{-1},\qquad \omega\neq 0, \text{ and some } \varepsilon>0.$$

The problem concerning the asymptotic size of solutions of (1.1) has earlier been considered by Staffans [3] (where references to prior work can be found) and our

* Received by the editors December 4, 1984; accepted for publication (in revised form) November 5, 1985. This work was done while the author was visiting the Institute of Mathematics of the University of Graz, Austria.

† Institute of Mathematics, Helsinki University of Technology, 02150 Espoo 15, Finland.

theorem extends his result. Thus we show by an elementary proof that the additional assumption

$$a' \in BV(R^+)$$

which is made in [3] to obtain (1.7) is in fact unnecessary.

As in [3] one may slightly extend Theorem 1:

THEOREM 2. *Let x be as in Theorem* 1 *and suppose* (1.2), (1.3), (1.5) *hold. Assume $a = b + c$ where $b \in L^1(R^+)$ and satisfies*

$$|\hat{b}(\omega)|^2 \leqq \beta \operatorname{Re} \hat{b}(\omega), \qquad \omega \in R,$$

for some $\beta \geqq 0$,

$$c \textit{ is positive definite}, \qquad c' \in L^1(R^+).$$

Finally let $f = f_1 + f_2 + f_3$ where $f_1 \in L^2(R^+)$, $f_2 \in BV(R^+)$, $f_3 \in L^\infty(R^+)$, $f_3' \in L^2(R^+)$. Then (1.7) *is satisfied.*

The proof of Theorem 2 closely follows that of Theorem 1 and we only note that by [3, Thm. 1] we may without loss of generality assume $c \not\equiv 0$ and that the integration by parts in (2.9) is performed only with $c(t)$.

To obtain $x \in L^\infty(R^+)$ (which is postulated in Theorem 1) one needs a somewhat different hypothesis. For completeness we state the following result which is a consequence of a result in [2].

THEOREM 3. *Let x be a solution of* (1.1) *on R^+ and assume that* (1.3) *holds. Suppose f, $f' \in L^2(R^+)$ and let $\limsup_{|x| \to \infty} G(x) = \infty$. Then $x \in L^\infty(R^+)$, and* (1.2) *holds.*

Above $G(x) = \int_0^x g(u)\,du$, $x \in R$. To prove Theorem 3 multiply (1.1) by $\varphi(t)$, integrate over $[0, T]$ and use [2, Prop. 4.1].

2. Proof of Theorem 1. Note first that $(d/dt)(a * \varphi)(t) \in L^\infty(R^+)$ from which (use (1.1), $x \in L^\infty(R^+)$ and $f \in L^2(R^+)$) follows $(a * \varphi)(t) \in L^\infty(R^+)$. Consequently x' is the sum of two functions belonging respectively to $L^\infty(R^+)$ and $L^2(R^+)$.

For any $T > 0$ let $y_T \in LAC(R^+)$ satisfy $y_T(t) = 0$, $t \notin (0, T)$ and be such that for $p = 1, 2$ and some c independent of T,

$$\int_0^T |x'(\tau) - y_T'(\tau)|^2 d\tau \leqq c, \qquad \int_0^T |x(\tau) - y_T(\tau)|^p d\tau \leqq c. \tag{2.1}$$

As $x \in L^\infty(R^+)$ and by the above decomposition of x this is possible. (Below c always denotes a constant independent of T. The actual value of c changes from line to line.)

Write $\varphi_T = \chi_{[0,T]}\varphi$. From (1.2) and by Parseval's identity (see e.g. [1, p. 258])

$$\sup_{T>0} \int_R |\hat{\varphi}_T(\omega)|^2 \operatorname{Re} \hat{a}(\omega)\, d\omega < \infty, \tag{2.2}$$

and so, by (1.3),

$$\sup_{T>0} \int_{|\omega| \leqq 1} |\hat{\varphi}_T(\omega)|^2 d\omega < \infty. \tag{2.3}$$

Multiply (1.1) by $y_T'(t)$ and integrate over $[0, T]$ to get

$$(2.4)\qquad \int_0^T |y_T'(\tau)|^2 d\tau = \int_0^T y_T'(\tau) f(\tau)\, d\tau + \int_0^T y_T'(\tau)[y_T'(\tau) - x'(\tau)]\, d\tau$$
$$- \int_0^T y_T'(\tau)(a * \varphi)(\tau)\, d\tau.$$

Thus, recalling (1.6) and (2.1),

$$(2.5)\qquad \|\hat{y}_T'(\omega)\|^2_{L^2(R)} \leqq c\left[1 + \left|\int_R \overline{\hat{y}_T'(\omega)}\hat{a}(\omega)\hat{\varphi}_T(\omega)\, d\omega\right|\right].$$

Obviously $\hat{y}_T'(\omega) = i\omega\hat{y}_T(\omega)$, $\omega \in R$. Also note that as $a' \in L^1(R^+)$ then $\sup_{\omega \in R\setminus\{0\}} |\omega\hat{a}(\omega)| < \infty$. Consequently (2.3) yields

$$(2.6)\qquad \left|\int_{|\omega| \leqq 1} \overline{\hat{y}_T'(\omega)}\hat{a}(\omega)\hat{\varphi}_T(\omega)\, d\omega\right| \leqq \int_{|\omega| \leqq 1} |\omega\hat{a}(\omega)|\,|\hat{y}_T(\omega)|\,|\hat{\varphi}_T(\omega)|\, d\omega$$
$$\leqq c\left(\int_{|\omega| \leqq 1} |\hat{y}_T(\omega)|^2 d\omega\right)^{1/2}.$$

Since $|\hat{a}(\omega)| = O(|\omega|^{-1})$ and a is strongly positive definite there exists $q < \infty$ such that $|\hat{a}(\omega)|^2 \leqq q^2 \operatorname{Re}\hat{a}(\omega)$, $|\omega| \geqq 1$. Therefore by (2.2)

$$(2.7)\qquad \left|\int_{|\omega| \geqq 1} \overline{\hat{y}_T'(\omega)}\hat{a}(\omega)\hat{\varphi}_T(\omega)\, d\omega\right|$$
$$\leqq q\left(\int_{|\omega| \geqq 1} |\hat{y}_T'(\omega)|^2 d\omega\right)^{1/2}\left(\int_{|\omega| \geqq 1} \operatorname{Re}\hat{a}(\omega)|\hat{\varphi}_T(\omega)|^2 d\omega\right)^{1/2}$$
$$\leqq c\left(\int_{|\omega| \geqq 1} |\hat{y}_T'(\omega)|^2 d\omega\right)^{1/2}.$$

The estimates (2.5)–(2.7) imply

$$(2.8)\qquad \int_R |\hat{y}_T'(\omega)|^2 d\omega \leqq c\left[1 + \left(\int_{|\omega| \leqq 1} |\hat{y}_T(\omega)|^2 d\omega\right)^{1/2}\right].$$

Consider the last term on the right side of (2.4). Add and subtract an extra term and integrate parts to get

$$(2.9)\qquad \int_0^T y_T'(\tau)(a * \varphi)(\tau)\, d\tau$$
$$= -a(0)\int_0^T x(\tau)\varphi(\tau)\, d\tau + a(0)\int_0^T [x(\tau) - y_T(\tau)]\varphi(\tau)\, d\tau$$
$$- \int_0^T y_T(\tau)(a' * \varphi)(\tau)\, d\tau.$$

The absolute value of the second term on the right side of (2.9) is obviously uniformly bounded in T. For the third term we have $(d=\int_{R^+}|a'(\tau)|\,d\tau)$

(2.10)

$$\begin{aligned}\left|\int_R \overline{\hat{y}_T(\omega)}\hat{a}'(\omega)\hat{\varphi}_T(\omega)\,d\omega\right| &\leqq d\left(\int_{|\omega|\geqq 1}+\int_{|\omega|\leqq 1}\{|\hat{y}_T(\omega)||\hat{\varphi}_T(\omega)|\}\,d\omega\right)\\ &\leqq d\left(\int_{|\omega|\geqq 1}\omega^2|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2}\left(\int_{|\omega|\geqq 1}\omega^{-2}|\hat{\varphi}_T(\omega)|^2\,d\omega\right)^{1/2}\\ &\quad +c\left(\int_{|\omega|\leqq 1}|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2}\\ &\leqq c\left(\int_{|\omega|\geqq 1}|\hat{y}_T'(\omega)|^2\,d\omega\right)^{1/2}+c\left(\int_{|\omega|\leqq 1}|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2},\end{aligned}$$

where we have used (2.2), (2.3). From (2.4), (2.9), (2.10) follows (note that by positive definiteness $a(0)>0$)

$$\left|\int_0^T x(\tau)\varphi(\tau)\,d\tau\right|\leqq c\left[1+\int_R|\hat{y}_T'(\omega)|^2\,d\omega+\left(\int_{|\omega|\geqq 1}|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2}\right]$$

and hence by (2.8)

$$\left|\int_0^T x(\tau)\varphi(\tau)\,d\tau\right|\leqq c\left[1+\left(\int_{|\omega|\leqq 1}|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2}\right].$$

Together with (1.5) this implies $(x_T=\chi_{[0,T]}x)$

$$\int_R|\hat{x}_T(\omega)|^2\,d\omega\leqq c\left[1+\left(\int_R|\hat{y}_T(\omega)|^2\,d\omega\right)^{1/2}\right].$$

With the aid of (2.1) we conclude that $x\in L^2(R^+)$. From (2.1), (2.8) then follows $x'\in L^2(R^+)$.

REFERENCES

[1] J. A. Nohel and D. F. Shea, *Frequency domain methods for Volterra equations*, Adv. Math., 22 (1976), pp. 278–304.

[2] O. J. Staffans, *Boundedness and asymptotic behavior of solutions of a Volterra equation*, Michigan Math. J., 24 (1977), pp. 77–95.

[3] ———, *A Volterra equation with square integrable solution*, Proc. Amer. Math. Soc., 78 (1980), pp. 213–217.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

SOLUTIONS OF A NONLOCAL CONSERVATION LAW ARISING IN COMBUSTION THEORY*

ROBERT GARDNER†

Abstract. We prove a local existence theorem for smooth solutions with Sobolev space data of a nonlinear conservation law which contains a nonlocal operator. In a certain parameter range we also prove that smooth solutions develop shocks in finite time. This gives further confirmation of the validity of the equation as an asymptotic approximation for the prediction of spontaneous Mach stem formation in solutions of the equations of reactive gas dynamics.

Key words. conservation laws, Mach stem, combustion shock formation

AMS(MOS) subject classifications. 35L, 35R

Introduction. This paper is concerned with solutions of the equation

$$(1)\qquad \begin{aligned} &u_t+Ruu_x+(R-1)\left[\int_0^\infty u(x+\beta s)u_x(x+s)\,ds\right]_x=0,\\ &u(x,0)=u_0(x), \end{aligned}$$

where $x\in\mathbb{R}^1$ (we shall suppress the t-dependence of u in the integral term). Equation (1) arises as an asymptotic approximation which governs the growth of multi-dimensional perturbations in planar detonation front solutions of the equations of reactive gas dynamics in two space variables (see Majda and Rosales [5]). In particular, if $\phi_x=u$, then ϕ describes the evolution of a 2-D perturbation in the primary planar front. In this context the formation of shocks in smooth, rapidly decaying solutions of (1) is associated with the onset of spontaneous Mach stem formation in solutions of the original system. The appearance of the additional shock and contact discontinuity at the Mach node is predicted by the first order asymptotics for the fluid components. This, together with the mechanism leading to the presence of the nonlocal term in (1), is discussed in depth in [5].

Conservation laws with nonlocal terms also arise in other contexts (see e.g. Majda and Rosales [7]), and it has been suggested that (1) may serve as an important canonical model in asymptotic approximations to multidimensional shock wave theory (see Majda [4]). It therefore seems desirable to present a framework for this type of problem in which analytical questions can be investigated. The local existence theory presented here generalizes in a straightforward manner to equations with more complicated nonlocal terms, such as those appearing in the general discussion in [5], and perhaps to systems of the type appearing in [7]. However, we shall limit the discussion to equations of the form (1). Our proof follows lines similar to those of the proof of [4, Thm. 2.1], the main difference being that the nonlocal term forces us to work in Sobolev spaces based on L^1 rather than L^2. In regard to the application of (1) to combustion theory, the existence of solutions with integrable decay in x was postulated in [5] and confirmed numerically in [6]. Such solutions were needed to construct suitable matched asymptotic expansions. Our result provides a rigorous proof of the local existence of such solutions.

* Received by the editors July 22, 1985, and in revised form December 9, 1985. This research was supported by the Deutsche Forschungsgemeinschaft and by the National Science Foundation under grant DMS 8320485.

† Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01003 and Universität Heidelberg, 6900 Heidelberg, Federal Republic of Germany.

We also prove that smooth solutions of (1) develop shocks in finite time. In the application to combustion theory, the relevant parameter range is

$$R>0, \quad R\neq 1, \quad \beta>1;$$

here, R and β depend in a complicated manner on several physical parameters (see [5], [6]). We are not aware of any analytical results on breakdown for this problem; however, there is strong numerical evidence that breakdown does indeed occur (see [6], [7]). Moreover, these numerical results indicate that solutions of (1) exhibit a strikingly rich, parameter dependent variety of phenomena. For example, when $R>1$, solutions with nonnegative data exhibit amplitude growth and oscillation through negative values in the "tail," in marked contrast to solutions of the Burgers equation. Our proof of shock formation is confined to the range $0<R<1$, where the numerics indicate that the mechanism leading to breakdown is somewhat closer to that occurring in the Burgers equation. There is also another requirement for R and β (see (iii) of Theorem 2.1), which we discuss later.

By a *solution* we shall mean a C^1 function which together with its first derivatives is integrable, and which satisfies the equation in the classical sense (actually, we require that u satisfies an equivalent equation, (10b) below, which only involves the first derivatives of u). Since (1) is in conservation form it is also possible to consider weak solutions; however, this is beyond the scope of the present discussion. Natural questions about such solutions, for example, global existence and decay, are probably not routine in character (see Dafermos [1] for a survey of this general area). For example, it does not appear that the equation admits the symmetry of a centered rarefaction wave; since it is not clear how to solve Riemann problems, the Glimm scheme is not immediately available. It may be more appropriate to consider a conservative finite difference approximation of the type used in the numerics in [6]. It would be interesting to investigate the applicability of the method of compensated compactness to the convergence of such approximate solutions (see Tartar [9] for a survey of this area and DiPerna [2] for a recent application to conservation laws). An important ingredient in such an approach is an a priori L^∞ bound for solutions. The numerics suggest that such bounds are available in some parameter ranges but not in others.

1. Local existence of smooth solutions.

A. We will present a local existence theory for smooth solutions of (1) with Sobolev space data. The proof employs an iteration scheme (see e.g. Kato [3], Majda [4]) in which contractiveness is obtained in a low derivative norm while uniform control is maintained over the iterates in a high derivative norm. The new aspect of the present discussion is the presence of the nonlocal operator, which makes it necessary to work in Sobolev spaces based on L^1 rather than L^2. We remark that a proof could also be given based on the Nash-Moser implicit function theorem. However, the above method is simpler and therefore seems preferable.

B. Notation. Let H^s denote the space of functions with s $L^1(\mathbb{R}^1)$ derivatives and with norm

$$\|u\|_s=\sum_{j\leqq s}\int_{-\infty}^{\infty}|\partial_x^j u|\,dx.$$

Also, let $X_{s,T}$ denote the space $L^\infty([0,T]\times H^s)$ with norm

$$\|u\|_{s,T}=\sup_{0\leqq t\leqq T}\|u(\cdot,t)\|_s.$$

If $D \subset \mathbb{R}^1$ and $\Omega \subset \{x \in \mathbb{R}^1, t \geqq 0\}$ define

$$\|u\|_{D,s} = \sum_{j \leqq s} \|\partial_x^j u\|_{L^1(D)}, \qquad \|u\|_{\Omega,s,T} = \sup_{0 \leqq t \leqq T} \|u\|_{\Omega(t),s},$$

where $\Omega(r) = \Omega \cap \{t = r\}$. Finally we will denote the supremum norm by $\|u\|_\infty$; the domain over which the supremum is being taken will be clear from the context.

C. Here are the main results of this section.

THEOREM 1.1. (i) *Suppose that* $u_0 \in H^s$ *with* $s \geqq 2$. *There exists* $T = T(\beta, R, s, \|u_0\|_s, \|u_0^0\|_s) > 0$, *where* u_0^0*is defined in* (8), *below, such that there exists a solution* $u \in X_{s,T}$ *of* (1). *For* $s \geqq 3$, $u \in C^1([0, T_2] \times \mathbb{R}^1)$.

(ii) *Suppose that* $u_0 \in H^s$ *for all* $s \geqq 2$ *and let* T_2 *be the* T *obtained in* (i) *with* $s = 2$. *Then* $u \in C^\infty([0, T_2] \times \mathbb{R}^1)$.

We note that if $u_0 \in H^{s+1}$ then $\|u_0^0\|_s \leqq \|u_0\|_{s+1}$. Thus the precise dependence of T on u_0 can be obtained by sacrificing one derivative.

COROLLARY 1.2. *Let* u *and* v *be solutions of* (1) *such that* $u(x, 0) = v(x, 0)$ *for* $x \geqq \alpha$. *If* $x(t; \alpha)$ *is the solution of* $x_t = u$, $x(0, \alpha) = \alpha$, *define*

$$\Omega = \{(x, t)\colon x \geqq x(t; \alpha),\ u \text{ and } v \text{ are smooth for } t \leqq T\}.$$

Then $u \equiv v$ *in* Ω. *In particular, if* $u \equiv 0$, *then* $v = 0$ *for* $0 \leqq t \leqq T$, $x \geqq x(t; \alpha)$.

Corollary 1.2 implies that signals propagate to the right with finite speed; this plays a crucial role in our proof of breakdown in § 2.

The proofs of Theorem 1.1 and Corollary 1.2 are concluded in §§ G and H, below.

D. L^1 estimates. We first present an L^1 estimate for solutions of local linear equations. Although this is a standard result, we include a proof for completeness.

LEMMA 1.3. (i) *Suppose that* $u_0 \in L^1(\mathbb{R}^1) \cap C^\infty$, $f \in X_{0,T} \cap C^\infty$, *and that* $a(x, t)$, $b(x, t)$ *are smooth bounded functions such that*

$$C = \|a_x\|_\infty + \|b\|_\infty < \infty.$$

If u *is the solution of*

$$u_t + au_x + bu = f, \qquad u(x, 0) = u_0(x), \tag{2}$$

then

$$\|u\|_{0,T} \leqq e^{CT}[\|u_0\|_0 + T\|f\|_{0,T}].$$

(ii) *Let* $x(t; \alpha)$ *be the characteristic curve defined by* $x_t = a(x, t)$, $x(0, \alpha) = \alpha$, *and let*

$$\Omega = \{(x, t)\colon x(t, \alpha) \leqq x \leqq \infty, 0 \leqq t \leqq T\},$$

$$I = \{x\colon x \geqq \alpha\}.$$

Then with notation and hypotheses as in (i), *we have that*

$$\|u\|_{\Omega,0,T} \leqq e^{CT}[\|u_0\|_{I,0} + T\|f\|_{\Omega,0,T}].$$

Proof. (i) First assume that $f = 0$. It suffices to consider data $u_0(x)$ such that the set D where $u_0(x) \neq 0$ consists of a finite number of components, since such functions are dense in L^1. Thus let D be the union of $I_j(0)$, $1 \leqq j \leqq N$, where

$$I_j(0) = (x_j^-(0), x_j^+(0)), \qquad x_j^+(0) \leqq x_{j+1}^-(0).$$

Since $f = 0$, u is of one sign along characteristics of (2). Let $x_j^\pm(t)$ be the solution of

$$\dot{x} = a(x, t), \qquad x(0) = x_j^\pm(0),$$

and let $I_j(t)$ be the interval $(x_i^-(t), x_j^+(t))$. Since $u=0$ on $\partial I_j(t)$ it follows that

$$\partial_t \int_{I_j(t)} u\,dx = \int_{I_j(t)} u_t\,dx.$$

Multiply (2) by the sign of u on I_j, integrate over $I_j(t)$ and sum over j. From the above identity it follows (after integrating au_x by parts) that

$$\partial_t \|u\|_0 \leqq C\|u\|_0, \qquad C = \|a_x\|_\infty + \|b\|_\infty;$$

this proves the lemma when $f=0$.

If $f \neq 0$ but $u_0 = 0$, let $U(x, t; s)$ be the solution of

$$U_t + a(x,t)U_x + b(x,t)U = 0, \qquad U(x,s;s) = f(x,s).$$

Then $u(x,t)$ is a solution of (2), where

$$u(x,t) = \int_0^t U(x,t;s)\,ds,$$

and it follows from the first step that

$$\|u\|_{0,T} \leqq T\|U\|_{0,T} \leqq Te^{CT}\|f\|_{0,T}.$$

If $u_0 \neq 0$, $f \neq 0$, the lemma follows from the above and linearity.

(ii) The proof is virtually the same as (i) and we omit the details.

E. Linear nonlocal equations. In order to define our iteration scheme for (1) it will be necessary to solve equations of the form

$$(3) \qquad u_t + a^2u_x + a^1u + I(u, a^0) = 0, \qquad u(x,0) = u_0(x)$$

where the operator I is defined by

$$(4) \qquad I(u,v) = \int_0^\infty u(x+\beta s)v(x+s)\,ds.$$

LEMMA 1.4. *Suppose that* $s \geqq 2$, $u_0 \in H^s$, $\partial_t^k a^i \in X_{s-k,T}$, $i=1,2$, $k=0,1$. *For every* $T>0$ *there exists a solution* $u \in X_{s,T}$ *of* (3) *which satisfies the estimate*

$$(5) \qquad \|u\|_{s,T} \leqq e^{KT}\|u_0\|_s,$$

where K depends only on s, $\|a^i\|_{s,T}$, $i=1,2$, $(\beta-1)^{-1}\|a^0\|_{s,T}$ *and* R. *If* $s \geqq 3$, *then* $u \in C^1([0,T]\times\mathbb{R}^1)$.

Proof. Define $u^0 \equiv u_0(x)$ and let u^k, $k \geqq 1$ be the solution of

$$u_t^k + a^2u_x^k + a^1u^k = -I(u^{k-1}, a^0), \qquad u^k(x,0) = u_0(x).$$

Let $v^{k+1} = u^{k+1} - u^k$, so that

$$v_t^{k+1} + a^2v_x^{k+1} + a^1v^{k+1} = -I(v^k, a^0), \qquad v^{k+1}(x,0) = 0.$$

Define $w^j = \partial_x^j v^{k+1}$, $\bar{w}^j = \partial_x^j v^k$. For each $r \leqq s$, w^r satisfies

$$(6)_r \qquad w_t^r + a^2w_x^r = -\sum_{j=0}^{r-1}(\partial_x^{j+1}a^2 + \partial_x^j a^1)w^{r-j} - (\partial_x^r a^1)w^0 + \sum_{j=0}^{r}\lambda_j I(\bar{w}^j, \partial_x^{r-j}a^0) \equiv F_r$$

where λ_j depends only on r, and $w^j = 0$ initially. From Lemma 1.3, we have that

$$\|w^r\|_{0,T} \leqq \{Te^{T\|a_x^2\|_\infty}\}\|F_r\|_{0,T}.$$

A simple computation shows that

$$(7) \qquad \|I(u,v)\|_0 \leqq (\beta-1)^{-1}\|u\|_0\|v\|_0.$$

Also, for any $g \in H^1$ it is immediate that $\|g\|_\infty \leqq \|g\|_1$. Using these remarks to estimate F_r in the above we obtain

$$\|w^r\|_{0,T} \leqq \alpha_r T e^{T\|a_x^2\|_\infty}[(\|a^2\|_{r,t} + \|a^1\|_{r,T})\|w^0\|_{r,T} + (\beta-1)^{-1}\|a^0\|_{r,T}\|\bar{w}^0\|_{r,T}],$$

where α_r is a constant depending only on r.

For $s \geqq 2$ we have that $\|a_x^2\|_\infty \leqq \|a^2\|_{s,T}$. Finally, we note that for $r = 0, 1$ we obtain

$$\|w^r\|_{0,T} \leqq \alpha_s T e^{T\|a_x^2\|_\infty}(\|a^2\|_{2,T} + \|a^1\|_{2,T})\|w^0\|_{2,T} + (\beta-1)^{-1}\|a^0\|_{2,T}\|\bar{w}^0\|_{2,T}).$$

Thus we can sum the above for all $r \leqq s$ to obtain

$$\|v^{k+1}\|_{s,T} \leqq \alpha_s T e^{T\|a_x^2\|_\infty}[(\|a^2\|_{s,T} + \|a^1\|_{s,T})\|v^{k+1}\|_{s,T} + (\beta-1)^{-1}\|a^0\|_{s,T}\|v^k\|_{s,T}].$$

It follows that for sufficiently small T, say $T \leqq T^*$, that

$$\|v^{k+1}\|_{s,T^*} \leqq \rho\|v^k\|_{s,T^*}$$

for some $\rho \in (0, 1)$, and so $\{u^k\}$ is Cauchy in X_{s,T^*}. If $u \in X_{s,T^*}$ and $s \geqq 2$ it follows that u and u_x are continuous functions of x, and using the equation, that u is uniformly Lipschitz in t. From the differential equation it also follows that $u_{tx} \in X_{s-1,T^*}$ and so $u_{tt} \in X_{s-1,T^*}$. It then follows that u_t is continuous in x and t also, so that $u \in C^1([0, T^*] \times \mathbb{R}^1)$.

The estimate (5) follows from an estimate similar to the one obtained for v^{k+1}; this clearly allows us to continue the solution globally in T.

F. An iteration scheme for the nonlinear equation. We now use solutions of (3) to define our iteration scheme. It will first be necessary to mollify the data; to this end let $j_\varepsilon(x) = \varepsilon^{-1}j(\varepsilon^{-1}x)$ be a C_0^∞ function with unit mass and let

$$u_0^k(x) = j_{\varepsilon_k} * u_0, \qquad \varepsilon_k = \varepsilon_0 2^{-k}. \tag{8}$$

It is well known that

$$\|J_\varepsilon u - u\|_s \to 0 \qquad \text{as } \varepsilon \to 0, \tag{9a}$$

$$\|J_\varepsilon u - u\|_0 \leqq C\varepsilon\|u\|_1 \quad \text{for } u \in H^1. \tag{9b}$$

Next, note that integration by parts yields two equivalent and useful forms of (1), namely

$$\begin{aligned} &u_t + auu_x + bI(u, u_{xx}) = 0, \\ &a = R - (R-1)\beta^{-1}, \qquad b = (R-1)(1-\beta^{-1}), \end{aligned} \tag{10a}$$

$$u_t + uu_x + BI(u_x, u_x) = 0, \qquad B = (R-1)(1-\beta), \tag{10b}$$

where I is the operator defined in (4).

The approximation scheme is defined by setting $u^0 \equiv u_0^0(x)$, and for $k \geqq 1$, setting u^k to be the solution of

$$0 = u_t^k + au^{k-1}u_x^k + bI(u^k, u_{xx}^{k-1}), \qquad u^k(x, 0) = u_0^k(x), \tag{11a}$$

$$[0 = u_t^k + u^{k-1}u_x^k + BI(u_x^k, u_x^{k-1})]. \tag{11b}$$

From Lemma 1.4 it follows that u^k exists, and since each u_0^k is actually C^∞, that u^k is smooth. Thus u^k also satisfies the equivalent equation (11b).

The main task is to show that $\{u^k\}$ is uniformly bounded in a high derivative norm on some uniform t interval.

LEMMA 1.5. *Suppose that $s \geqq 2$. There exists $T^* > 0$ depending only on s, R, β and $\|u_0\|_{s+1}$ such that for sufficiently small ε_0 in* (8),

$$\|u^k - u_0^0\|_{s,T^*} \leqq 1, \qquad k = 0, 1, 2, \cdots. \tag{12}$$

Proof. The lemma clearly holds when $k=0$; we proceed by induction on k. We shall always view u^k as a solution of (11b).

We again use the notation w^j (resp. $\bar{w}^j$) for $\partial_x^j u^{k+1}$ (resp. $\partial_x^j u^k$). We first establish the following formulas for $j \geqq 1$:

$$\partial_x^{2j} I(w^1, \bar{w}^1) = (2-\beta-\beta^{-1})^j I(w^{j+1}, \bar{w}^{j+1}) + R_j, \tag{13a}$$

$$\partial_x^{2j+1} I(w^1, \bar{w}^1) = (2-\beta-\beta^{-1})^j [I(w^{j+2}, \bar{w}^{j+1}) + I(w^{j+1}, \bar{w}^{j+2})] + \hat{R}_j, \tag{13b}$$

(13c)$_j$ R_j (resp. $\hat{R}_j$) consists of local terms of the form $\lambda w^\alpha \bar{w}^\gamma$, where $1 \leqq \alpha, \gamma \leqq 2j$ (resp. $2j+1$) and $3 \leqq \alpha+\gamma \leqq 2j+1$ (resp. $2j+2$). The coefficients λ depend only on β and j.

We establish (13a) by induction on j; (13b) is obtained by taking ∂_x of (13a). For $j=1$ we have that

$$\begin{aligned}\partial_x^2 I(w^1, \bar{w}^1) &= I(w^3, \bar{w}^1) + 2I(w^2, \bar{w}^2) + I(w^1, \bar{w}^3)\\ &= (2-\beta-\beta^{-1}) I(w^2, \bar{w}^2) - \beta^{-1} w^2 \bar{w}^1 - \beta w^1 \bar{w}^2.\end{aligned}$$

Now assume that (13a) holds for $j \geqq 1$; then

$$\begin{aligned}\partial_x^{2j+2} I(w^1, w^1) &= (2-\beta-\beta^{-1})^j \partial_x^2 I(w^{j+1}, \bar{w}^{j+1}) + \partial_x^2 R_j\\ &= (2-\beta-\beta^{-1})^{j+1} I(w^{j+2}, \bar{w}^{j+2})\\ &\quad -(2-\beta-\beta^{-1})^j [\beta^{-1} w^{j+2} \bar{w}^{j+1} + \beta w^{j+1} \bar{w}^{j+2}] + \partial_x^2 R_j.\end{aligned}$$

If R_j satisfies (13c)$_j$ then $\partial_x^2 R_j$ satisfies (13c)$_{j+1}$; the other local terms satisfy (13c)$_{j+1}$ as well. This establishes (13).

For each j we obtain an equation for w^j by taking ∂_x^j of (11b); using (13) we obtain

$$w_t^j + \bar{w}^0 w_x^j = -\sum_{\alpha,\beta=0}^{j} \lambda_{\alpha,\beta} w^\alpha \bar{w}^\beta + B\tilde{I}_j \equiv F_j, \tag{14a}$$

where $\tilde{I}_j$ is defined by

$$\tilde{I}_j = \begin{cases} I(w^{j/2+1}, \bar{w}^{j/2+1}) & j \text{ even},\\ I(w^{(j+3)/2}, \bar{w}^{(j+1)/2}) + I(w^{(j+1)/2}, \bar{w}^{(j+3)/2}) & j \text{ odd}.\end{cases} \tag{14b}$$

The crucial observation is that for $j \geqq 2$ all terms on the right-hand side of (14a) involve only w^k and $\bar{w}^k$ with $k \leqq j$ (this is false when $j=0, 1$). For example, it follows from (13c) that if $\lambda_{\alpha,\beta} \neq 0$ then at least one of α or β is strictly less than j. It follows from the embedding of H^1 in L^∞ we have that for such α, β,

$$\|w^\alpha \bar{w}^\beta\|_0 \leqq \|w^0\|_j \|\bar{w}^0\|_j. \tag{15a}$$

Similarly, from (7) we have that for all $j \geqq 2$,

$$\|\tilde{I}_j\|_0 \leqq 2(\beta-1)^{-1} \|w^0\|_j \|\bar{w}^0\|_j. \tag{15b}$$

We can now estimate $u^k - u_0^0$. To this end let $U^j = \partial_x^j u_0^0$ and define z^j, $\bar{z}^j$ by

$$w^j = z^j + U^j, \qquad \bar{w}^j = \bar{z}^j + U^j, \qquad 0 \leqq j \leqq s.$$

From (14a) we have for $j \geqq 2$ that z^j satisfies

$$z_t^j + \bar{w}^0 z_x^j = -\bar{w}^0 U^{j+1} + F_j, \tag{16}$$

where from (15), it follows that F_j satisfies the estimate

$$\|F_j\|_0 \leqq C[\|z^0\|_j \|\bar{z}^0\|_j + \|u_0^0\|_j (\|z^0\|_j + \|\bar{z}^0\|_j) + \|u_0^0\|_j^2],$$

where C depends only on j, β and R.

Let $\bar{C} = \|\bar{w}^0\|_{s,T}$ so that for $s \geqq 2$, $\|\bar{w}_x^0\|_\infty \leqq \bar{C}$. For $j \geqq 2$ we obtain from Lemma 1.3 that

$$\|z^j\|_{0,T} \leqq e^{\bar{C}T}[\|u_0^k - u_0^0\|_j + T(\bar{C}\|u_0^0\|_{j+1} + C\|u_0^0\|_j^2) \\ + CT\|z^0\|_{0,T}\|\bar{z}^0\|_{j,T} + \|u_0^0\|_j(\|z^0\|_{j,T} + \|\bar{z}^0\|_{j,T})]. \tag{17$_j$}$$

For $j = 0, 1$ we obtain equations for z^0, z^1, where the right-hand side contains z^j, z^i with $j, i \leqq 2$. This yields an estimate of the form $(17)_j$ with $j = 0, 1$ on the left and $j \leqq 2$ on the right. Finally, sum $(17)_j$ over j, $0 \leqq j \leqq s$ to obtain

$$\|u^k - u_0^0\|_{s,T} \leqq K e^{\bar{C}T}[\|u_0^k - u_0^0\|_s + T(\bar{C}\|u_0^0\|_{s+1} + C\|u_0^0\|_s^2) \\ + CT(\|u^k - u_0^0\|_{s,T}\|u^{k-1} - u_0^0\|_{s,T} \\ + \|u_0^0\|_s(\|u^k - u_0^0\|_{s,T} + \|u^{k-1} - u_0^0\|_{s,T})], \tag{18}$$

where K depends only on s. The induction hypothesis for the $k-1$ step and (9) imply that

$$\bar{C} \leqq 1 + 2\|u_0\|_s \equiv \hat{C} \quad \text{(for small } \varepsilon_0).$$

From (8) and (9) we have for sufficiently small ε_0 that

$$K e^{\hat{C}}\|u_0^k - u_0^0\|_s < \tfrac{1}{3}, \qquad k = 0, 1, 2, \cdots.$$

Now choose $T^* \leqq 1$ so small that

$$T^* K e^{\hat{C}}((\hat{C}+1)\|u_0^0\|_{s+1} + C\|u_0^0\|_s^2) < \tfrac{1}{3}, \qquad T^* K e^{\hat{C}}(C + \|u_0^0\|_s) < \tfrac{1}{2}.$$

For such T^* the induction step will hold at the kth step; this completes the proof.

We remark that if $u_0 \in H^{s+1}$, then $\|u_0^0\|_{s+1} \leqq \|u_0\|_{s+1}$, so that T^* can be chosen uniformly for all small ε_0. If $u_0 \notin H^{s+1}$ the choice of T^* must tend to zero with decreasing ε_0.

Armed with Lemma 1.5, we can now prove contactiveness in the low norm.

LEMMA 1.6. *Suppose that* $\alpha \in (0, 1)$. *There exists* $T^{**} \in (0, T^*]$ *such that*

$$\|u^{k+1} - u^k\|_{0,T^{**}} \leqq \alpha\|u^k - u^{k-1}\|_{0,T^{**}} + \|u_0^k - u_0^0\|_0. \tag{19}$$

T^{**} *depends only on* β, R *and* T^*.

Proof. Let $v^k = u^{k+1} - u^k$, $k = 0, 1, \cdots$. After an integration by parts, it can be seen that v^k satisfies

$$0 = v_t^k + u^k v_x^k - B u_x^k v^k + (1-B)u_x^k v^{k-1} - B[\beta I(u_{xx}^k, v^k) + \beta^{-1} I(v^{k-1}, u_{xx}^k)],$$
$$v^k(x, 0) = u_0^{k+1}(x) - u_0^k(x).$$

By Lemma 1.3, (7), and the embedding of H^1 into L^∞, we obtain

$$\|v^k\|_{0,T} \leqq e^{CT}[CT\|v^{k-1}\|_{0,T} + CT\|v^k\|_{0,T} + \|v^k(\cdot, 0)\|_0],$$

where C depends only on $\|u^k\|_{2,T}$, β and R. This determines a suitable $T^{**} \leqq T^*$ such that (19) holds.

G. Completion of the proof. (i) Lemmas 1.5 and 1.6 imply that for $T \leqq T^{**}$, $\{u^k\}$ converges to a limit $u \in x_{0,T}$. These lemmas, together with a standard interpolation inequality for Sobolev space imply that the convergence occurs in the space $X_{s-1,T}$. The equation implies that u_t^k converges to u_t in $X_{s-2,T}$, so that for $s \geqq 3$ the Sobolev embedding theorem implies that u is $C^1([0, T] \times \mathbb{R}^1)$.

(ii) If $u_0 \in H^s$ for all $s \in \mathbb{Z}_+$, we claim that there exists T independent of s such that $u \in C^\infty([0, T] \times \mathbb{R}^1)$. Pick $T_2 > 0$ such that the existence holds when $s = 2$. If $w = \partial_x^2 u$, then w satisfies

$$(20)\quad w_t + uw_x + [3 - (\beta + \beta^{-1})(R-1)(1-\beta)]u_x w = (2 - \beta - \beta^{-1})(R-1)(1-\beta)I(w, w).$$

From Lemma 1.3 and (7) we see that

$$\partial_t \|w\|_0 \leqq \text{const } \|w\|_0^2;$$

this inequality and $\|u_0''\|_0$ determine the maximal T_2.

For $j > 2$ we let $w^j = \partial_x^j w$. From (13) with $\bar{w}^j = w^j$ it follows that the equation for w^j analogous to (14a) is *linear* in w^j. Thus, given a bound for $\|w^k\|_{0,T}$ for $k < j$ and $T < T_2$ we obtain a bound for $\|w^j\|_{0,T}$ since by Lemma 1 and the previous remark $\|w^j\|_{0,T}$ can grow at most exponentially fast for $T \leqq T_2$. This proves (ii).

H. Proof of Corollary 1.2. Let $w = u - v$; using form (11b) of (1) we have that

$$(21)\qquad 0 = w_t + uw_x + v_x w + BI(w_x, u_x) + BI(v_x, w_x),$$

and after an integration by parts this becomes

$$(22)\qquad 0 = w_t + uw_x + (v_x - \beta^{-1}Bu_x - Bv_x)w - \beta^{-1}BI(w, u_{xx}) - BI(v_{xx}, w).$$

We now use (ii) of Lemma 1.3 with Ω as in the statement of the corollary and (7) to obtain

$$\|w\|_{\Omega,0,T} \leqq CTe^{CT}\|w\|_{\Omega,0,T},$$

where C depends only on $\|u\|_{\Omega,2,T}$, $\|v\|_{\Omega,2,T}$, β and R. Thus for sufficiently small T, $\|w\|_{\Omega,0,T} = 0$. The argument can be repeated to obtain the result in all of Ω.

2. Formation of shocks.

A. We will show that for certain parameter values and rather general data the solution of (1) breaks down in finite time, i.e., there exists $T_b > 0$ such that $u_x(x, t)$ becomes infinite at some $t = \hat{t} \leqq T_b$. The result is not sharp in that $\hat{t}$ may be substantially smaller than T_b, although for solutions with a single positive pulse (such as those studied numerically in [6]) T_b may provide a more accurate estimate for $\hat{t}$.

The analysis of breakdown is facilitated by having a suitable continuation theorem for smooth solutions. For example, for local conservation laws it can be shown that a bound for $\|\nabla u\|_\infty$ implies a bound on the higher derivatives (see [4, Thm. 2.2]). The difficulty in this regard with equation (1) can be seen in equation (20) for $w = \partial_x^2 u$. Even if a uniform bound is postulated for $\|u_x\|_\infty$ the quadratic term $I(w, w)$ could conceivably cause blow-up in the second derivative.

The key to continuation and breakdown is in controlling this term. To this end we consider C_0^∞ data u_0 supported in $\{x \leqq 0\}$ which, together with u_0' and u_0'', satisfy certain sign conditions near $x = 0$. Using maximum principle type arguments it can be shown that these sign conditions persist into a certain region $\Omega \subset \{t > 0, x < 0\}$ as long as the solution remains C^∞. It can then be shown that if $|u_x|$ is uniformly bounded for $t \leqq T_b$, then Ω intersects the line $t = T_b$. This is our continuation principle. Breakdown is obtained from the explicit knowledge of the signs of the derivatives of u in Ω.

In summary, breakdown could occur in two different ways. Either a shock forms from the data in Ω in time of order T_b or a shock forms somewhere to the left of Ω and rapidly moves into Ω.

THEOREM 2.1. *Suppose that $u_0 \in C_0^\infty(x \leqq 0)$ and that u_0, β and R satisfy the conditions*

(i) *There exists $y < 0$ such that $u_0''(x) > 0$ for $y < x < 0$.*

(ii) $0 < R < 1$, $\beta > 1$.

(iii) $\gamma > 0$, *where* $\gamma = 1 - (1-R)(\beta - \beta^{-1})/2$.

Then u_x becomes unbounded at a time $\hat{t} \leqq T_b$, where $T_b = (\gamma |u_0'(y)|)^{-1}$.

We have been unable to determine whether (iii) is a genuine threshold for breakdown, or whether it is an artifact of our techniques. In the two numerical examples reported in [6] with $R < 1$, condition (iii) is satisfied. The second example has $\gamma = 10^{-2}$. However, the authors mention other experiments in the parameter ranges $0.5 \leqq R \leqq 1$, $1 \leqq \beta \leqq 10$ in which they still found evidence of breakdown. It may be that for $\gamma < 0$ breakdown can be inhibited if the data are sufficiently small. Further numerical experiments in this parameter range might provide a better indication of what sort of results to expect.

In the following we will assume that u_x is uniformly bounded for $t \leqq T_b$. A contradiction is obtained in § D, below.

B. An a priori estimate. Given a solution u of (1) we define "characteristic" curves $x(t; \alpha)$ by

$$\frac{dx}{dt} = u(x, t), \qquad x(0, \alpha) = \alpha \quad \text{for } \alpha \in \mathbb{R}^1. \tag{23}$$

The region Ω referred to earlier is defined by setting

$$\begin{aligned} T_0 &= \sup\{T \geqq 0\colon u \text{ is } C^\infty \text{ on } 0 \leqq t \leqq T, x(t; y) \leqq x \leqq 0\}, \\ \Omega &= \{(x, t)\colon 0 < t < T_0, x(t; y) < x < 0\}. \end{aligned} \tag{24}$$

Note that by (ii) of Theorem 1.1, $T_0 > 0$.

THEOREM 2.2. (i) *For all $\hat{t} < T_0$, the set $\Omega \cap \{t = \hat{t}\}$ is an interval of positive length.*

(ii) *The inequalities $u > 0$, $u_x < 0$, and $u_{xx} > 0$ hold for all $(x, t) \in \Omega$.*

Proof. (i) Since u is assumed to be smooth in $\bar{\Omega} \cap \{t \leqq \hat{t}\}$ solution curves of (23) cannot cross in the (x, t) plane, so that the map $\alpha \to x(t; \alpha)$ is a diffeomorphism of $[y, 0]$ onto $[x(t; y), 0]$.

(ii) We will show that $u_{xx} > 0$ in Ω; the other inequalities follow immediately from this.

By hypothesis (i) of Theorem 2.1 and Corollary 1.2, u must decay rapidly near $x = 0$. It is therefore difficult to control the sign of u_{xx} in this region. This problem is overcome by perturbing the equation slightly to

$$u_t^\varepsilon + u^\varepsilon u_x^\varepsilon + (R-1)(1-\beta)I(u_x^\varepsilon, u_x^\varepsilon) = \varepsilon f(x), \qquad u^\varepsilon(x, 0) = u_0(x),$$

where f is specified below. If f decays rapidly at $|x| = \infty$ and is reasonably smooth, the existence theory of § 1 proceeds exactly as before. In particular, the proofs of Lemmas 1.5 and 1.6 are exactly the same. Thus for such f, u^ε will depend continuously on ε in a high derivative norm, and it will exist on some interval $[0, T_1]$ where T_1 is uniform for all small ε. We omit the details.

We now choose $f(x)$ so that $f(x) = 0$ for $x \geqq 0$, f and f' are continuous, and so that $f''(x) = \rho(x)|x|^k$ for $x < 0$, where k is a large positive integer and $\rho(x) \geqq 0$ is C^∞ with $\rho = 0$ for $x < y$ and $\rho = 1$ for $x \geqq y/2$. Clearly $f \in H^{k+2}$ but $f \notin H^{k+1}$.

We next define "characteristic" curves $x^\varepsilon(t; y)$ and a region Ω_ε for the solution u^ε as in (23) and (24). We will show that $u_{xx}^\varepsilon > 0$ in $\Omega_\varepsilon \cap [0, T_1]$ for all $\varepsilon > 0$.

To this end we make the following claim: there exists a (relatively open) border $B_\varepsilon \subset \bar{\Omega}_\varepsilon \cap [0, T_1]$ containing the line segments in $\partial\Omega_\varepsilon$,

$$S = \{t = 0, y \leqq x \leqq 0\} \cup \{x = 0, 0 \leqq t \leqq T_1\},$$

such that $u_{xx}^\varepsilon = w^\varepsilon > 0$ in $B_\varepsilon \cap \Omega_\varepsilon$. We assume the claim for the moment and proceed with the proof.

If the lemma were false there would exist a smallest $t = \hat{t} \leqq T_1$ such that $w^\varepsilon(\hat{x}, \hat{t}) = 0$ for some $\hat{x} \in [x^\varepsilon(\hat{t}; y), 0]$; since $w^\varepsilon > 0$ in B_ε it follows that $\hat{x} < 0$. The equation for w^ε is (see (20))

$$\begin{aligned} &w_t^\varepsilon + u^\varepsilon w_x^\varepsilon + \Gamma u_x^\varepsilon w^\varepsilon = (2 - \beta - \beta^{-1})(R - 1)(\beta - 1)I(w^\varepsilon, w^\varepsilon) + \varepsilon f''(x), \\ &w(x, 0) = u_0''(x), \qquad \Gamma = 3 - (\beta + \beta^{-1})(R - 1)(1 - \beta). \end{aligned} \tag{25}$$

For $\beta > 1$ and $0 < R < 1$ the coefficient of I in the above is positive. Also, by minimality of $\hat{t}$, $w = w_x = 0$ at $(\hat{x}, \hat{t})$. Since $\varepsilon f''$, $w^\varepsilon \geqq 0$ and $w^\varepsilon > 0$ near $x = 0$ it follows that the right-hand side of (25) is positive, whence $w_t(\hat{x}, \hat{t}) > 0$. This contradicts the minimality of $\hat{t}$. Thus $w^\varepsilon > 0$ in $\Omega_\varepsilon \cap [0, T_1]$ for all $\varepsilon > 0$. Since w^ε converges to $w = u_{xx}$ as ε tends to zero in $[0, T_1] \times \mathbb{R}^n$, we have that $w \geqq 0$ in Ω. It follows that $u \geqq 0$ and $u_x \leqq 0$ in Ω. Thus the free boundary of Ω, $x = x(t; y)$, is an increasing function of t. If, for some $(\hat{x}, \hat{t}) \in \Omega$, $w(\hat{x}, \hat{t}) = 0$, then the above argument applied to w at $(\hat{x}, \hat{t})$ shows that w must vanish identically on the segment $x \geqq \hat{x}$, $t = \hat{t}$. Thus u must be a linear function of x for $x \geqq \hat{x}$, $y = \hat{t}$, and since u lies in a Sobolev space, it follows that $u \equiv 0$ on this half line. Now the equation is reversible in time as long as u remains smooth, so that we may apply Corollary 1.2 in backward time to conclude that $u(x, 0) = u_0(x) = 0$ for $x \geqq \hat{x}$. Since our assumption was that $\hat{x} < 0$, we have obtained a contradiction to (ii) of Theorem 2.1. Thus $w > 0$ in $\Omega \cap [0, T_1]$.

Fix $T < T_0$, where T_0 is as in (24); we next extend the result to the region $\Omega \cap [0, T]$. By hypothesis, u is C^∞ in $\bar{\Omega} \cap [0, T]$ so that $\|u(\cdot, t)\|_{\Omega(r),s}$ is uniformly bounded for $r \in [0, T]$ and each s, where $\Omega(r)$ is as in 1.B. We next note that given a bound for $\|u(\cdot, r)\|_{\Omega(r),s}$ we can alter $u(x, r)$ for $x < x(r; y)$ to a (C^∞) function $\tilde{u}(\cdot, r)$ such that $\|\tilde{u}(\cdot, r)\|_s \leqq C\|u(\cdot, r)\|_{\Omega(r),s}$ for some constant C depending only on s. Now fix $s > 2$; for $r \in [0, T]$ we can determine a T_1 as in the beginning of the proof using $\tilde{u}(\cdot, r)$ as "data," and by construction, T_1 will depend only on T. If v is the resulting solution we have that $u \equiv v$ in Ω by Corollary 1.2. Thus the mollification does not affect u in Ω. (Note that such a modification may be necessary since a singularity could form rapidly in the region to the left of Ω.) Proceeding by induction, we assume $u(\cdot, nT_1)$ satisfies (ii) of Theorem 2.1, with y replaced with $x(nT_1, y)$, modify u to $\tilde{u}$ in the manner described above at $t = nT_1$, and apply the argument of the preceding paragraph to conclude that $u_{xx} > 0$ in $\Omega \cap \{t \leqq (n+1)T_1\}$. Since T_1 is independent of n provided that $nT_1 \leqq T$, we obtain the result in $\Omega \cap [0, T]$ for any $T < T_0$.

To complete the proof it only remains to construct the border B_ε. We first claim that there exists $x(\varepsilon) < 0$ such that for $x(\varepsilon) \leqq x \leqq 0$ and $t \leqq T_1$ we have that

$$|(2 - \beta - \beta^{-1})(R - 1)(\beta - 1)I(w^\varepsilon, w^\varepsilon)| < \varepsilon f(x). \tag{26}$$

Since $f \in H^{k+2}$ and $u_0 \in C_0^\infty$ it follows that $u \in X_{k+2,T}$, whence w is C^{k-1} in near $x = 0$ uniformly in t. Since $w(x, t) \equiv 0$ for $x \geqq 0$ it follows that $\partial_x^j w = 0$ for $j \leqq k - 1$, at $x = 0$. Thus by Taylor's theorem $|w(x, t)| \leqq C|x|^{k-1}$ for x near zero, uniformly in t. Thus for x near zero we have that

$$|\text{const } I(w^\varepsilon, w^\varepsilon)| \leqq c|x|^{2k-1}$$

which clearly implies (26) for x close enough to zero (depending on ε) and $k \geqq 2$.

Now let $x^\varepsilon(t;\alpha)$ be the characteristics for u^ε and define

$$p^\varepsilon(t;\alpha)=u^\varepsilon_{xx}(x^\varepsilon(t;\alpha),t),$$
$$q^\varepsilon(t;\alpha)=u^\varepsilon_x(x^\varepsilon(t;\alpha),t).$$

The equation for p^ε is

$$\begin{aligned}&p^\varepsilon_t+\Gamma q^\varepsilon p^\varepsilon=(2-\beta-\beta^{-1})(R-1)(\beta-1)I(w^\varepsilon,w^\varepsilon)+\varepsilon f''(x),\\&p(0;\alpha)=u_0''(\alpha).\end{aligned}\tag{27}$$

From (26) it follows that the terms on the right-hand side are positive provided that α is close enough to zero. Moreover by (ii) of Theorem 2.1, $p^\varepsilon(\alpha,0)>0$ for $\alpha\in(y,0)$. At the smallest $t=\hat t$ where $p^\varepsilon=0$ it would follow from (27) that $p_t(\hat t,\alpha)>0$ for α near zero, contradicting the minimality of $\hat t$. Thus $p^\varepsilon(\hat t;\alpha)>0$ on some region of the form $[0,T]\times[\alpha,0)$ for $|\alpha|$ small enough. Finally, we note that from (i) of this theorem, the intersection of B_ε with each line $t=\text{constant}$ is an interval containing zero in its closure, for $t\leqq T_1$. Thus B_ε has nonempty interior in x for each fixed t.

C. A continuation principle. Given a sign condition on u_{xx} and u_x, it is relatively easy to prove a continuation theorem.

LEMMA 2.3. *Suppose that $|u_x|$ is uniformly bounded for $t\leqq T_b$. Then $T_0\geqq T_b$ where T_0 is as in* (24).

Proof. We need to show that a bound for $|u_x|$ implies that u is C^∞ in Ω. Clearly $\|u\|_{\Omega,1,T_b}$ is finite. Since $u_{xx}>0$ in Ω we also have that

$$\|u_{xx}\|_{\Omega,0,T_b}=\sup_{0\leqq t\leqq T_b}\int_{x(t;y)}^0 u_{xx}(x,t)\,dx=\sup_{0\leqq t\leqq T_b}-u_x(x(t;y),t)<\infty.$$

Thus $\|u\|_{\Omega,2,T_b}<\infty$. Proceeding by induction, suppose that $\|u\|_{\Omega,j,T_b}$ is bounded for $j\geqq 2$. We then use (13) to obtain (14a) with $\partial_x^{j+1}u=w^{j+1}$ and $\bar w^k=w^k$. From (14b) and $(13\text{c})_l$ where $l=[(j+1)/2]$, this equation is linear in w^{j+1} if $j\geqq 2$. Thus we can apply (ii) of Lemma 1.3 to obtain a bound for $\|u\|_{\Omega,j+1,T_b}$. This completes the proof.

D. A proof of breakdown. We finally derive a contradiction to the hypothesis that u_x is uniformly bounded for $t\leqq T_b$.

Let $q(t)=u_x(x(t;y),t)$; then the equation for q is

$$\begin{aligned}q_t&=-q^2+(R-1)(\beta-1)[I(u_{xx},u_x)+I(u_x,u_{xx})]\\&=[-1-(R-1)(\beta-1)]q^2-(R-1)(\beta-1)^2I(u_{xx},u_x).\end{aligned}\tag{28}$$

Since $u_x<0$ and $u_{xx}>0$ in Ω, u_x is monotone, and we have that

$$\begin{aligned}I(u_{xx},u_x)&=\int_0^\infty u_{xx}(x(t;y)+\beta s)u_x(x(t;y)+s)\,ds\\&<\int_0^\infty u_{xx}(x(t;y)+\beta s)u_x(x(t;y)+\beta s)\,ds\\&=\beta^{-1}\int_0^\infty\frac{\partial}{\partial x}[u_x(x(t;y)+\beta s)^2/2]\,ds\\&=-\beta^{-1}q^2/2.\end{aligned}$$

Thus from the above and (28) we obtain

$$q_t<[-1-(R-1)(\beta-1)+(R-1)(\beta-1)^2\beta^{-1}/2]q^2=-\gamma q^2$$

where γ is as in (ii) of Theorem 2.1. Thus $q(t)$ is less than $(-T_b+t)^{-1}$, which completes the proof.

REFERENCES

[1] C. DAFERMOS, *Hyperbolic systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations (Oxford, 1982), pp. 25-70; NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci., Reidel, Boston, 1983.

[2] R. DIPERNA, *Convergence of the viscosity method for isentropic gas dynamics*, Comm. Pure Appl. Math., 91 (1983).

[3] T. KATO, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Arch. Rational Mech. Anal., 58 (1975), pp. 181-205.

[4] A. MAJDA, *Compressible fluid flow and systems of conservation laws in several space variables*, Lecture Notes in Applied Mathematics 53, Springer-Verlag, New York, 1984.

[5] A. MAJDA AND R. ROSALES, *A theory for spontaneous Mach-stem formation in reacting shock fronts*. I, *The basic perturbation analysis*, SIAM J. Appl. Math., 43 (1983), pp. 1310-1334.

[6] ———, *A theory for spontaneous Mach-stem formation in reacting shock fronts*. II, *Steady wave bifurcations and the evidence for breakdown*, Stud. Appl. Math., 71 (1984), pp. 117-148.

[7] ———, *Resonantly interacting weakly nonlinear hyperbolic waves*. I. *A single space variable*, Stud. Appl. Math., 71 (1984), pp. 149-179.

[8] A. MAJDA, R. ROSALES AND E. THOMANN, *Numerical computation of solutions of an integro-differential conservation law arising in MACH-stem formation*, to appear.

[9] L. TARTAR, *Compacité par compensation: résultats et perspectives*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. IV (Paris, 1981/1982), pp. 350-369, Research Notes in Mathematics 84, Pitman, London.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

014

APPLICATION OF TOPOLOGICAL TECHNIQUES TO THE ANALYSIS OF ASYMPTOTIC BEHAVIOR OF NUMERICAL SOLUTIONS OF A REACTION-DIFFUSION EQUATION*

SAT NAM S. KHALSA†

Abstract. The initial boundary value problem for a reaction-diffusion equation

$$(*) \qquad u_t = u_{xx} + f(u), \qquad f(u) = -u(u-b)(u-1), \qquad 0 < b < \tfrac{1}{2},$$

was analysed in [2], [10] by using the Conley index. In this paper we study the asymptotic behavior of solutions of the semidiscrete approximations

$$(**) \qquad \dot{u}_i = (u_{i-1} - 2u_i + u_{i+1})/h^2 + f(u_i), \qquad i = 1, \cdots, n.$$

We show that for large n the spectrum of the linearized discrete steady-state problem is a "good" approximation for the spectrum of the linearized continuous steady-state problem. Using the interpretation of the Conley index as the dimension of an unstable manifold of a steady-state solution, we establish that the properties of solutions of (**) are completely analogous to those of the solutions of (*). The asymptotic, as $t \to \infty$, second order convergence of the approximate solutions is proved.

Key words. reaction-diffusion, finite differences, Conley index

AMS(MOS) subject classifications. 65M20, 65M10, 35K57

1. Introduction. Many phenomena in biology and physiology can be modeled by certain nonlinear reaction-diffusion equations. The asymptotic state of a solution specifies its ultimate behavior while ignoring transient effects. The stable asymptotic states may be represented by solutions which can be perturbed by a uniformly small function without destroying their long-time behavior. These are the solutions generally seen in applied contexts. It is therefore crucial to know that the asymptotic behavior of the approximate solutions mimic those of the exact solution.

The asymptotic behavior of solutions for the problem

$$(1.1a) \qquad u_t = u_{xx} + f(u), \qquad |x| < L, \quad t > 0,$$

$$(1.1b) \qquad u(x, 0) = u^0(x), \qquad |x| < L,$$

$$(1.1c) \qquad u(\pm L, t) = 0, \qquad t > 0,$$

where

$$(1.2) \qquad f(u) = -u(u-b)(u-1), \qquad 0 < b < \tfrac{1}{2},$$

was analysed in [2] and [10] (see Theorem 2.2 below). In particular, it was shown there that for $L > L_0$ the steady-state problem

$$(1.3a) \qquad u'' + f(u) = 0, \qquad |x| < L,$$

$$(1.3b) \qquad u(\pm L) = 0$$

has exactly three solutions: u_0, u_1 and u_2 with $u_0 \equiv 0 < u_1(x) < u_2(x)$, which are nondegenerate (i.e., zero is not in the spectrum of the linearized problem). u_0 and u_2 are attractors for the associated parabolic problem (1.1a-c). By this we mean that if the initial data u^0 is sufficiently close (in C) to either u_0 or u_2, then the corresponding

* Received by the editors May 28, 1985; accepted for publication (in revised form) September 5, 1985.

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

solution of (1.1a–c) converges (in C) to the corresponding solution of (1.3a–b). Similarly, u_1 is unstable and has a one-dimensional unstable manifold. These results are based on computation of the Conley index (the Morse index of an isolated invariant set) for each of u_0, u_1 and u_2. The Conley index generalizes the classical Morse index of a nondegenerate critical point of a vector field in that the classical index is a nonnegative integer n, where n is the dimension of the unstable manifold to the critical point, and, considered as an isolated invariant set, the homotopy index of the critical point is the homotopy type of a pointed n-sphere.

In this paper we analyse the asymptotic behavior of solutions of semidiscrete approximations

$$\text{(1.4a)} \qquad \dot{u}_i=(u_{i-1}-2u_i+u_{i+1})/h^2+f(u_i), \qquad i=1,\cdots,n, \quad t>0,$$

$$\text{(1.4b)} \qquad u_i(0)=u_{i,0}, \qquad i=1,\cdots,n,$$

$$\text{(1.4c)} \qquad u_0(t)=u_{n+1}(t)=0, \qquad t>0,$$

of (1.1a–c). Here $h=2L/(n+1)$, and $u_i(t)$ is an approximation to $u(x_i,t)$, with an appropriate choice of $u_{i,0}$, say, $u_{i,0}=u^0(x_i)$, $x_i=ih-L$, $i=0,\cdots,n+1$. We first study the approximate steady-state problem

$$\text{(1.5a)} \qquad (u_{i-1}-2u_i+u_{i+1})/h^2+f(u_i)=0, \qquad i=1,\cdots,n,$$

$$\text{(1.5b)} \qquad u_0=u_{n+1}=0.$$

In Theorem 3.1 we prove that for large n (1.5a, b) also has exactly three solutions $\bar{0}=\bar{u}^0<\bar{u}^1<\bar{u}^2$ and establish the second order convergence in the sup-norm of the approximate solutions.

In Theorem 3.2 we show that the properties of solutions of the problem (1.4a–c) are completely analogous to those of (1.1a–c) and establish the asymptotic, as $t\to\infty$, second order convergence of the approximate solutions.

Our approach is based on an (intuitively natural) fact that for large n the spectrum of the linearized operator in the right-hand side of (1.4a) is a "good" approximation of the spectrum of the linearized operator in the right-hand side of (1.1a). This gives existence and convergence of the approximate steady-state solutions, and also implies that the Conley index of a rest point $\bar{u}^k$ of the approximate problem is the same as the one of the corresponding rest point $u_k(x)$, $k=0,1,2$, of the exact problem. The latter implies the existence of orbits connecting the rest points of the approximate problem.

The results of this paper have been extended [7] to the finite element approximation of (1), with interpolation of coefficients for nonlinear terms.

In § 2 we collect the results from [8] and [9] which we need to analyse the approximate steady-state problem and the results from [10] about the continuous problem (1.1a–c).

In § 3 we prove our principal result, Theorem 3.2, which analyses the properties of the approximate problem (1.4a–c) and establishes the asymptotic convergence of the approximate solutions.

Recently, asymptotic convergence of numerical solutions of systems of reaction-diffusion equations to constant and zero rest points was analysed by several authors. In [4], [5] finite difference approximations were shown to converge with a time-independent error bound by imposing a monotonicity condition on the reaction term or under the conditions that the reaction term is "slowly varying." The conditions imposed guarantee the exponential decay of the exact solution together with its derivatives. The results in [4], [5] were obtained for both the Dirichlet and Neumann problems. For the Neumann problem similar results were obtained in [3] and in the linear case, using finite elements, in [6].

2. Background and preliminaries. As we mentioned in the Introduction, in our analysis of the approximate steady-state problem we follow [8]. The approach in [8] is to replace the problem of solving the steady-state problem using finite differences by an equivalent problem: solving the integral equation

$$u(x)+\int_{-L}^{L} G(x,y)f(u(y))\,dy=0 \tag{2.1}$$

using the method of mechanical quadratures

$$u_i+\sum_{j=1}^{n} G(x_i,x_j)f(u_j)=0, \qquad i=1,\cdots,n, \tag{2.2}$$

and then to analyse the latter problem. Here $G(x,y)$ is the Green's function of the operator d^2/dx^2 for the boundary condition (1.3b).

Equation (2.1) can be regarded as an operator equation

$$u+Tu=0, \tag{2.3}$$

where

$$Tu=\int_{-L}^{L} G(x,y)f(u(y))\,dy, \tag{2.4}$$

in the Banach space E of bounded measurable functions $u(x)$ on $[-L, L]$.

From [8, p. 307, proof of Thm. 19.5 and Thm. 19.6] we have

THEOREM 2.1. (i) *The boundary value problem* (1.3a, b) *in* $C^2([-L, L])$ *is equivalent to the integral equation* (2.1) *in* E.

(ii) *Let* $u_*(x)$ *be an isolated solution of the problem* (1.3a, b). *Let also* $f(u)$ *be continuously differentiable in the domain*

$$|u-u_*(x)|\leqq \delta, \qquad \delta=\text{const}>0, \tag{2.5}$$

and set

$$g(x)=f'(u_*(x)). \tag{2.6}$$

Assume that the linearized problem

$$u''+g(x)u=0 \tag{2.7}$$

with the boundary conditions (1.3b) *has no nontrivial solutions.*

Then there exist n_0 *and* δ_0 *such that for* $n\geqq n_0$ *the system* (1.5a, b) *has a unique solution satisfying the inequalities*

$$|u_i-u_*(x_i)|\leqq \delta_0, \qquad i=1,\cdots,n. \tag{2.8}$$

If $f(u_*(x))$ *is twice continuously differentiable, then the rate of convergence is given by*

$$\max_{1\leqq i\leqq n} |u_i-u_*(x_i)|\leqq ch^2, \qquad c=\text{const}. \tag{2.9}$$

Remark. Note that by Theorem 2.2 below all solutions $u_*(x)$ of (1.3a, b) are isolated, and (2.7) has no nontrivial solutions.

Theorem 2.1 together with the argument in [8, p. 302] imply

LEMMA 2.1. *Under the assumptions of Theorem* 2.1, *for any given* δ_0 *there exists* n_0 *such that for* $n \geqq n_0$ *any solution* $\bar{u}^* = (u_0^*, \cdots, u_{n+1}^*)^T$ *of* (1.5a, b) *satisfies*

$$\max_{0 \leqq i \leqq n+1} |u_i^* - u_*(x_i)| \leqq \delta_0, \tag{2.10}$$

for some solution $u_*(x)$ *of* (1.3a, b).

The following lemma is contained in [9, Thm. 6.2] and can also be obtained using [8, Thm. 18.1].

LEMMA 2.2. *Under the assumptions of Theorem* 2.1 *every limit point of any sequence* λ_n *of eigenvalues of equations*

$$(v_{i-1} - 2v_i + v_{i+1})/h^2 + g(x_i)v_i = \lambda v_i, \qquad i = 1, \cdots, n, \tag{2.11}$$

$$v_0 = v_{n+1} = 0, \tag{2.12}$$

is an eigenvalue of the equation

$$v'' + g(x)v = \lambda v, \tag{2.13}$$

$$v(\pm L) = 0. \tag{2.14}$$

We shall use the notation $\bar{u} \equiv \bar{u}_n = (u_0, \cdots, u_{n+1})^T$ for $\bar{u} \in R^{n+2}$ and the notation $\bar{u} < \bar{v}$ when $u_i < v_i$, $i = 1, \cdots, n$; $u_0 \leqq v_0$, $u_{n+1} \leqq v_{n+1}$. With some abuse of notation we shall also write $\bar{u}$ for the vector $(u_1, \cdots, u_n)^T$, in the case that $u_0 = u_{n+1} = 0$.

We shall also need a comparison principle by Kamke in the following form:

LEMMA 2.3 [1]. *Let* $\bar{u}(t)$ *and* $\bar{v}(t)$ *be solutions of* (1.4a, c), *defined for* $a \leqq t \leqq b$. *Then* $\bar{u}(a) < \bar{v}(a)$ *implies* $\bar{u}(b) \leqq \bar{v}(b)$.

A solution $u_*(x)$ of a steady-state problem (1.3a, b) is called an attractor for the associated parabolic problem (1.1a, c) if, for the initial data $u(x, 0)$ sufficiently close (in C) to u_*, the corresponding solution of (1.1a–c) converges (in C) to u_*. We shall use the notation $h(I)$ for the Conley index of an isolated invariant set I, and Σ^k for the pointed k-sphere.

THEOREM 2.2 [10, Thm. 24.13]. *Let* f *be defined by* (1.2), *and let* $L > L_0$. *Then there are exactly three steady-state solutions* $u_k \in C^\infty$, $k = 0, 1, 2$, *of* (1.1a, c): $0 \equiv u_0(x) < u_1(x) < u_2(x) \leqq 1$, $|x| < L$. *They are isolated invariant sets,* $h(u_0) = h(u_2) = \Sigma^0$, *in particular,* u_0 *and* u_2 *are attractors for* (1.1a, c), *and the linearized operators* Q_0 *and* Q_2, *where*

$$Q_k = \frac{d^2}{dx^2} + g_k, \qquad g_k(x) = f'(u_k(x)), \qquad k = 0, 1, 2, \tag{2.15}$$

together with the boundary conditions (1.3b), *have only negative eigenvalues.* $h(u_1) = \Sigma^1$, *in particular* Q_1 *has precisely one positive eigenvalue, and* u_1 *has a one-dimensional unstable manifold which consists of orbits connecting* u_1 *to each of the other rest points. Initial data* $u(x, 0)$ *which satisfies* $u_1(x) < u(x, 0) < u_2(x)$ (*resp.* $0 < u(x, 0) < u_1(x)$) *on* $|x| < L$ *is in the stable manifold of* u_2 (*resp.* 0).

3. Convergence.

THEOREM 3.1. *Let* $u_k(x)$, $k = 0, 1, 2$, *be the solutions of* (1.3a, b). *There exists* n_0 *such that for* $n \geqq n_0$ *the system* (1.5a, b) *has exactly three solutions* $\bar{u}_n^k \equiv \bar{u}^k = (u_0^k, \cdots, u_{n+1}^k)^T$, $k = 0, 1, 2$, *satisfying*

$$\max_{0 \leqq i \leqq n+1} |u_i^k - u_k(x_i)| \leqq ch^2, \qquad c = \text{const}, \quad k = 1, 2, \tag{3.1}$$

$$\bar{0} \equiv \bar{u}^0 < \bar{u}^1 < \bar{u}^2, \tag{3.2}$$

where $\bar{0} = (0, \cdots, 0)^T$.

Proof. By Theorem 2.2, for $u_*(x) = u_k(x)$, $k = 0, 1, 2$, the conditions of Theorem 2.1 are satisfied. This gives the existence of solutions $\bar{u}^k$, $k = 0, 1, 2$, of (1.5a, b), satisfying (3.1). By (2.9) and Lemma 2.1 $\bar{u}^k$, $k = 0, 1, 2$, are the only solutions of (1.5a, b).

By Theorem 2.2 $0 \equiv u_0 < u_1 < u_2$ are isolated solutions of (1.3a, b). Together with (3.1) this implies (3.2) for n_0 sufficiently large. □

Define

$$F(u) = \int_0^u f(t)\,dt, \tag{3.3}$$

$$\Phi(\bar{u}) = \sum_{k=1}^{n} [u_k(u_{k-1} - 2u_k + u_{k+1})/2h^2 + F(u_k)]. \tag{3.4}$$

Then we have

LEMMA 3.1. *The system* (1.4a-c) *is a gradient one with respect to the function* Φ, *i.e.*

$$\dot{u} = \nabla\Phi. \tag{3.5}$$

Let $u_k(x)$ solve (1.3a, b) and $\bar{u}^k$ solve (1.5a, b). For $k = 0, 1, 2$, define the linearized matrix operators $Q_n^k = (q_{ij})$ and $\tilde{Q}_n^k = (\tilde{q}_{ij})$, $i, j = 1, \cdots, n$, by

$$q_{ij} = \begin{cases} -2/h^2 + f'(u_i^k), & i = j, \\ 1/h^2, & |i - j| = 1, \\ 0, & |i - j| > 1, \end{cases} \tag{3.6}$$

$$\tilde{q}_{ij} = \begin{cases} q_{ij} + \gamma_i, & i = j, \\ q_{ij}, & i \neq j, \end{cases} \tag{3.7}$$

where

$$\gamma_i = f'(u_k(x_i)) - f'(u_i^k). \tag{3.8}$$

LEMMA 3.2. *Suppose the assumptions of Theorem* 3.1 *hold. Let* n_0 *be chosen as in Theorem* 3.1. *There exists* $n_0' \geqq n_0$ *such that for* $n \geqq n_0'$ *we have*:

(i) Q_n^0 *and* Q_n^2 *have only negative eigenvalues, and* Q_n^1 *has precisely one positive eigenvalue, and* $\bar{u}_1$ *has a one-dimensional unstable manifold.*

(ii) $\bar{u}^k$, $k = 0, 1, 2$, *are isolated invariant sets of the system* (1.4a, c) *and* $h(\bar{u}^0) = h(\bar{u}^2) = \Sigma^0$, $h(\bar{u}^1) = \Sigma^1$.

Proof. By Lemma 2.2 with $u_*(x) = u_k(x)$, $k = 0, 1, 2$, every limit point of any sequence λ_n of eigenvalues of (2.11)-(2.12) is an eigenvalue of the problem (2.13)-(2.14). Using the notation (2.15) and (3.6)-(3.8), we rewrite (2.11)-(2.12) and (2.13)-(2.14), respectively, as

$$\tilde{Q}_n^k \bar{v} = \lambda \bar{v}, \tag{3.9}$$

$$Q_k^v = \lambda v, \qquad v(\pm L) = 0. \tag{3.10}$$

Let us prove the lemma for $k = 0$. For $k = 1, 2$, the proof is completely analogous.

Let $\tilde{\lambda}_n$ be a sequence of eigenvalues of (3.9). Since by Theorem 2.2 the largest eigenvalue λ of Q_0 is negative, say, $\lambda = -3\varepsilon$, $\varepsilon > 0$, by the above there exists n_0'' such that for $n \geqq n_0''$, $\tilde{\lambda}_n \leqq -2\varepsilon$. To complete the proof of (i) it is sufficient to verify that the eigenvalues λ_n of Q_n^0 satisfy $\lambda_n \leqq -\varepsilon$ for sufficiently large n. For $n \geqq n_0$ from (3.8) and (3.1), using that f'' is bounded on $[-L, L]$, we have

$$|\gamma_i| = \left| \int_{u_i^0}^{u^0(x_i)} f''(t)\,dt \right| \leqq c|u^0(x_i) - u_i^0| \leqq \frac{c}{n^2}. \tag{3.11}$$

From (3.7) and (3.11)

$$\begin{aligned}\lambda_n &= \sup_{\|\bar{v}\|=1} (Q_n^0 \bar{v}, \bar{v}) = \sup_{\|\bar{v}\|=1} \left[(\tilde{Q}_n^0 \bar{v}, \bar{v}) + \sum_{i=1}^{n} \gamma_i v_i^2 \right] \\ &\leqq \sup_{\|\bar{v}\|=1} \left[(\tilde{Q}_n^0 \bar{v}, \bar{v}) + \frac{c}{n^2} \right] = \tilde{\lambda}_n + \frac{c}{n^2}, \qquad c = \text{const},\end{aligned} \tag{3.12}$$

where

$$\|\bar{v}\|^2 = \sum_{i=1}^{n} v_i^2, \qquad (\bar{u}, \bar{v}) = \sum_{i=1}^{n} u_i v_i.$$

Now the above implies that choosing $n_0' \geqq \max\{n_0, n_0'', \sqrt{c/\varepsilon}\}$, we have for $n \geqq n_0'$, $c/n^2 < \varepsilon$, and therefore $\lambda_n < -\varepsilon$.

By (i) $\bar{u}^k$, $k = 0, 1, 2$, are nondegenerate rest points of (1.4a, c). Since from Lemma 3.1 the system (1.4a, c) is a gradient one, then by [10, pp. 151-152] the Morse index of $\bar{u}^k$, which is the number of positive eigenvalues of Q_n^k, is defined and is equal to the Conley index $h(\bar{u}^k)$. Thus by (i) $h(\bar{u}^0) = h(\bar{u}^2) = \Sigma^0$ and $h(\bar{u}^1) = \Sigma^1$. Alternatively, the latter result follows from [10, § 4, pp. 503-504]. By [10, p. 468 and Thm. 23.32] $\bar{u}^k$, $k = 0, 1, 2$, are isolated invariant sets for (1.4a, c). □

LEMMA 3.3. *The rectangle*

$$R = \bigcap_{i=1}^{n} \{\bar{u}: 0 \leqq u_i \leqq 1\} \tag{3.13}$$

is attracting for the problem (1.4a-c), *i.e., all solutions* $\bar{u}(t)$ *of* (1.4a-c) *tend to* R *as* $t \to \infty$.

Proof. Let $R_\tau = \tau R$, $\tau > 1$, be a family of contracting rectangles about R. We say that $\bar{u}(t)$ is in the jth right-hand face on R_τ if $u_j(t) = \tau$, with a similar definition for the jth left-hand face. We shall also use the notations $\bar{f} = \{f(u_1), \cdots, f(u_n)\}^T$ and $\nabla\Phi$ (see (3.5)) for the vector field which is the right-hand side of the system (1.4a).

If now $\bar{u}(T) \in \partial R_\tau$, e.g., $\bar{u}(T)$ is in the jth right-hand face, then from the definition (1.2) of $f(u)$ we have

$$(\bar{f}, \bar{n}(\bar{u})) = f(u_j) < -\eta,$$

where $\bar{n}$ is the outward-pointing normal at $\bar{u}$, $\eta = \text{const} > 0$. And therefore by (1.4a) and Lemma 3.1 there holds

$$(\nabla\Phi, \bar{n}) = (u_{j+1} - 2u_j + u_{i-1})/h^2 + f(u_j) < -\eta. \tag{3.14}$$

Thus (3.14) shows that $\bar{u}(t)$ must lie in a smaller rectangle for $T < t < T + \delta$, for some $\delta > 0$. □

Using Lemmas 3.2 and 3.3 and repeating the proof of [10, Lemma 24.12] with $u_k(x)$ replaced by $\bar{u}^k$ we arrive at the next lemma.

LEMMA 3.4. *Under the assumptions of Lemma* 3.2, *there exist solutions* $\bar{v}^0$ *and* $\bar{v}^2$ *of* (1.4a, c), *which connect* $\bar{u}^1$ *to* $\bar{u}^0$ *and* $\bar{u}^1$ *to* $\bar{u}^2$, *respectively*; *i.e.*,

$$\lim_{t \to -\infty} \bar{v}^0(t) = \bar{u}^1, \qquad \lim_{t \to \infty} \bar{v}^0(t) = u^0,$$

$$\lim_{t \to -\infty} \bar{v}^2(t) = \bar{u}^1, \qquad \lim_{t \to \infty} \bar{v}^2(t) = \bar{u}^2.$$

Combining the above results, we have the following:

THEOREM 3.2. *Let f be defined by* (1.2) *and $L > L_0$. Then there exists n_0 such that for $n \geqq n_0$*

(i) *The steady-state problem* (1.5a, b) *has exactly three solutions*: $\bar{0} \equiv \bar{u}^0 < \bar{u}^1 < \bar{u}^2 \leqq \bar{1}$.

(ii) $\bar{u}^k$, $k = 0, 1, 2$, *are isolated invariant sets for the associated parabolic problem* (1.4a, c). $h(\bar{u}^0) = h(\bar{u}^2) = \Sigma^0$, *in particular, $\bar{u}^0$ and $\bar{u}^2$ are the attractors, and the linearized operators Q_n^0 and Q_n^2, where Q_n^k are defined by* (3.6), *have only negative eigenvalues.* $h(\bar{u}^1) = \Sigma^1$, *in particular, $\bar{u}^1$ has a one-dimensional unstable manifold which consists of orbits connecting $\bar{u}^1$ to each of the other rest points.*

(iii) *Initial data which satisfies $\bar{u}^1 < \bar{u}(0)$ is in the stable manifold of $\bar{u}^2$, and there holds*:

$$\lim_{t \to \infty} \max_{0 \leqq i \leqq n+1} |u_i(t) - u(x_i, t)| \leqq ch^2. \tag{3.15}$$

Initial data which satisfies $\bar{u}(0) < \bar{u}^1$ is in the stable manifold of $\bar{0}$, and there holds:

$$\lim_{t \to \infty} \max_{0 \leqq i \leqq n+1} |u_i(t)| = 0. \tag{3.16}$$

Proof. Condition (i) follows from Theorem 3.1 and Lemma 3.3. Condition (ii) follows from Lemmas 3.2 and 3.4. Condition (iii) follows from (i), (ii) and Lemmas 2.3, 3.4 by repeating the argument in [10, pp. 535-536]. □

Acknowledgments. The author wishes to thank Prof. S. Grossberg for bringing to his attention [10] and Prof. J. Smoller for bringing to his attention [5].

REFERENCES

[1] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1965.

[2] C. CONLEY AND J. SMOLLER, *Remarks on the stability of steady state solutions of reaction-diffusion equations*, in Bifurcation Phenomena in Mathematical Physics and Related Phenomena, C. Bardos and D. Bessis, eds., Reidel, Dordrecht, 1980, pp. 47-56.

[3] L. GALEONE, *The use of positive matrices for the analysis of the large time behavior of the numerical solution of reaction-diffusion systems*, Math. Comp., 41 (1983), pp. 461-472.

[4] D. HOFF, *Approximation and decay of solutions of systems of nonlinear diffusion equations*, Rocky Mountain J. Math., 7 (1977), pp. 547-556.

[5] ———, *Stability and convergence of finite difference methods for systems of nonlinear reaction-diffusion equations*, SIAM J. Numer. Anal., 15 (1978), pp. 1161-1177.

[6] K. ISHIHARA, *On numerical asymptotic behavior of finite element solutions for parabolic equations*, Numer. Math., 44 (1984), pp. 285-300.

[7] S. N. S. KHALSA, *Finite element approximation of a reaction-diffusion equation. Part* I: *Application of topological techniques to the analysis of the asymptotic behavior of the semi-discrete approximations*, Quart. Appl. Math., to appear.

[8] M. A. KRASNOSELSKII, G. M. VAINIKKO, P. P. ZABREIKO, J. B. RUTITCHI AND V. JA STECENKO, *Approximate Solution of Operator Equations*, Walters-Noordhoof, Gronigen, 1972.

[9] H.-O. KREISS, *Difference approximations for boundary and eigenvalue problems for ordinary differential equations*, Math. Comp., 26 (1972), pp. 605-624.

[10] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

015

ON THE ZEROS OF THE ASKEY-WILSON POLYNOMIALS, WITH APPLICATIONS TO CODING THEORY*

LAURA CHIHARA†

Abstract. In a symmetric association scheme that is (P and Q)-polynomial, the P and Q eigenmatrices are given by balanced ${}_4\phi_3$ Askey-Wilson polynomials. In this paper, the parameters of the Askey-Wilson polynomial are classified so that its zeros are not contained in its spectrum. These results, together with theorems of Biggs and Delsarte, imply the nonexistence of perfect codes and tight designs in the classical association schemes of type A_N, B_N, C_N, D_N and the affine matrix schemes.

Key words. Askey-Wilson polynomials, association schemes, perfect codes, tight designs

AMS(MOS) subject classifications. 5, 33, 94

1. Introduction. In classical coding theory there is a well-known relationship between the existence of perfect codes and the properties of certain orthogonal polynomials. Lloyd's theorem [7, p. 179] states that if a perfect code exists in the Hamming metric, then the Krawtchouk polynomial must have integral zeros. In this paper we consider a very general set of orthogonal polynomials, the Askey-Wilson polynomials [1], and give sufficient conditions on their parameters so that they do not have the corresponding property from Lloyd's theorem. Our main result is Theorem 4.8, which shows that perfect codes do not exist in the families of association schemes [10] defined by Chevalley groups over $GF(q)$. The only possible exceptions are perfect 1-codes for B_N and C_N when $N=2^m-1$. Since B_N and C_N provide q-analogues of N-tuples of 0's and 1's, this is the natural condition for perfect Hamming 1-codes [7, p. 23].

The fact that the Askey-Wilson polynomials are relevant is due to work of Biggs [3], Delsarte [5] and Leonard [6]. Biggs showed that Lloyd's theorem generalized to distance regular graphs, while Delsarte generalized it to metric P-polynomial association schemes. Both theorems stated that related polynomials must have zeros in a very restricted set. Leonard proved that the polynomials for a (P and Q)-polynomial association scheme must be the Askey-Wilson polynomials. The set for the zeros of these polynomials is easily identified. Since the Askey-Wilson polynomials have an explicit formula, it is possible to show that the zeros do not lie in the set. Since the parameters of the association scheme are related to the parameters of the polynomial, this proves nonexistence of perfect codes. A dual theorem of Delsarte [5, p. 76] also proves the nonexistence of tight designs.

Our paper is organized in the following way. In § 2, we give the Askey-Wilson polynomials and identify the generalized Lloyd polynomials in Proposition 2.1 as other Askey-Wilson polynomials. Properties of the zeros are given in § 3. Under the hypothesis that q is integral and a, b, d rational, sufficient conditions on a, b and d are listed in Table 1 so that the zeros do not lie in the spectrum. In § 4, we show how to use the results from § 3 to prove nonexistence of perfect codes and tight designs in the classical association schemes. These results are summarized in Table 2. Finally, in § 5 we show that if a perfect 1-code exists in types B_N and C_N, then $N=2^m-1$. We also give another proof of Proposition 2.1.

* Received by the Editors November 12, 1984; accepted for publication November 20, 1985.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. Present address, Department of Mathematics, St. Olaf College, Northfield, Minnesota 55057.

2. Preliminaries. In this section, we review the basic facts about (P and Q)-polynomial association schemes and the Askey-Wilson polynomials. References for this material are [1], [2], [5].

A symmetric association scheme X is called P-polynomial, where P is the eigenmatrix of X with (i, k) entry $P_k(i)$, if for a given set of distinct nonnegative real numbers $z_0=0, z_1, \cdots, z_n$ and each integer k, $0 \leqq k \leqq N$, there exists a polynomial $\Phi_k(z)$ over $\mathbb{R}$ of degree k such that

$$\Phi_k(z_i)=P_k(i), \qquad i=0, 1, \cdots, N.$$

A Q-polynomial scheme is defined analogously from the Q eigenmatrix. We denote the corresponding polynomials by Φ_k^*.

If X is a (P and Q)-polynomial association scheme, Leonard showed [6] that for $N \geqq 9$, Φ_k and Φ_k^* are given by Askey-Wilson polynomials, including certain limiting cases. These polynomials $P_k(\theta(x), a, b, c, d; q)$ of degree k in $\theta(x)$ are defined in terms of basic hypergeometric series and have the explicit formula [1]

$$P_k(\theta(x), a, b, c, d; q)=P_k(\theta(x))={}_4\phi_3\left(\begin{matrix} q^{-k}, abq^{k+1}, q^{-x}, cdq^{x+1} \\ aq, bdq, cq \end{matrix} \;\middle|\; q; q\right) \tag{2.1}$$

where

$$\theta(x)=(1-q^{-x})(1-cdq^{x+1}) \tag{2.2}$$

and

$${}_{r+1}\phi_r\left(\begin{matrix} q^{-k}, a_1, \cdots, a_r \\ b_1, \cdots, b_r \end{matrix} \;\middle|\; q; q\right)=\sum_{j=0}^{k} \frac{(q^{-k})_j(a_1)_j \cdots (a_r)_j q^j}{(b_1)_j \cdots (b_r)_j(q)_j} \tag{2.3}$$

with

$$(a)_j=(a; q)_j=\begin{cases} (1-a)(1-aq) \cdots (1-aq^{j-1}), & j=1, 2, \cdots, \\ 1, & j=0. \end{cases} \tag{2.4}$$

Askey and Wilson showed that if aq, bdq or cq is assumed to be q^{-N}, then $\{P_k(\theta(x))\}_{k=0}^N$ are orthogonal polynomials on $\{\theta(0), \cdots, \theta(N)\}$ and have the discrete orthogonality relation

$$\sum_{x=0}^{N} P_n(\theta(x), a, b, c, d; q)P_m(\theta(x), a, b, c, d; q)w(x)=0, \tag{2.5}$$

$$m \neq n, \quad 0 \leqq m, \quad n \leqq N$$

where

$$w(x)=w(x, a, b, c, d; q)=\frac{(cdq)_x(1-cdq^{2x+1})(aq)_x(bdq)_x(cq)_x(abq)^{-x}}{(q)_x(1-cdq)(cda^{-1}q)_x(b^{-1}cq)_x(dq)_x}. \tag{2.6}$$

Leonard's theorem [6] states that for $N \geqq 9$, given a (P and Q)-polynomial association scheme X, there exists parameters a, b, c, d, q such that

$$\Phi_k(\theta(x), a, b, c, d; q)=\Phi_k(\theta(x))=\nu_k P_k(\theta(x), a, b, c, d; q) \tag{2.7}$$

and

$$\Phi_k^*(\theta^*(x), a, b, c, d; q)=\Phi_k^*(\theta^*(x))=\mu_k P_k(\theta^*(x), c, d, a, b; q) \tag{2.8}$$

where $\theta^*(x)=(1-q^{-x})(1-abq^{x+1})$. The nonzero constants ν_k and μ_k are the valencies and multiplicities, respectively, of the association scheme. Moreover, up to a constant, $\nu_k=w(k, c, d, a, b; q)$ and $\mu_k=w(k, a, b, c, d; q)$.

For such an association scheme X, recall that the generalized Lloyd and Wilson polynomials [5, p. 58] are given by

$$\Psi_e(\theta(x))=\sum_{k=0}^{e}\Phi_k(\theta(x)), \tag{2.9}$$

$$\Psi_e^*(\theta^*(x))=\sum_{k=0}^{e}\Phi_k^*(\theta^*(x)) \tag{2.10}$$

where e is a positive integer.

We now state a proposition that identifies $\Psi_e(\theta(x))$ and $\Psi_e^*(\theta^*(x))$ also as Askey-Wilson polynomials.

PROPOSITION 2.1. *Suppose X is a (P and Q)-polynomial association scheme such that* (2.7) *and* (2.8) *hold. Then*

$$\Psi_e(\theta(x))=AP_e(\tilde{\theta}(x-1), aq, b, cq, dq; q), \tag{2.11}$$

$$\Psi_e^*(\theta^*(x))=A^*P_e(\tilde{\theta}^*(x-1), cq, d, aq, bq; q), \tag{2.12}$$

where A and A^* are nonzero constants, $\tilde{\theta}(x-1)=(1-q^{-x+1})(1-cdq^{x+2})$ and $\tilde{\theta}^*(x-1)=(1-q^{-x+1})(1-abq^{x+2})$.

Remark. Recall that for $P_k(\theta(x), a, b, c, d; q)$, we have defined $\theta(x)$ as in (2.2). Thus, for polynomials with shifted parameters, $P_k(\theta(x), aq, b, cq, dq; q)$, we get $\theta(x)=(1-q^{-x})(1-cdq^{x+3})$. From now on, $\theta(x)$ will always stand for the variable in $P_k(\theta(x), a, b, c, d, q)$, and whenever the parameters shift, we will denote the corresponding variable as $\tilde{\theta}(x)$.

Proof of Proposition 2.1. Delsarte shows [5, p. 58] that $\Psi_0(\theta(x)),\cdots,\Psi_{N-1}(\theta(x))$ form an orthogonal set of polynomials on $\{\theta(1),\cdots,\theta(N)\}$ with respect to the weight $\theta(x)w(x)$. From (2.2) and (2.3), we see that

$$\theta(x)w(x, a, b, c, d; q)=\alpha w(x-1, aq, b, cq, dq; q) \tag{2.13}$$

where $\alpha\neq 0$ is a constant. Since $\Psi_e(\theta(x))$ is a polynomial in $\theta(x)$ and orthogonal polynomials are unique with respect to a weight, we have (2.11). Similarly, we get (2.12).

Remark. In § 5, we will present another proof of Proposition 2.1 in which we explicitly determine the constant A.

3. Properties of the roots of the Askey-Wilson polynomials. There is a generalized Lloyd's theorem for perfect e-codes due to Delsarte [5, p. 63] and Biggs [3, p. 294]. It states that if the association scheme X has a perfect e-code, then the Lloyd polynomial of degree e, $\Psi_e(\theta(x))$, must have e distinct roots among $\{\theta(1),\cdots,\theta(N)\}$. A similar result holds for tight t-designs ($t=2e$) and Wilson polynomials [5, p. 76]. By Proposition 2.1, $\Psi_e(\theta(x))$ and $\Psi_e^*(\theta^*(x))$ are explicitly given by an Askey-Wilson polynomial. In this section, we state properties of $P_e(\tilde{\theta}(x-1), aq, b, cq, dq; q)$.

PROPOSITION 3.1. *Let $\theta_1,\cdots,\theta_e$ be the e roots of $P_e(\tilde{\theta}(x-1), aq, b, cq, dq; q)$ as a polynomial in $\theta(x)$. Then*

(i) $\theta_1=\theta(1)+\dfrac{(1-aq^2)(1-bdq^2)(1-cq^2)q^{-1}}{(1-abq^3)}$ *if* $e=1$,

(ii) $\displaystyle\sum_{j=1}^{e}\theta_j=\sum_{j=1}^{e}\theta(j)+\frac{(1-aq^{e+1})(1-bdq^{e+1})(1-cq^{e+1})(1-q^{e-1})(1-q^{e})q^{-e}}{(1-abq^{2e+1})(1-q)}$

if $e\geqq 2$.

Proof. Note that $\tilde{\theta}(x-1)=q\theta(x)+1-q-cdq^2+cdq^3$. Let $F_e(\theta(x))=P_e(\tilde{\theta}(x-1), aq, b, cq, dq; q)$. Then clearly $F_e(\theta(x))=\beta(\theta(x)-\theta_1)\cdots(\theta(x)-\theta_e)$ for some constant $\beta\neq 0$. Thus, to get a formula for the sum of the roots, we need to consider the coefficient of $\theta^{e-1}(x)$ in $F_e(\theta(x))/\beta$. From (2.1) and (2.3), we have

$$F_e(\theta(x))=\sum_{j=0}^{e}\frac{(q^{-e})_j(abq^{e+2})_j(q^{-x+1})_j(cdq^{x+2})_jq^j}{(aq^2)_j(bdq^2)_j(cq^2)_j(q)_j}. \tag{3.1}$$

An explicit computation yields $(1-q^{-x+k})(1-cdq^{x+k+1})=q^k(\theta(x)-\theta(k))$. Hence, for $j\geqq 1$,

$$(q^{-x+1})_j(cdq^{x+2})_j=\prod_{k=1}^{j}(1-q^{-x+k})(1-cdq^{x+k+1})=\prod_{k=1}^{j}q^k(\theta(x)-\theta(k)).$$

Thus,

$$\begin{aligned}F_e(\theta(x))&=1+\sum_{j=1}^{e}\frac{(q^{-e})_j(abq^{e+2})_j[\prod_{k=1}^{j}\theta(x)-\theta(k)]q^{j+j(j+1)/2}}{(aq^2)_j(bdq^2)_j(cq^2)_j(q)_j}\\&=\beta\left\{\frac{1}{\beta}+\sum_{j=1}^{e}\frac{(q^{-e})_j(abq^{e+2})_j[\prod_{k=1}^{j}\theta(x)-\theta(k)]q^{j+j(j+1)/2}}{\beta(aq^2)_j(bdq^2)_j(cq^2)_j(q)_j}\right\},\end{aligned} \tag{3.2}$$

where

$$\beta=\frac{(q^{-e})_e(abq^{e+2})_eq^{e+e(e+1)/2}}{(aq^2)_e(bdq^2)_e(cq^2)_e(q)_e}=\frac{(-1)^eq^e(abq^{e+2})_e}{(aq^2)_e(bdq^2)_e(cq^2)_e}.$$

Thus, the expression between the brackets in (3.2) is now monic in $\theta(x)$. The coefficient of $\theta(x)^{e-1}(e\geqq 2)$ in $F_e(\theta(x))/\beta$ is

$$-\sum_{i=1}^{e}\theta_i=-\sum_{j=1}^{e}\theta(j)-\frac{(1-aq^{e+1})(1-bdq^{e+1})(1-cq^{e+1})(1-q^{e-1})(1-q^e)q^{-e}}{(1-abq^{e+1})(1-q)},$$

where the sum on the right side of the equality is the contribution when $j=e$ in (3.2), and the second is the $j=e-1$ contribution. For $e=1$, $F_1(\theta(x))$ is of degree 1, so solving explicitly for θ_1 yields (i).

PROPOSITION 3.2. *If $\theta_1,\cdots,\theta_e$ are the e roots of $P_e(\tilde{\theta}(x-1), aq, b, cq, dq; q)$ as a polynomial in $\theta(x)$, then*

$$\prod_{i=1}^{e}\theta_i=\frac{(aq^2)_e(bdq^2)_e(cq^2)_e}{(abq^{e+2})_eq^e}\,{}_4\phi_3\left(\begin{matrix}q^{-e}, abq^{e+2}, cdq^2, q\\ aq^2, bdq^2, cq^2\end{matrix}\;\middle|\; q;q\right). \tag{3.3}$$

Proof. Using the same notation as in the proof of Proposition 3.1 we see that we want the constant term in $F_e(\theta(x))/\beta$. From (3.2) we have

$$(-1)^e\prod_{i=1}^{e}\theta_i=\frac{1}{\beta}+\sum_{j=1}^{e}\frac{(q^{-e})_j(abq^{e+2})_j(-1)^j\theta(1)\cdots\theta(j)q^{j+j(j+1)/2}}{\beta(aq^2)_j(bdq^2)(cq^2)_j(q)_j}.$$

Noting that $\theta(k)=-(1-q^k)(1-cdq^{k+1})q^{-k}$ and recalling (2.3) and the definition of β, the result follows.

From now on, we assume $cq=q^{-N}$.

PROPOSITION 3.3. *Suppose $P_e(\tilde{\theta}(x-1), aq, b, cq, dq; q)$ has e roots θ_i such that $\theta_i=\theta(x_i)$ for some x_i, $i=1,2,\cdots,e$. Then*

$$\begin{aligned}-(1-abq^{2e+1})\sum_{i=1}^{e}(dq^{x_i}+q^{N-x_i})&=-(1-abq^{2e+1})(dq+q^{N-e})(1+p'(q))\\&\quad+(1-aq^{e+1})(1-bdq^{e+1})(-1+p(q))\end{aligned} \tag{3.4}$$

where $p'(q)$ and $p(q)$ are polynomials in q with integer coefficients and $p'(q) \equiv 0 \bmod q$ and $p(q) \equiv 0 \bmod q$.

Proof. Recalling $\theta(x) = (1-q^{-x})(1-dq^{-N+x})$, we have

$$\sum_{i=1}^{e} \theta(x_i) = \sum_{i=1}^{e} (1 - dq^{-N+x_i} - q^{-x_i} + dq^{-N})$$
$$= e(1+dq^{-N}) - \sum_{i=1}^{e} (dq^{-N+x_i} + q^{-x_i})$$

and

$$\sum_{j=1}^{e} \theta(j) = e(1+dq^{-N}) - \sum_{j=1}^{e} (dq^{-N+j} + q^{-j})$$
$$= e(1+dq^{-N}) - (dq^{-N} + q^{-e})\frac{(1-q^e)}{(1-q)}$$

where the last equality is obtained by the sum for the geometric progression. Substituting the above expressions into (i) and (ii) of Proposition 3.1 and multiplying the resulting identity by $q^N(1-abq^{2e+1})$ yields the desired result, where

$$1+p'(q) = \begin{cases} 1, & e=1, \\ \dfrac{1-q^e}{1-q}, & e \geqq 2, \end{cases}$$

and

$$-1+p(q) = \begin{cases} -1+q^{N-1}, & e=1, \\ (q^{N-e}-1)(1-q^{e-1})\dfrac{(1-q^e)}{(1-q)}, & e \geqq 2. \end{cases}$$

Remarks. The existence of a perfect e-code in X implies that the e distinct roots of $\Psi_e(\theta(x))$ are contained in $\{\theta(1), \cdots, \theta(N)\}$. In fact, we can eliminate $\theta(1)$ and $\theta(N)$ as possible roots of $\Psi_e(\theta(x))$. From (2.11) and (3.2), we see that $\Psi_e(\theta(1)) = 1 \neq 0$. On the other hand, if we let $x = N$, then

$$\Psi_e(\theta(N)) = {}_3\phi_2\left(\begin{matrix} q^{-e}, abq^{e+2}, dq \\ aq^2, bdq^2 \end{matrix} \,\middle|\, q; q\right) = \frac{(1/bq^e)_e(aq/d)_e}{(aq^2)_e(1/bdq^{e+1})_e}$$

where the last equality was obtained by the ${}_3\phi_2$ evaluation [8, p. 96]

$$(3.5) \qquad {}_3\phi_2\left(\begin{matrix} q^{-e}, A, B \\ C, ABq^{1-e}/C \end{matrix} \,\middle|\, q; q\right) = \frac{(C/A)_e(C/B)_e}{(C)_e(C/AB)_e}.$$

Hence, $\Psi_e(\theta(N)) = 0$ implies that either

$$(3.6) \qquad b = q^{-k}, \quad 1 \leqq k \leqq e \quad \text{or} \quad a/d = q^{-k}, \quad 1 \leqq k \leqq e.$$

However, since the Askey-Wilson polynomials are orthogonal on $\{\theta(1), \cdots, \theta(N)\}$, certain positivity conditions must be satisfied, putting some restrictions on a, b and d [1, p. 1015]. These considerations rule out (3.6). Thus $\Psi_e(\theta(N)) \neq 0$.

We assume that the (P and Q)-polynomial association scheme X has a perfect e-code ($2e+1 \leqq N$). Thus, the generalized Lloyd's theorem and the previous remarks tell us that Ψ_e has all its roots in $\{\theta(2), \cdots, \theta(N-1)\}$. For integral $q \neq 1$ this will imply conditions on the parameters a, b and d. For the known association schemes we will record those values a, b and d and find they do not satisfy these conditions.

The main idea is to assume q integral and a, b and d rational. Then (3.4), upon clearing denominators, is a polynomial identity in q. We get an expression that is $0 \bmod q$ on one side, and not $0 \bmod q$ on the other. This contradiction establishes that $\Psi_e(\theta(x))$ cannot have e distinct roots among $\{\theta(i) | i = 2, \cdots, N-1\}$ and hence X cannot have a perfect e-code. These results are given in Table 1. For simplicity's sake, we only list those parameters for which a contradiction can be obtained without detailed knowledge of the zeros themselves. We give two examples to illustrate the techniques.

We assume q is integral, $q \neq 1$, and e is an integer, $2e+1 \leqq N$. Also, recall $cq = q^{-N}$.

Example 1. Suppose $d = r/s$, $a = m/nq^k$, $b = m'/n'q^l$, where r, s, m, n, k, m', n', l are integers, $r, s, m, n, m', n' \not\equiv 0 \bmod q$, and $k \leqq e$, $l \leqq e$. From (3.4), we get

$$-\left(1-\frac{mm'}{nn'}q^{2e+l-l-k}\right)\sum_{i=1}^{e}\left(\frac{r}{s}q^{x_i}+q^{N-x_i}\right)$$
$$=-\left(1-\frac{mm'}{nn'}q^{2e+1-l-k}\right)\left(\frac{r}{s}q+q^{N-e}\right)(1+p'(q))$$
$$+\left(1-\frac{m}{n}q^{e+1-k}\right)\left(1-\frac{m'r}{n's}q^{e+1-l}\right)(-1+P(q)),$$

or upon clearing fractions,

$$\begin{aligned}&-(nn'-mm'q^{2e+1-l-k})\sum_{i=1}^{e}(rq^{x_i}+sq^{N-x_i})\\ &=-(nn'-mm'q^{2e+1-l-k})(rq+sq^{N-e})(1+p'(q))\\ &+(n-mq^{e+1-k})(n's-m'rq^{e+1-l})(-1+p(q)).\end{aligned}\tag{3.7}$$

Since we have e distinct zeros $\theta(x_i) \in \{\theta(2), \cdots, \theta(N-1)\}$, we let $2 \leqq x_1 < x_2 < \cdots < x_e \leqq N-1$. Noting that all powers of q are positive ($k \leqq e, l \leqq e, 2e+1 \leqq N$), we see that the expression on the left of the equality in (3.6) is $0 \bmod q$. On the right side, the first term is $0 \bmod q$. Thus, we get a contradiction if the second term on the right side is not $0 \bmod q$, or in particular, $nn's \not\equiv 0 \bmod q$. This is listed in Table 1 under 2B(ii).

Example 2. Suppose $a = m/nq^{e+1}$, $b = m'/n'q^{e+1}$, $d = r/s$ where m, n, m', n', r, s are integers not congruent to $0 \bmod q$. Substituting these values into (3.4) and clearing fractions yield

$$-(nn'q-mm')\sum_{i=1}^{e}(rq^{x_i}+sq^{N-x_i}) = -(nn'q-mm')(rq+sq^{N-e})(1+p'(q))$$
$$+q(n-m)(n's-m'r)(-1+p(q)).$$

If we have $N - x_i \geqq 2$ for all i (recall we always have $x_i \geqq 2$), then the expression on the left of the equality is $0 \bmod q^2$, so we get a contradiction if we insist that $mm'r - (n-m)(n's-m'r) \not\equiv 0 \bmod q$. This case is not listed in Table 1 since in general, we do not know if $N - x_i \geqq 2$ for all i.

To consider results on tight t-designs ($t = 2e$) in the known (P and Q) association schemes, the following limiting case is very important. We will see (Corollary 3.5) that knowledge of the x_i's is crucial.

PROPOSITION 3.4. *Suppose the parameters of* $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$ *have the following limiting values*: $d = 0$, $a \to 0$, $b \to \infty$, $ab \to u/vq^p$, *where* u, v *and* p *are integers*, $u, v \not\equiv 0 \bmod q$. *Then the* e *roots of* $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$ *satisfy*

$$-(v-uq^{2e+1-p})\sum_{i=1}^{e}q^{N-x_i} = -(v-uq^{2e+1-p})q^{N-e}(1+p'(q))+v(-1+p(q)).\tag{3.8}$$

Proof. Using (3.4), the proof is straightforward.

TABLE 1

Parameters such that $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$ does not have e distinct zeros in $\{\theta(j) \mid j=2,\cdots,N-1\}$.

d	$a=m/hq^k$ $m, n \not\equiv 0 \bmod q$, k integral	$b=m'/n'q^l$ $m', n' \not\equiv 0 \bmod q$, l integral
1. $d=0$	A. $a \to 0$	(i) b no restriction ($\lvert b\rvert<\infty$)
		(ii) $b\to\infty$, $ab \to u/vq^p$, $u, v \not\equiv 0 \bmod q$ $p \leqq 2e+1$
	B. $k \leqq e$	(i) $b=0$
		(ii) $l \leqq e+1$, $nn' \not\equiv 0 \bmod q$
		(iii) $l>e+1$, $l+k \leqq 2e+1$, $nn' \not\equiv 0 \bmod q$
	C. $k=e+1$	(i) $b=0$, $n-m \not\equiv 0 \bmod q$
		(ii) $l \leqq e$, $n'(n-m) \not\equiv 0 \bmod q$
		(iii) $l \geqq N$, $mm' \not\equiv 0 \bmod q$
	D. $k \geqq e+2$	(i) $b=0$
		(ii) $l \leqq e$, $mn' \not\equiv 0 \bmod q$
		(iii) $l \geqq N$, $mm' \not\equiv 0 \bmod q$
	E. $a \to \infty$	(i) $\lvert b\rvert<\infty$, $b \neq 0$, $l \leqq e$
		(ii) $\lvert b\rvert<\infty$, $l \geqq N$
2. $d=r/s$ $r, s \not\equiv 0 \bmod q$	A. $a=0$	(i) $b=0$
		(ii) $l<e+1$, $n's \not\equiv 0 \bmod q$
		(iii) $l=e+1$, $n's-m'r \not\equiv 0 \bmod q$
		(iv) $l>e+1$, $rm' \not\equiv 0 \bmod q$
	B. $k \leqq e$	(i) $b=0$, $ns \not\equiv 0 \bmod q$
		(ii) $l \leqq e$, $nn's \not\equiv 0 \bmod q$
		(iii) $l=e+1$, $n(n's-m'r) \not\equiv 0 \bmod q$
		(iv) $l>e+1$, $nrm' \not\equiv 0 \bmod q$
		(v) $b \to \infty$, $nr \not\equiv 0 \bmod q$
	C. $k=e+1$	(i) $b=0$, $s(n-m) \not\equiv 0 \bmod q$
		(ii) $l \leqq e$, $n's(n-m) \not\equiv 0 \bmod q$
	D. $k \geqq e+2$	(i) $b=0$, $sm \not\equiv 0 \bmod q$
		(ii) $l \leqq e$, $mn's \not\equiv 0 \bmod q$
	E. $a \to \infty$	(i) $\lvert b\rvert<\lvert b\rvert<\infty$, $l \leqq e$, $sn' \not\equiv 0 \bmod q$
3. $d=rq^t/s$ $t \geqq 1$ $r, s \not\equiv 0 \bmod q$	A. $a=0$	(i) $b=0$
		(ii) $l<e+t+1$, $sn' \not\equiv 0 \bmod q$
		(iii) $l=e+t+1$, $sn'-rm' \not\equiv 0 \bmod q$
		(iv) $l>e+t+1$, $rm' \not\equiv 0 \bmod q$
	B. $k \leqq e$	(i) $b=0$
		(ii) $l \leqq e+t$, $l+k \leqq 2e+1$, $nn's \not\equiv 0 \bmod q$
		(iii) $l=e+t+1$, $k \leqq e-t$, $n(sn'-rm') \not\equiv 0 \bmod q$
		(iv) $l \geqq e+t+2$, $k \leqq e-t$, $nrm' \not\equiv 0 \bmod q$
		(v) $b \to \infty$, $k \leqq e-t$, $nr \not\equiv 0$ *mod q*
	C. $k=e+1$	(i) $b=0$, $n-m \not\equiv 0 \bmod q$
		(ii) $l<e+t+1$, $(n-m)sn' \not\equiv 0 \bmod q$
	D. $k \geqq e+2$	(i) $b=0$, $sm \not\equiv 0 \bmod q$
		(ii) $l \leqq e$, $msn' \not\equiv 0 \bmod q$

TABLE 1 (cont.).

d	$a = m/hq^k$ $m, n \not\equiv 0 \bmod q$, k integral	$b = m'/n'q^l$ $m', n' \not\equiv 0 \bmod q$, l integral
	E. $a \to \infty$	(i) $l < e$, $n's \not\equiv 0 \bmod q$ (ii) $l = e$, $n's - m'r \not\equiv 0 \bmod q$ (iii) $l > e$, $m'r \not\equiv 0 \bmod q$
4. $d = r/sq^t$ $t \geqq 2$ $r, s \not\equiv 0 \bmod q$	A. $a = 0$	(i) $b = 0$ (ii) $l < e$, $n'r \not\equiv 0 \bmod q$ (iii) $l = e$, $r(m' - n') \not\equiv 0 \bmod q$ (iv) $l > e$, $m'r \not\equiv 0 \bmod q$
	B. $k \leqq e$	(i) $b = 0$, $nr \not\equiv 0 \bmod q$ (ii) $l < e$, $nn'r \not\equiv 0 \bmod q$ (iii) $l = e$, $nr(m' - n') \not\equiv 0 \bmod q$ (iv) $l > e$, $nm'r \not\equiv 0 \bmod q$ (v) $b \to \infty$, $rn \not\equiv 0 \bmod q$
	C. $k = e+1$	(i) $b = 0$ (ii) $l < e$, $nn'r \not\equiv 0 \bmod q$ (iii) $l = e$, $nr(m' - n') \not\equiv 0 \bmod q$ (iv) $l > e$, $nm'r \not\equiv 0 \bmod q$ (v) $b \to \infty$, $rn \not\equiv 0 \bmod q$
	D. $e+2 \leqq k < e+t$	(i) $b = 0$, $nr \not\equiv 0 \bmod q$ (ii) $l < e$, $k + l \leqq 2e+1$, $rnn' \not\equiv 0 \bmod q$
	E. $k = e+t$	(i) $b = 0$, $nr + ms \not\equiv 0 \bmod q$ (ii) $l \leqq e - t + 1$, $mn's - nn'r \not\equiv 0 \bmod q$
	F. $K > e+t$	(i) $b = 0$, $sm \not\equiv 0 \bmod q$ (ii) $l < e - t + 2$, $k + l > 2e+1$, $mn's \not\equiv 0 \bmod q$ (iii) $l < e$, $k + l \leqq 2e+1$, $mn's \not\equiv 0 \bmod q$
	G. $a \to \infty$	(i) $\|b\| < \infty$, $l \leqq e - t + 1$, $n's \not\equiv 0 \bmod q$
5. $d = r/sq$ $r, s \not\equiv 0 \bmod q$	A. $a = 0$	(i) $b = 0$, $r + s \not\equiv 0 \bmod q$ (ii) $l < e$, $n'(r + s) \not\equiv 0 \bmod q$ (iii) $l = e$, $n'r + n's - m'r \not\equiv 0 \bmod q$ (iv) $l > e$, $m'r \not\equiv 0 \bmod q$
	B. $k \leqq e$	(i) $b = 0$. $n(r + s) \not\equiv 0 \bmod q$ (ii) $l < e$, $nn'(r + s) \not\equiv 0 \bmod q$ (iii) $l = e$, $nn'r + nn's - nm'r \not\equiv 0 \bmod q$ (iv) $l > e$, $nm'r \not\equiv 0 \bmod q$ (v) $b \to \infty$, $nr \not\equiv 0 \bmod q$
	C. $k = e+1$	(i) $b = 0$, $rn + sn - sm \not\equiv 0 \bmod q$ (ii) $l < e$, $n'(ms - ns - nr) \not\equiv 0 \bmod q$ (iii) $l = e$, $n's(m - n) + nr(m' - n') \not\equiv 0 \bmod q$ (iv) $l > e$, $nm'r \not\equiv 0 \bmod q$ (v) $b \to \infty$, $nr \not\equiv 0 \bmod q$
	D. $k \geqq e+2$	(i) $b = 0$ (ii) $l \leqq e$, $mn's \not\equiv 0 \bmod q$
	E. $a \to \infty$	(i) $l \leqq e$, $n's \not\equiv 0 \bmod q$

TABLE 1 (cont.).

d	$a=m/hq^k$ $m, n \not\equiv 0 \bmod q$, k integral	$b=m'/n'q^l$ $m', n' \not\equiv 0 \bmod q$, l integral
6. $d \to \infty$	A. $a=0$	$\lvert b\rvert<\infty$
		(i) $l<e$
		(ii) $l=e$, $m'-n'\not\equiv 0 \bmod q$
		(iii) $l>e$
	B. $a\neq 0, \lvert a\rvert<\infty$	$\lvert b\rvert<\infty$
	α) $k\leqq e+1$	(i) $l\leqq e-1$, $nn'\not\equiv 0 \bmod q$
		(ii) $l=e$, $n(n'-m')\not\equiv 0 \bmod q$
		(iii) $l\geqq e+1$, $m'n\not\equiv 0 \bmod q$
	β) $k\geqq e+2$	(i) $l<e$, $k+l\leqq 2e+1$, $nn'\not\equiv 0 \bmod q$
	C. $a\to\infty$	$\lvert b\rvert<\infty$, $b\neq 0$
	D. $a\to\infty$	$b\to 0$
		$ab\to g/hq^x$; $g, h\not\equiv 0 \bmod q$, x integral
		$bd\to u/vq^p$; $u, v\not\equiv 0 \bmod q$, p integral
		(i) $x<e+1$
		(α) $p<1$, $p+e<x$, $hu\not\equiv 0 \bmod q$
		(β) $p<1$, $p+e>x$, $gv\not\equiv 0 \bmod q$
		(γ) $p\geqq 1$, $hu\not\equiv 0 \bmod q$
		(ii) $x=e+1$
		(α) $p<1$, $gv\not\equiv 0 \bmod q$
		(β) $p=1$, $gv-hu\not\equiv 0 \bmod q$
		(γ) $p>1$, $hu\not\equiv 0 \bmod q$
		(iii) $e+1<x<2e+1$
		(α) $p\leqq 1$, $gv\not\equiv 0 \bmod q$
		(β) $p>1$, $x<p+e$, $hu\not\equiv 0 \bmod q$
		(γ) $p>1$, $x=p+e$, $gv-hu\not\equiv 0 \bmod q$
		(δ) $p>1$, $x>p+e$, $gv\not\equiv 0 \bmod q$
		(iv) $x=2e+1$
		(α) $p<e+1$, $gv\not\equiv 0 \bmod q$
		(β) $p=e+1$, $gv-hu\not\equiv 0 \bmod q$
		(γ) $p>e+1$, $hu\not\equiv 0 \bmod q$
		(v) $x>2e+1$
		(α) $p\leqq e+1$, $gv\not\equiv 0 \bmod q$

COROLLARY 3.5. *Under the hypothesis of Proposition* 3.4, $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$ *does not have* e *distinct zeros* $\theta(x_i)$, $x_1<x_2<\cdots<x_e$, *in* $\{\theta(j)\mid j=2,\cdots,N-1\}$ *if*

(i) $p\leqq 2e+1$, *or*

(ii) $p>2e+1$ *and* $N-x_e<\min\{p-2e-1, N-e\}$.

Proof. (i) Since $2\leqq x_i\leqq N-1$, the left side of (3.8) is $0 \bmod q$, whereas on the right, $v\not\equiv 0 \bmod q$ by assumption. This is a contradiction.

(ii) We see that (3.8) is equivalent to

$$-(vq^{p-2e-1}-u)\sum_{i=1}^{e} q^{N-x_i}=-(vq^{p-2e-1}-u)q^{N-e}(1+p'(q))+vq^{p-2e-1}(-1+p(q)).$$

Both sides are now $0 \bmod q$. However, if $N-x_e<\min\{p-2e-1, N-e\}$, then we can divide both sides by q^{N-x_e}. The left side will not be 0 mod q (since $u \not\equiv 0 \bmod q$), but the right side will still be 0 mod q. This is a contradiction.

4. Applications. We shall show that the results of the previous section establish the nonexistence of perfect e-codes and tight e-designs in the classical association schemes of types $A_{v-1}, {}^2A_{2N}, {}^2A_{2N-1}, B_N, C_N, D_N, {}^2D_{N+1}$, and the affine matrix schemes. We refer the reader to [2] for a brief description.

As an example, consider the association scheme of type D_N. The (i, k) entry of the P eigenmatrix is given by

$$P_k(i) = v_k P_k(\theta(i), 0, 0, q^{-N}, -q; q).$$

Hence we have $a=0$, $b=0$, $d=-1$. These parameters lie in Table 1 under 2A(i). We conclude that the Lloyd polynomial does not have e distinct roots in $\{\theta(j) | j=1, \cdots, N\}$ and so there are no perfect e-codes.

The (i, k) entry of the Q eigenmatrix in type D_N is given by

$$Q_k(i) = \mu_k {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, -q^{-N+k} \\ 0, q^{-N} \end{matrix} \middle| q; q\right).$$

The ${}_3\phi_2$ is derived from (2.1) by considering the limiting case $d=0$, $a \to \infty$, $b \to 0$, $ab \to -q^{-N-1}$. To apply Corollary 3.5, we will need to establish some lemmas giving us more information on the roots of the corresponding Wilson polynomial. (For this reason, Table 1 is inapplicable here.)

LEMMA 4.1. *If $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, 0; q)$ has e distinct roots $\theta(x_1), \cdots, \theta(x_e)$ in $\{\theta(j) | j=2, \cdots, N-1\}$, $2 \leqq x_1 < \cdots < x_e \leqq N-1$, then $x_e > 2e$ except in the following cases*:

(i) *If $q>0$, $x_1=2$, $x_2=4, \cdots, x_e=2e$.*

(ii) *If $q<0$ and e is even, then either*

$$x_1=2, \quad x_2=3, \quad x_3=6, \quad x_4=7, \cdots, x_{e-1}=2(e-1), \quad x_e=2e-1, \textit{ or}$$

$$x_1=3, \quad x_2=4, \quad x_3=7, \quad x_4=8, \cdots, x_{e-1}=2(e-1)+1, \quad x_e=2e.$$

(iii) *If $q<0$ and e is odd, then*

$$x_1=2, \quad x_2=3, \quad x_3=6, \quad x_4=7, \cdots, x_{e-1}=2(e-1)+1, \quad x_e=2e.$$

Proof. Recall that the open interval between two consecutive zeros of an orthogonal polynomial must contain a point of the spectrum [4, p. 59]. Also, since $d=0$, we have $\theta(x) = 1-q^{-x}$.

(i) If $q>0$, then $\theta(k) < \theta(k+1)$ for all positive integers k. Hence $x_e > 2e$ unless $x_1=2$, $x_2=4, \cdots, x_2=2e$.

(ii) If $q<0$, then for nonnegative integers n and k, we have

$$\theta(2n) < \theta(2n+2), \quad \theta(2n+1) > \theta(2n+3), \quad \theta(2n) < \theta(2k+1).$$

If e is even, then there are only two ways to choose the x_i $(x_1 < x_2 < \cdots < x_e)$ so that $x_e \leqq 2e$. We take $(e/2)$ of the x_i to be even and $(e/2)$ of the x_i to be odd to get either

$$\theta(2) < \theta(6) < \theta(10) < \cdots < \theta(2e-2) < \theta(2e-1) < \theta(2e-5) < \cdots < \theta(7) < \theta(3),$$

or

$$\theta(4) < \theta(8) < \cdots < \theta(2e-4) < \theta(2e) < \theta(2e-1) < \theta(2e-5) < \cdots < \theta(7) < \theta(3).$$

(iii) If $q<0$ and e is odd, we choose $((e+1)/2)$ of the x_i to be even and $((e-1)/2)$ of the x_i to be odd

$$\theta(2)<\theta(6)<\cdots<\theta(2e-4)<\theta(2e)<\theta(2e-1)<\cdots<\theta(7)<\theta(3).$$

LEMMA 4.2. *For the association scheme D_N, $\Psi_e^*(\theta^*(3))\neq 0$ for $e\geqq 1$.*

Proof. If $e\geqq 2$ and $\Psi_e^*(\theta^*(3))=0$, then

$$A^*\sum_{j=0}^{2}\frac{(q^{-e})_j(-q^{-N+e+1})_j(q^{-2})_j q^j}{(q^{-N+1})_j(q)_j}=0.$$

Multiplying the above expression by $(1-q^{-N+1})(1-q^{-N+2})(1-q)(1-q^2)q^{2N-2}$ yields $qf(q)-(q^{N-2}-1)(q^{N-e-1}+1)(q^e-1)(q^2-1)^2=0$ where $f(q)$ is a polynomial in q with integer coefficients. Since $2e+1\leqq N$ and $e\geqq 2$, the second term on the left side of the equality is not 0 mod q. This contradiction implies $\Psi^*(\theta^*(3))\neq 0$. A similar calculation establishes the result for $e=1$.

THEOREM 4.3. *The association scheme of type D_N has no tight t-design of order $e(t=2e)$ unless $2e+1=N$.*

Proof. Since $\Psi_e^*(\theta^*(x))=P_e(\tilde{\theta}(x-1),\infty,0,q^{-N},0;q)$, we apply Corollary 3.5 with $d=0$, $a\to\infty$, $b\to 0$, $ab\to q^{-N-1}$. We take $u=-1$, $v=1$ and $p=N+1$. Since $2e+1\leqq N$, we have $p>2e+1$ and so $\Psi_e^*(\theta^*(x))$ does not have e distinct roots in $\{\theta(j)\,|\,j=2,\cdots,N-1\}$ if $N-x_e>\min\{P-2e-1,N-e\}$, or equivalently $x_e>2e$. By Lemma 4.1 and Lemma 4.2, the only possibility is $x_1=2$, $x_2=4,\cdots,x_e=2e$ if $q>0$. Checking $\Psi_e^*(\theta(2))=0$ yields $e=(N-1)/2$. Tight designs where $N=2e+1$ are known to exist [11, p. 661].

We summarize our results in Table 2. We list the P and Q polynomials for the classical association schemes and affine matrix schemes along with the values of the parameters for the Askey-Wilson polynomials. Finally, we give the references which establish the nonexistence of perfect e-codes and tight t-designs. Note that Table 1 gives the nonexistence of perfect e-codes for all the schemes except types B_N and C_N (these will be done later in this paper). The results for e-designs in types A_{v-1}, ${}^2A_{2N}$, ${}^2A_{2N-1}$, B_N, C_N, D_N and ${}^2D_{N+1}$ follow from Corollary 3.5 and Lemma 4.1, and for ${}^2A_{2N-1}$, B_N and C_N, a calculation similar to Lemma 4.2 showing $\Psi_e^*(\theta^*(2))\neq 0$.

We conclude this section by showing that there are no perfect e-codes for $e\geqq 2$ in the association schemes of types B_N and C_N. The sum of the roots formula for the Lloyd polynomial fails to yield a contradiction in this case, so we will use the product of the roots formula.

The polynomial for this scheme is

$$P_k(i)=v_k p_k(\theta(i),0,0,q^{-N-1},-q^{-1};q).$$

Assuming $\Psi_e(\theta(x))$ has e distinct roots $\theta(x_i)$, $x_1<\cdots<x_e$, in $\{\theta(j)\,|\,j=2,\cdots,N-1\}$, Proposition 3.2 yields

$$\prod_{i=1}^{e}\theta(x_i)=\frac{(q^{-N+1})_e}{q^e}\,{}_3\phi_2\left(\begin{matrix}q^{-e},-q^{-N},q\\ 0,q^{-N+1}\end{matrix}\,\middle|\,q;q\right). \tag{4.1}$$

We shall need a transformation of a terminating ${}_3\phi_2$ [10, p. 101]

$${}_3\phi_2\left(\begin{matrix}q^{-e},A,B\\ C,D\end{matrix}\,\middle|\,q;q\right)=\frac{(C/A)_e A^e}{(C)_e}\,{}_3\phi_2\left(\begin{matrix}q^{-e},A,D/B\\ Aq^{1-e}/C,D\end{matrix}\,\middle|\,q,\frac{Bq}{C}\right).$$

TABLE 2
Askey-Wilson parameters.

Scheme	$(cq = q^{-N})$	Reference for no codes or designs
1. $N \times M$ matrices/$GF(q)$, $N \leqq M$		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, 0 \\ q^{-M}, q^{-N} \end{matrix}\,\middle\vert\, q; q\right)$	$a = q^{-M-1}$ $b = 0$ $d = 0$	Table 1 1D(i)
$Q_k(i) = P_k(i)$	same as above	
2. $N \times N$ Hermitian matrices/$GF(q^2)$ $(q \to -q)$		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} (-q)^{-k}, (-q)^{-i}, 0 \\ (-1)^{N+1}q^{-N}, (-q)^{-N} \end{matrix}\,\middle\vert\, q; q\right)$	$a = (-1)^N q^{-N-1}$ $b = 0$ $d = 0$	Table 1 1D(i)
$Q_k(i) = P_k(i)$	same as above	
3. $N \times N$ Skew symmetric matrices/$GF(q)$, $(q \to q^2), 2 \neq q$		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} q^{-2k}, q^{-2i}, 0 \\ q^{-N}, q^{-N+1} \end{matrix}\,\middle\vert\, q^2; q^2\right)$	$a = q^{-N-1}$ $b = 0$ $d = 0$	Table 1 1D(i)
$Q_k(i) = P_k(i)$	same as above	
4. Type A_{v-1} (q-Johnson)		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, q^{-v-1+i} \\ q^{N-v}, q^{-N} \end{matrix}\,\middle\vert\, q; q\right)$	$a = 0$ $b = 1$ $d = q^{N-v-1}$	Table 1 4A(ii)
$2N \leqq V$	$(v - N + 1 \geqq 2)$	
$Q_k(i) = \mu_k\,{}_3\phi_2\left(\begin{matrix} q^{-i}, q^{-k}, q^{-v-1+k} \\ q^{N-v}, q^{-N} \end{matrix}\,\middle\vert\, q; q\right)$	$a = q^{N-v-1}$ $b = q^{-N-1}$ $d = 0$	Table 1 1C(iii), 1D(iii)
5. Type ${}^2A_{2N}m$ $(q \to q^2)$		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} q^{-2k}, q^{-2i}, -q^{-2N+2i-3} \\ 0, q^{-2N} \end{matrix}\,\middle\vert\, q^2; q^2\right)$	$a = 0$ $b = 0$ $d = -q^{-3}$ $r = -1, \quad s = q$ $s \not\equiv 0 \bmod q^2$	Table 1 5A(i)
$Q_k(i) = \mu_k\,{}_3\phi_2\left(\begin{matrix} q^{-2i}, q^{-2k}, -q^{-2N+2k-3} \\ 0, q^{-2N} \end{matrix}\,\middle\vert\, q^2; q^2\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-2N-5}$ $u = -1, \quad v = q$ $p = N + 2$	Corollary 3.5(ii) Lemma 4.1(i)
6. Type ${}^2A_{2N-1}$, $(q \to q^2)$		
$P_k(i) = \nu_k\,{}_3\phi_2\left(\begin{matrix} q^{-2k}, q^{-2i}, -q^{-2N+2i-1} \\ 0, q^{-2N} \end{matrix}\,\middle\vert\, q^2; q^2\right)$	$a = 0$ $b = 0$ $d = -1/q$ $r = -1, \quad s = q$ $s \not\equiv 0 \bmod q^2$	Table 1 2A(i)

TABLE 2 (cont.).

Scheme	$(cq = q^{-N})$	Reference for no codes or designs
$Q_k(i) = \mu_k\, {}_3\phi_2\left(\begin{matrix} q^{-2i}, q^{-2k}, -q^{-2N+2k-1} \\ 0, q^{-2N} \end{matrix} \,\middle\vert\, q^2; q^2\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-2N-3}$ $u = -1$, $v = q \not\equiv 0 \bmod q^2$ $p = N+1$	Corollary 3.5(ii) Lemma 4.1(i) $\Psi_e^*(\theta^*(2)) \neq 0$
7. Types B_N and C_N		
$P_k(i) = \nu_k\, {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, -q^{-N+i-1} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a = 0$ $b = 0$ $d = -q^{-1}$	Theorem 4.8
$Q_k(i) = \mu_k\, {}_3\phi_2\left(\begin{matrix} q^{-i}, q^{-k}, -q^{-N+k-1} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-N-2}$ $u = -1, \quad v = 1$ $p = N+2$	Corollary 3.5(ii) Lemma 4.1 $\Psi_e^*(\theta^*(2)) \neq 0$
8. Type D_N		
$P_k(i) = \nu_k\, {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, -q^{-N+i} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a = 0$ $b = 0$ $d = -1$	Table 1 2A(i)
$Q_k(i) = \mu_k\, {}_3\phi_2\left(\begin{matrix} q^{-i}, q^{-k}, -q^{-N+k} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-N-1}$	Theorem 4.3
9. Type D_N		
$P_k(i) = \nu_k\, {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-2i}, -q^{-N+k} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-N-1}$ $u = -1, \quad v = 1$ $p = N+1$	Corollary 3.5(ii) $x \to 2x$
$Q_k(i) = \mu_k\, {}_3\phi_2\left(\begin{matrix} q^{-2k}, q^{-i}, q^{-N+i} \\ 0, q^{1-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a = 0$ $b = 0$ $d = -1$ $k \to 2k$	Table 1 2A(i)
10. Type ${}^2D_{N+1}$		
$P_k(i) = \nu_3\, {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-i}, -q^{-N+i-2} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a = 0$ $b = 0$ $d = -q^{-2}$	Table 1 4A(i)
$Q_k(i) = \mu_k\, {}_3\phi_2\left(\begin{matrix} q^{-i}, q^{-k}, -q^{-N+k-2} \\ 0, q^{-N} \end{matrix} \,\middle\vert\, q; q\right)$	$a \to 0$ $b \to \infty$ $d = 0$ $ab \to -q^{-N-3}$ $u = -1, \quad v = 1$ $p = N+3$	Corollary 3.5(ii) Lemma 4.1

Thus, (4.1) is equivalent to

$$\prod_{i=1}^{e} \theta(x_i) = (q^{-N})_e \sum_{j=0}^{e} \frac{(q^{-e})_j(-1)^j}{(q^{N-e+1})_j}, \tag{4.2}$$

or, upon multiplying by q^{Ne+e},

$$q^{Ne+e} \prod_{i=1}^{e} \theta(x_i) = q^{Ne+e}(q^{-N})_e \sum_{j=0}^{e} \frac{(q^{-e})_j(-1)^j}{(q^{N-e+1})_j}. \tag{4.3}$$

Next, we state without proof, a result from [11, p. 630].

PROPOSITION 4.4. *Let k, y and M be integers, $0 \leqq k, y \leqq M$ and $d \neq 0$. Then*

$$ {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-y}, -d^{-1}q^{-2M+y-1} \\ 0, q^{-M} \end{matrix} \middle| q; q\right)$$
$$= (-d)^{-k}q^{-k(M+1)} {}_3\phi_2\left(\begin{matrix} q^{-k}, q^{-M+y}, -dq^{M-y+1} \\ 0, q^{-M} \end{matrix} \middle| q; q\right).$$

In particular, setting $k=e$, $y=x-1$, $M=N-1$ and $d=q^{-N}$ in Proposition 4.4 and checking the expression for $\Psi_e(\theta(x))$, we see that $\Psi_e(\theta(x)) = \Psi_e(\theta(N-x+1))$. Hence we record

COROLLARY 4.5. *In types B_N and C_N, $\theta(x)$ is a root of Ψ_e if and only if $\theta(N-x+1)$ is.*

LEMMA 4.6. *Set $f(q) = q^{Ne+e}\prod_{i=1}^{e} \theta(x_i)$. Then $f(q) = (-1)^e + (-1)^{e-1}q^{2x_1} +$ higher ordered terms in q.*

Proof. Suppose e is even. We take the $e/2$ smallest x_i: $x_1 < x_2 < \cdots < x_{e/2}$. Hence, by Corollary 4.5,

$$\begin{aligned} f(q) &= q^{Ne+e} \prod_{i=1}^{e/2} \theta(x_i)\theta(N-x_i+1) \\ &= \prod_{i=1}^{e/2} (q^{2x_i}-1)(q^{2N-2x_i+2}-1) \\ &= 1-(q^{2x_1}+q^{2x_2}+\cdots)+\text{higher order terms in } q. \end{aligned}$$

If e is odd, then Corollary 4.5 implies that $\theta(k) = \theta(N-k+1)$ for some k. Hence $k=(N+1)/2$ and so N must be odd also.

$$\begin{aligned} f(q) &= q^{Ne+e}\theta\left(\frac{N+1}{2}\right) \prod_{i=0}^{(e-1)/2} \theta(x_i)\theta(N-x_i+1) \\ &= (q^{N+1}-1) \prod_{i=0}^{(e-1)/2} (q^{2x_i}-1)(q^{2N-2x_i+2}-1) \\ &= 1-(q^{2x_1}+q^{2x_2}+\cdots)+\text{higher ordered terms in } q. \end{aligned}$$

LEMMA 4.7. *Let*

$$h(q) = q^{Ne+e}(q^{-N})_e \sum_{j=0}^{e} \frac{(q^{-e})_j(-1)^j}{(q^{N-e+1})_j}.$$

Then $h(q) = (-1)^e + (-1)^{e-1}q^2 +$ higher ordered terms in q.

Proof. If $h(q) = \sum_{j=0}^{e} C(j)$, then

$$\begin{aligned} C(e) &= (q^e-1)(q^{e-1}-1)\cdots(q-1) \\ &= (-1)^e + (-1)^{e+1}(q+q^2+\cdots q^e) + \text{terms in } q \text{ of order} \geqq 3, \end{aligned}$$

$$C(e-1)=q(q^N-1)(q^e-1)(q^{e-1}-1)\cdots(q^2-1)$$
$$=(-1)^e q+(-1)^{e-1}(q^3+\cdots+q^{e+1}+q^{N+1})+\text{terms in } q \text{ of order} \geqq 6,$$

and for $0 \leqq j \leqq e-2$, we have

$$C(j)=q^{e(e+1)/2}(q^N-1)(q^{N-1}-1)\cdots(q^{N-e+j-1}-1)(q^{-e})_j$$
$$=(-1)^e \cdot q^{(e+1)e/2-ej+(j-1)j/2}+\text{higher ordered terms in } q.$$

In particular, for $0 \leqq j \leqq e-2$, we have $e(e+1)/2-ej+(j-1)j/2 \geqq 3$ (recall $e \geqq 2$). Thus $h(q)=(-1)^e+(-1)^{e-1}q^2+$higher ordered terms in q.

From (4.3) we have $f(q)=h(q)$ and hence $f(q)-(-1)^e=h(q)-(-1)^e$. But since $x_1 \geqq 2$, we have $f(q)-(-1)^e \equiv 0 \bmod q^3$, whereas $h(q)-(-1)^3 \equiv (-1)^{e-1}q^2 \bmod q^3$. This contradiction shows that for $e \geqq 2$, there are no perfect e-codes in the association schemes of type B_N and C_N.

We summarize our main results:

THEOREM 4.8. (i) *There are no perfect e-codes for $e \geqq 1$ in the dual polar spaces of types A_{v-1}, ${}^2A_{2N}$, ${}^2A_{2N-1}$, D_N, ${}^2D_{N+1}$, and the affine matrix schemes, and no perfect e-codes for $e \geqq 2$ in the spaces B_N and C_N.*

(ii) *There are no tight designs of order e in the spaces of types A_{v-1}, ${}^2A_{2N}$, ${}^2A_{2N-1}$, B_N, C_N, D_N, ${}^2D_{N+1}$, and the affine matrix schemes with the exception of tight designs of order $e=(N-1)/2$ in D_N.*

5. Conclusion. We have determined the nonexistence of perfect e-codes and tight t-designs ($t=2e$) in the classical association schemes and affine matrix schemes for $e \geqq 1$, except in type D_N where tight designs exist if $e=(N-1)/2$ and in types B_N and C_N for $e=1$ where perfect codes are still undetermined. For types B_N and C_N, the sphere packing bound says that if a perfect 1-code exists, then [11, p. 659]

$$\frac{q^{N+1}-1}{q-1} \,\Big|\, (1+q)(1+q^2)\cdots(1+q^N). \tag{5.1}$$

We are grateful to J. Shearer for the following proposition.

PROPOSITION 5.1. *For q integral, $q \neq 1$, (5.1) implies $N+1=2^m$ for some integer m.*

Proof. Suppose (5.1) holds and $N+1$ has an odd prime factor p, then $(q^p-1)/(q-1) \mid (1+q)\cdots(1+q^N)$. Let $r \mid (q^p-1)/(q-1)$ for r prime. Then $r \mid 1+q^s$ for some s (pick s minimal). Hence $q^s \equiv -1 \bmod r$ and thus $q^{2s} \equiv 1 \bmod r$. We also have $q^p \equiv 1 \bmod r$. If $r \neq 2$ then clearly $s<p$ and $p \mid 2s$. Since p is odd, we see that $p \mid s$ which contradicts $p<s$. If $r=2$, then $1+q+\cdots+q^{p-1} \equiv 0 \bmod 2$ which contradicts p being odd.

Thus, if there are perfect 1-codes in types B_N and C_N, then $N+1=2^m$. This is the same condition for perfect 1-codes in the Hamming scheme. Perfect 1-codes are known to exist in C_3 [11, p. 660], [12].

To conclude this paper, we present another proof of Proposition 2.1. We establish that

$$P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$$
$$=B\sum_{n=0}^{e} w(n, q^{-N-1}, d, a, b)P_n(\theta(x), a, b, q^{-N-1}, d; q) \tag{5.2}$$

for a nonzero constant B (to be determined).

We shall need the following ${}_4\phi_3$ transformation [8, p. 167]

(5.3)
$$
\begin{aligned}
&{}_4\phi_3\left(\begin{matrix} q^{-e}, q^{e-1}ABCD, A/Z, AZ \\ AB, AC, AD \end{matrix}\;\middle|\; q; q\right) \\
&= \frac{(BC)_e(BD)_e(AB^{-1})^e}{(AC)_e(AD)_e}\,{}_4\phi_3\left(\begin{matrix} q^{-e}, q^{e-1}ABCD, B/Z, BZ \\ AB, BC, BD \end{matrix}\;\middle|\; q; q\right).
\end{aligned}
$$

Recalling (2.1), we see that (5.3) implies

(5.4)
$$
\begin{aligned}
&P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q) \\
&= \frac{(aq/d)_e(bq)_e}{(aq^2)_e(bdq^2)_e}(dq)^e P_e(\tilde{\theta}(N-x), b, aq, q^{-N}, 1/dq; q).
\end{aligned}
$$

Note that $(1-q^{x-N+s})(1-d^{-1}q^{-x+s}) = d^{-1}q^s(\theta(x)-\theta(N-s))$, so that $(q^{x-N})_m (d^{-1}q^{-x})_m$ is a polynomial of degree m in $\theta(x)$. Since $\{p_0(\theta(x)), \cdots, p_m(\theta(x))\}$ forms a basis for the vector space of polynomials of degree at most m in $\theta(x)$, we have

(5.5)
$$
(q^{x-N})_m(d^{-1}q^{-x})_m = \sum_{n=0}^{m} A_{m,n}P_n(\theta(x), a, b, q^{-N-1}, d; q)
$$

for constants $A_{m,n}$. Askey and Wilson have calculated $A_{m,n}$[1, p. 1014].

(5.6)
$$
\begin{aligned}
&A_{m,n} \\
&= \frac{(abq)_n(1-abq^{2n+1})(aq)_n(bdq)_n(a^{-1}dq^{-m})_m(b^{-1}q^{-m})_m(q^{-N})_m(q^{-m})_n(ab)^m q^{2\binom{m+1}{2}}}{(q)_n(1-abq)(bq)_n(ad^{-1}q)_n(abq^{m+2})_n(abq^2)_m d^{m+n}}.
\end{aligned}
$$

Thus, from (5.4), (5.5) and (5.6), we clearly have that the coefficient $B_{m,n}$ of $P_n(\theta(x), a, b, q^{-N-1}, d; q)$ in $P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)$ is

$$
\begin{aligned}
&B_{m,n} = \\
&\frac{(aq/d)_e(bq)_e(dq)^e(abq)_n(1-abq^{2n+1})(aq)_n(bdq)_n(q^{-e})_n(q^{e+2}ab)_n(q^{-N})_n q^{n+\binom{n}{2}}(-1)^n}{(aq^2)_e(bdq^2)_e(q)_n(1-abq)(bq)_n(ad^{-1}q)_n(abq^2)_n(q^{n+2}ab)_n(q^{-N+1})_n d^n} \\
&\quad\times {}_3\phi_2\left(\begin{matrix} q^{-e+n}, q^{e+n+2}ab, q^{-N+n} \\ abq^{2n+2}, q^{-N+n+1} \end{matrix}\;\middle|\; q; q\right) \\
&= (dq)^e\frac{(aq/d)_e(bq)_e(q^{-e})_e(abq^{N+2})_e}{(aq^2)_e(bdq^2)_e(q^{N-e})_e(abq^2)_e} w(n, q^{-N-1}, d, a, b)
\end{aligned}
$$

where we have recalled the evaluation (3.5) and the definition of the weights (2.6).

Hence,

$$
\begin{aligned}
P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q) &= (dq)^e\frac{(aq/d)_e(bq)_e(q^{-e})_e(abq^{N+2})_e}{(aq^2)_e(bdq^2)_e(q^{N-e})_e(abq^2)_e} \\
&\quad\times \sum_{n=0}^{e} w(n, q^{-N-1}, d, a, b)P_n(\theta(x), a, b, q^{-N-1}, d; q)
\end{aligned}
$$

and so finally, recalling (2.7) and (2.9), we have

$$
\Psi_e(\theta(x)) = (dq)^{-e}\frac{(aq^2)_e(bdq^2)_e(q^{N-e})_e(abq^2)_e}{(aq/d)_e(bq)_e(q^{-e})_e(abq^{N+2})_e}P_e(\tilde{\theta}(x-1), aq, b, q^{-N}, dq; q)
$$

which establishes (2.11).

Acknowledgment. The author gratefully acknowledges Dennis Stanton for his many helpful comments and suggestions.

REFERENCES

[1] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, this Journal, 10 (1979), pp. 1008-1016.

[2] E. Bannai and T. Ito, *Algebraic Combinatorics. Part* 1, Benjamin-Cummings, Menlo Park, CA, 1984.

[3] N. Biggs, *Perfect codes in graphs*, J. Comb. Theory Ser. B, 15 (1973), pp. 289-296.

[4] T. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[5] P. Delsarte, *An algebraic approach to the association schemes of coding theory*, Philips Research Reports Suppl. 10, 1973.

[6] D. Leonard, *Orthogonal polynomials, duality and association schemes*, this Journal, 13 (1982), pp. 656-663.

[7] F. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland Mathematical Library, New York, Vol. 16, 1977.

[8] D. Sears, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc., 53 (1951), pp. 158-180.

[9] L. Slater, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[10] D. Stanton, *Product formulas for q-Hahn polynomials*, this Journal, 11 (1980), pp. 100-107.

[11] ———, *Some q-Krawtchouk polynomials on Chevalley groups*, Amer. J. Math., 102 (1980), pp. 625-662.

[12] J. Thas, *Two infinite classes of perfect codes in metrically regular graphs*, J. Comb. Theory Ser. B, 23 (1977), pp. 236-238.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

016

TIME TO REACH STATIONARITY IN THE BERNOULLI-LAPLACE DIFFUSION MODEL*

PERSI DIACONIS† AND MEHRDAD SHAHSHAHANI‡

Abstract. Consider two urns, the left containing n red balls, the right containing n black balls. At each time a ball is chosen at random in each urn and the two balls are switched. We show it takes $\frac{1}{4}n \log n + cn$ switches to mix up the urns. The argument involves lifting the urn model to a random walk on the symmetric group and using the Fourier transform (which in turn involves the dual Hahn polynomials). The methods apply to other "nearest neighbor" walks on two-point homogeneous spaces.

Key words. Markov chains, eigenvalues, Gelfand pairs, Hahn polynomials

AMS(MOS) subject classifications. Primary 60B15; secondary 60J20

1. Introduction. Daniel Bernoulli and Laplace introduced a simple model to study diffusion. Consider two urns, the left containing n red balls, the right containing n black balls. A ball is chosen at random in each urn and the two balls are switched. It is intuitively clear that after many such switches the urns will be well mixed, about half red and half black. The process is completely determined by the number of red balls in one of the urns. The stationary distribution may be described as the law of the composition of n balls drawn without replacement from n red and n black balls

$$\pi_n(j)=\binom{n}{j}\binom{n}{n-j}\bigg/\binom{2n}{n}, \quad 0 \leqq j \leqq n. \tag{1.1}$$

The main question addressed here is the rate of convergence to the stationary distribution. Let P_k be the law of the process after k steps. Distance to stationarity will be measured by variation distance

$$\|P_k-\pi_n\| = \frac{1}{2}\sum_j |P_k(j)-\pi_n(j)|. \tag{1.2}$$

THEOREM 1. *Let P_k be the probability distribution of the number of red balls in one urn of the Bernoulli-Laplace diffusion model based on n of each color.*

(1.3) *Let $k=\frac{1}{4}n\log n+cn$ for $c \geqq 0$. Then for a universal constant a, $\|P_k-\pi_n\| \leqq ae^{-2c}$.*

(1.4) *With k as above, and arbitrary negative c in $[-\frac{1}{4}\log n, 0]$ there is a universal positive b such that $\|P_k-\pi_n\| \geqq 1-be^{4c}$.*

Remarks. Theorem 1 gives a sharp sense in which $\frac{1}{4}n \log n$ switches are needed: for somewhat fewer switches, the variation distance is essentially at its maximum value of 1. For somewhat more switches, the distance tends to zero exponentially fast. There is a fairly sharp cut-off at $\frac{1}{4}n \log n$.

* Received by the editors August 26, 1985, and in revised form October 30, 1985.

† Department of Statistics, Stanford University, Stanford, California 94305.

‡ Jet Propulsion Laboratories, Pasadena, California 91109.

A somewhat stronger result, starting with r red balls and $n-r$ black balls is proved in §§3 and 4 of this paper. The argument uses Fourier analysis on the symmetric group S_n and the homogeneous space $S_n/S_r \times S_{n-r}$. This last space is a "Gelfand pair" so the Fourier analysis is essentially commutative, involving spherical functions that turn out to be the dual Hahn polynomials. Section 2 develops the needed background material. Section 5 describes how essentially the same argument applies to nearest neighbor random walks on two-point homogeneous spaces. These include the Ehrenfest's model of diffusion (random walk on the "cube" Z_2^n) and random walk on the k dimensional subspaces of a vector space over a finite field.

The Bernoulli-Laplace process is discussed by Feller (1968, p. 423) who gives historical references. See also Johnson and Kotz (1977, pp. 205-207). We conclude this section by listing several "real world" problems where the model appears.

1) *r sets of an n set.* Let X be the set of r element subsets of $\{1, 2, \cdots, n\}$ so $|X| = \binom{n}{r}$. A random walk can be constructed on X as follows. Begin at $\{1, 2, \cdots, r\}$. Each time, pick an element from the present set and an element from its complement, and switch the two elements. This is a nearest neighbor walk using the metric: $d(x, y) = r - |x \cap y|$. The stationary distribution is the uniform distribution over X. Professor Laurel Smith points out that when $n = r = 2$, this becomes nearest neighbor random walk on the vertices of an octahedron. The rate of convergence to stationarity is the same as the rate for the Bernoulli-Laplace model with r red and $n-r$ black balls as shown in Lemma 1 below.

Walks of this type are an essential ingredient of the currently fashionable approach to combinatorial optimization called simulated annealing. Given a function $f: X \rightarrow \mathbb{R}$ annealing algorithms perform a stochastic search for the minimum of f based on the walk. Kirkpatrick et al. (1983) or Aragon et al. (1984) give further details.

2) *Moran's model in mathematical genetics.* Moran (1958) introduced a simple process to model the stochastic behavior of gene frequencies in a finite population. In one version, there is a population of n individuals each of whom is either of type A_1 or A_2. At each time, an individual is chosen at random to reproduce. After reproduction, an individual is chosen at random to die. The model allows mutation of the newborn (from A_1 to A_2 at rate u, from A_2 to A_1 at rate v). If $u = v = 1$, the transition mechanism of Moran's model becomes precisely the transition mechanism of the Bernoulli-Laplace diffusion. A clear discussion of Moran's model is in Ewens (1979, §3.3).

Ewens gives numerous references to eigenvalue-eigenvector analysis of this Markov chain. We will use part of this literature as an ingredient of our analysis.

3) *Piaget's randomization board.* In investigating children's ability to comprehend randomness, Piaget and Inhelder (1975, pp. 1-25) worked extensively with the physical device shown in Fig. 1. The left side of the box contains 8 red balls, the right side contains 8 white balls. When the box is tipped about an axis through its center (like a child's see-saw) the balls roll across to the other side. Usually one or two balls "change sides"—a red moving into the blacks or vice versa.

Piaget asked children of varying ages questions such as "how long will we have to wait until the balls are mixed up?" Answer: 5-10 switches for 8 reds. He also asked "how long will we have to wait until the balls return to the way they started?" Piaget offered an answer to the second problem for 10 reds and 10 blacks: about 185,000 moves are needed! Naturally, children (and most adults) do not guess it takes such a long time.

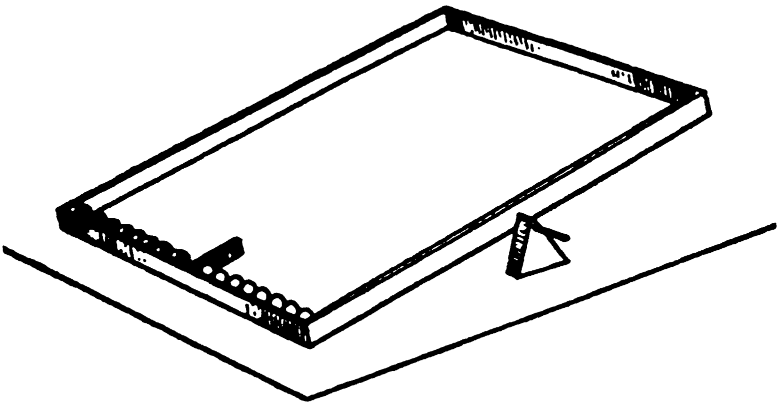


FIG. 1

One natural reaction for a mathematician is "how on earth does he know?" We began work on this paper by considering the Bernoulli–Laplace model (ignorant of its origins). Theorem 1 shows the random walk is "rapidly mixing," to use terminology of Aldous (1983). That is, the time to reach stationarity is of the order of the log of the number of states. Aldous (1983) shows that for rapidly mixing walks, the time to return to the original state has approximately the same distribution as a random walk with independent uniform steps.

If there are $|X|$ states, and W is the first time to return, then for large $|X|$,

$$P\left\{\frac{W}{|X|}>t\right\}\doteq e^{-t}.$$

For 8 red and 8 black balls, $|X| = \binom{16}{8} = 12{,}870$, so the median return time is about 9,000. For 10 red and 10 black balls $|X| = 184{,}756$; the median return time is about 128,000.

A referee points out that the associated Markov chain is doubly stochastic, irreducible, and aperiodic. Standard Markov chain theory shows that the expected number of steps to return to the starting state is $|X|$.

As explained in Diaconis and Shahshahani (1981), the analysis presented for this problem yields all the eigenvalues and eigenvectors of the associated Markov chain. Using these, it is straightforward to derive a closed form expression for the generating function of W as in Flatto, Odlyzko and Wales (1985). This can be used to get sharper asymptotic estimates for Piaget's problem.

2. Group theoretic preliminaries. One natural way to analyze the Bernoulli–Laplace process is by lifting to a random walk on the symmetric group. For integers r and b with $r + b = n$, let S_n be the symmetric group on n letters. Let $S_r \times S_b$ be the subgroup of permutations that permute the first r elements among themselves and the last b elements among themselves. Then $X = S_n/S_r \times S_b$ may be identified with the set of all $\binom{n}{r}$ r-element subsets. The random walk on X moves from x to y by choosing an element in the set x at random and an element of the complement of x at random and switching the two elements to form a new subset y. Choose a metric $d(x, y) = r - |x \cap y|$ on X. The walk is thus a nearest neighbor random walk on X. It is easy to see that the Bernoulli–Laplace process corresponds to the distance process of this walk on subsets.

It is useful to work in more generality: let G be a group and K a subgroup. Let $X = G/K$ be the associated space of cosets. Choose $x_0 = \text{id}, x_1, \cdots, x_m$ as fixed coset representatives, so $G = x_0K \cup x_1K \cdots \cup x_mK$. We will often identify X and $\{x_i\}$. Let Q be a K bi-invariant probability on G, so $Q(k_1gk_2) = Q(g)$ for all $k_1, k_2 \in K$, $g \in G$. The probability Q induces a random walk (more precisely a Markov chain) on X by the following recipe

$$Q(x, y) \overset{d}{=} Q(x^{-1}yK). \tag{2.1}$$

In (2.1), $Q(x, y)$ is the probability of going from x to y in one step. The definition comes from the following considerations: think of x_0 as the origin. Each time choose $g \in G$ from Q and move from x_0 to gx_0. This motion is then translated to motion around x via $y = xgx_0$. Thus, the chance of moving from x to y is $Q(x^{-1}yK)$.

Note that $Q(x, y)$ is well defined and satisfies

$$Q(x, y) = Q(gx, gy) \quad \text{for any} \quad g \in G. \tag{2.2}$$

Philippe Bougerol has pointed out a converse. If $Q(x, y)$ is a Markov chain on $X = G/K$ satisfying (2.2), then Q is induced by a bi-invariant probability defined by

$$Q(A) = Q(x_0, Ax_0) \quad \text{for} \quad A \subset G.$$

Alternatively, write a generic element of G as xk, then $Q(xk) = Q(x_0, x)/|K|$. This measure is K bi-invariant and $Q(x^{-1}yK) = |K|Q(x^{-1}y) = Q(x_0, x^{-1}y) = Q(x, y)$ as required. The following elementary lemma gives further connections between the random walk and Markov chain.

LEMMA 1. *Let $Q(x, y) = Q(gx, gy)$. For any $k \geqq 1$, the k step transition matrix of the Markov chain $Q(x, y)$ is induced by the kth convolution of the associated bi-invariant probability Q. The variation distance to the stationary distribution equals the variation distance to the uniform distribution.*

Because of Lemma 1, Fourier analysis on G can be used to approximate the convolution powers. We briefly review what we need from representation theory. Serre (1977) or Diaconis (1982) contain the details. Recall that a representation of G is a homomorphism $\rho: G \rightarrow GL(V)$ from G into invertible matrices on a vector space V. The dimension d_ρ of ρ is defined as the dimension of V. A representation ρ is irreducible if there are no nontrivial invariant subspaces of V. For Q a probability and ρ a representation, the Fourier transform of Q at ρ is defined by

$$\hat{Q}(\rho) = \Sigma\rho(g)Q(g).$$

The Fourier transform takes convolution into products through $P * Q(\rho) = \hat{Q}(\rho)\hat{P}(\rho)$. The uniform distribution of G: $U(g) = 1/|G|$, has $\hat{U}(\rho) = 0$ for every nontrivial irreducible representation ρ. For $X = G/K$, the set of all complex functions on X is denoted $L(X)$. The group acts on X and so $L(X)$ can be thought of as a representation as well.

The variation distance can be approximated by the following

LEMMA 2. *Let Q be a K bi-invariant probability on a finite group G*

$$\|Q - U\|^2 \leqq \tfrac{1}{4}\Sigma_\rho^* d_\rho \operatorname{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*)$$

where the sum is over all nontrivial representations that occur in the decomposition of $L(X)$.

Proof.

$$\|Q-U\|^2=\tfrac{1}{4}\{\Sigma|Q(g)-U(g)|\}^2\leqq\tfrac{1}{4}|G|\,\Sigma|Q(g)-U(g)|^2$$
$$=\tfrac{1}{4}\Sigma_\rho^* d_\rho \operatorname{Tr}(\hat{Q}(\rho)\hat{Q}(\rho^*)).$$

Here, the Cauchy-Schwarz inequality was used and then the Plancherel theorem as in Serre (1977, p. 49) applied to $Q(g) - U(g)$. Terms corresponding to representations ρ that do not appear in the decomposition of $L(X)$ have zero Fourier transform because of Frobenius reciprocity (Serre (1977, p. 56)): this implies that a representation ρ occurs in $L(X)$ with multiplicity corresponding to the dimension of the space of K fixed vectors of ρ. Thus if ρ does not occur, then the restriction of ρ to K does not contain the trivial representation. Thus $\hat{Q}(\rho) = \Sigma_x Q(x)\rho(x)\Sigma_k \rho(k)$. The inner sum is zero because of the orthogonality of the matrix entries of the irreducible representations (Serre (1977, p. 14)). □

The Fourier transform can simplify a great deal further. Indeed, for the cases treated here the matrix $\hat{Q}(\rho)$ has only one nonzero entry in a suitable basis. The simplification in general is discussed in Volume 6 of Dieudonné (1978).

DEFINITION. The pair (G, K) is called a *Gelfand pair* if each irreducible representation of G appears in $L(X)$ with multiplicity at most 1.

Remarks. Probability theory for bi-invariant probabilities on a Gelfand pair has an extensive literature. Readable overviews appear in Letac (1981), Bougerol (1983), or Dieudonné (1978). The Bernoulli-Laplace model can be treated directly in this framework. However, the more general framework developed here is needed if one is to attack more general problems such as the natural extension to three urns where $L(X)$ has multiplicity.

For a Gelfand pair, let

$$L(X) = V_0 \oplus V_1 \oplus \cdots \oplus V_\lambda \tag{2.3}$$

be the decomposition into distinct irreducibles. Frobenius reciprocity implies that each V_j has a one-dimensional subspace of K invariant functions. Let $s_j(x)$ be a K invariant function in V_j normed so that $s_j(x_0) = 1$. This is called the jth spherical function. The spherical functions have been explicitly computed for many Gelfand pairs.

LEMMA 3. *If (G, K) is a Gelfand pair and Q is a K bi-invariant probability, then*

$$\|Q-U\|^2\leqq\frac{1}{4}\sum_{j=0}^{\lambda} d_j|\hat{Q}(j)|^2,$$

where the sum is over the nontrivial irreducible representations occurring in (2.3) *and*

$$\hat{Q}(j)=\Sigma_g Q(g)s_j(g).$$

Proof. Fix i, and consider the vector space V_i of (2.3) as a representation ρ_i of G. Complete s_i to a basis for V_i, taking s_i as the first basis vector. With respect to this basis, the Fourier transform of any K bi-invariant function f on G becomes

$$\hat{f}(\rho)=\Sigma_g f(g)\rho_i(g)=\Sigma_x f(x)\rho_i(x)\Sigma_k \rho_i(k).$$

But the Schur orthogonality relations imply

$$\Sigma_k \rho_{ij}(k)=\begin{cases}|K| & \text{if } i=j=1,\\ 0 & \text{otherwise.}\end{cases}$$

Thus

$$\hat{f}(\rho)=\begin{pmatrix} a & & 0 \\ 0 & 0 & \\ 0 & & 0 \end{pmatrix} \quad \text{with } a=|K|\Sigma f(x)s_\rho(x).$$

Since the trace norm is invariant under unitary changes of basis, the result follows from Lemma 2. □

Remark. In the description above, the random walk associated to a bi-invariant probability Q can be represented somewhat curiously as a *right* action of G on X. Thus if the basic random elements chosen from Q on G are $g_1, g_2, g_3, \cdots$, and the walk starts at x, the successive steps have the representation $x, xg_1, xg_1g_2, xg_1g_2g_3, \cdots$, where xg denotes the coset containing xg. This is well defined for Q bi-invariant.

There is another natural way to associate a Markov chain to a probability Q, using the natural left action: $x, g_1x, g_2g_1x, g_3g_2g_1x, \cdots$. This process does not correspond to a nearest neighbor walk. It yields a Markov chain with transition matrix $\tilde{Q}(x, y) = Q(yKx^{-1})$. Chains defined in this way satisfy $\tilde{Q}(k_1x, k_2y) = \tilde{Q}(x, y)$ but we do not know a necessary and sufficient condition for a chain to lift to the left action of a bi-invariant Q. Of course, if a lifting can be found, the Fourier analysis is precisely as above, all the bounds and lemmas holding without essential change.

3. The upper bound. Let r and $n - r$ be positive integers. The stationary distribution for the Bernoulli-Laplace model based on two urns, one containing r red balls initially, the second containing $n - r$ black balls may be described as the distribution of the number of red balls in a random sample of size r from the total population of n

$$\pi_r^n(j)=\binom{r}{j}\binom{n-r}{r-j}\Big/\binom{n}{r}, \qquad 0 \leqq j \leqq r.$$

Let P_k be the probability distribution of the number of red balls in the urn containing r balls after k switches have been made.

THEOREM 2. *If*

$$k=\frac{r}{2}\left(1-\frac{r}{n}\right)[\log n+c] \quad \textit{for } c \geqq 0,$$

then, for a universal constant a,

$$\|P_k-\pi_r^n\| \leqq ae^{-c/2}.$$

Proof. Without loss, take $r \leqq n/2$. The decomposition of the space $L(X)$ is a standard result in the representation theory of the symmetric group. James (1978, p. 52) proves that $L(X) = V_0 \oplus V_1 \oplus \cdots \oplus V_r$ where V_i are distinct irreducible representations of the symmetric group corresponding to the partition $(n - i, i)$. In particular, the pair S_n, $S_r \times S_{n-r}$ is a Gelfand pair. Since this result holds for all $r \leqq n/2$, induction gives $\dim(V_i) = \binom{n}{i} - \binom{n}{i-1}$.

The spherical functions have essentially been determined by Karlin and McGregor (1961). Stanton (1984) contains this result in modern language. The spherical functions turn out to be classically studied orthogonal functions called the dual Hahn polynomials. As a function on X, the function $s_i(x)$ only depends on the distance $d(x, x_0)$ and is a polynomial in d given by

$$s_i(d)=\sum_{m=0}^{i}\frac{(-i)_m(i-n-1)_m(-d)_m}{(r-n)_m(-r)_m m!},\qquad 0\leqq i\leqq r, \tag{3.1}$$

where $(j)_m=j(j+1)\cdots(j+m-1)$. Thus

$$\begin{aligned} &s_0(d)=1, \qquad s_1(d)=1-\frac{dn}{r(n-r)},\\ &s_2(d)=1-\frac{2d(n-1)}{r(n-r)}+\frac{(n-1)(n-2)d(d-1)}{(n-r)(n-r-1)r(r-1)}. \end{aligned} \tag{3.2}$$

The basic probability Q for this problem may be regarded as the uniform distribution on the $r(n-r)$ sets of distance one from the set $\{1, \cdots, r\}$. Thus, the Fourier transform of Q at the ith spherical function is

$$\hat{Q}(i)=s_i(1)=1-\frac{i(n-i+1)}{r(n-r)},\qquad 0\leqq i\leqq r.$$

Using this information in the upper bound lemma (Lemma 3) yields

$$\|P_k-\pi_r^n\|^2\leqq\frac{1}{4}\sum_{i=1}^{r}\left\{\binom{n}{i}-\binom{n}{i-1}\right\}\left(1-\frac{i(n-i+1)}{r(n-r)}\right)^{2k}. \tag{3.3}$$

To bound the sum, consider first the term for $i=1$

$$(n-1)\left(1-\frac{n}{r(n-r)}\right)^{2k}.$$

This is smaller than

$$\exp\left(-\frac{2kn}{r(n-r)}+\log n\right).$$

Thus k must be at least

$$\frac{r}{2}\left(1-\frac{r}{n}\right)[\log n+c]$$

to drive this term to zero. With k of this form, the problem is reduced to bounding

$$\sum_{i=1}^{r} e^{a(i)+b(i)}$$

where

$$a(i)=ci\left(\frac{(i-1)}{n}-1\right),\qquad b(i)=\frac{i(i-1)\log n}{n}-\log(i!).$$

Calculus shows that $a(i)\leqq a(1)=-c$ for all $i=[2, n/2]$. Thus, to prove Theorem 1 it suffices to prove

$$\sum_{i=1}^{n/2} e^{b(i)}\leqq B \quad \text{independently of } n.$$

Clearly, the sum of $e^{b(i)}$ over $1\leqq i\leqq 21$ is uniformly bounded. For the remaining range, upper bound $b(i)$ by $i^2(\log n/n)-i\log i+i$. It will be argued that

$i^2(\log n/n) - i\log i + i < -i$ for $21 \leqq i \leqq n/2$, equivalently, $\log n/n < (\log i - 2)/i$. Now if $f(x) = (\log x - 2)/x$, $f'(x) = (3 - \log x)/x^2$. This is negative for $x > e^3$, so for $i > 21 > e^3$,

$$\frac{\log i - 2}{i} > \frac{\log(n/2) - 2}{n/2}.$$

This last is greater than $\log n/n$ for $n \geqq e^{4+2\log 2}$. Thus,

$$\sum_{i=22}^{n/2} e^{b(i)} \leqq \sum_{i=22}^{\infty} e^{-i} < B \quad \text{uniformly in } n. \qquad \square$$

Remarks. Change n to $2n$ and take $r = n$. The result becomes $(n/4)\log n + (c/2)n$ which gives (1.3) of Theorem 1. If $r = o(n)$, the result becomes $(r/2)\log n + (c/2)r$. As usual with approximations, some precision has been lost to get a clean statement. When $r = 1$, for example, there is only one term: $(n-1)(1/(n-1))^{2k}$. For $k = 1$ this gives $\frac{1}{2}(1/\sqrt{n-1})$ as an upper bound for the variation distance. Elementary considerations show that for this case the correct answer is $1/n$. Thus, the upper bound lemma gives the right answer for the number of steps required (namely, 1) but an over estimate for the distance.

4. The lower bound. A lower bound for the variation distance will be found by using the easily derived relation

$$\|P - Q\| = \sup_A |P(A) - Q(A)|. \tag{4.1}$$

Any specific set A thus provides a lower bound. Intuitively, if the number of steps k is too small, there will tend to be too many of the original color in the urn. The argument below gives a sharp form of this. For ease of exposition, we only prove the result for $r = n - r$ (for example (1.4)) but the proof works in the general case.

The idea of the proof is to again use spherical functions, but this time as random variables, not transforms. Thus for any Gelfand pair (G, K) with $X = G/K$, consider $s_j\colon X \to \mathbb{R}$ as a random variable. If Z is a point chosen uniformly in X, the orthogonality relations (as in Stanton (1984, eq. (2.9))) give

$$E\{s_j(Z)\} = \delta_{0j}, \qquad \operatorname{Var}(s_j(Z)) = \frac{1}{\dim(V_j)}. \tag{4.2}$$

If Z_k denotes an X valued random variable with distribution P^{*k} for P a bi-invariant probability on X the basic convolution property of spherical functions becomes

$$E\{s_j(Z_k)\} = E\{s_j(Z_1)\}^k. \tag{4.3}$$

On $S_{2n}/S_n \times S_n$ the first three spherical functions, as functions of the distance d are given (from (3.1)) as

$$s_0(d) = 1, \qquad s_1(d) = 1 - \frac{2d}{n},$$

$$s_2(d) = 1 - \frac{2(2n-1)d}{n^2} + \frac{(2n-1)(2n-2)d(d-1)}{[n(n-1)]^2}.$$

Since these are polynomials in d, it follows that for some a, b, c, $s_1^2 = a + bs_1 + cs_2$. After a computation

$$s_1^2 = \frac{1}{2n-1} + \frac{2n-2}{2n-1} s_2. \tag{4.4}$$

Remark. When working with general r, $n-r$ values the term s_1 appears in the expression for s_1^2.

To lower bound the variation distance, consider the normalized spherical function $f(x) \overset{d}{=} \sqrt{n-1}\, s_1(x)$. Now (4.2) implies for Z uniform on x,

$$E\{f(Z)\} = 0, \qquad \mathrm{Var}\,\{f(Z)\} = 1.$$

Under the convolution measure

$$E\{f(Z_k)\} = \sqrt{n-1}\left(1-\frac{2}{n}\right)^k, \tag{4.5}$$

$$\mathrm{Var}\{f(Z_k)\} = \frac{n-1}{2n-1} + \frac{(n-1)(2n-2)}{(2n-1)}\left(1-\frac{2(2n-1)}{n^2}\right)^k - (n-1)\left(1-\frac{2}{n}\right)^{2k}. \tag{4.6}$$

For k of the form $\frac{1}{4}n\log n - cn$, the mean becomes

$$E\{f(Z_k)\} = \exp\left(2c + O\left(\frac{\log n}{n}\right) + O\left(\frac{c}{n}\right)\right), \tag{4.7}$$

where $c > 0$, and all error terms are uniform in both n and c. Thus, for c large, the mean is large. Similarly

$$\begin{aligned}\mathrm{Var}\{f(Z_k)\} &= \frac{1}{2} + O\left(\frac{1}{n}\right) + \exp\left(4c + O\left(\frac{\log n}{n}\right) + O\left(\frac{c}{n}\right)\right)\\ &\quad - \exp\left(4c + O\left(\frac{\log n}{n}\right) + O\left(\frac{c}{n}\right)\right)\\ &= \frac{1}{2} + e^{4c}\left\{O\left(\frac{\log n}{n}\right) + O\left(\frac{c}{n}\right)\right\}.\end{aligned}$$

Thus, the variance is uniformly bounded for $O \leqq c \leqq \frac{1}{4}\log n$. Now use Chebyshev's inequality: if $A_\alpha = \{x\colon |f(x)| \leqq \alpha\}$, $\pi_n(A_\alpha) \geqq 1 - 1/\alpha^2$ while $P_k(A_\alpha) < B/(e^{2c} - \alpha)^2$ where B is uniformly bounded for $c \leqq \frac{1}{4}\log n$. Thus, for any fixed α and c, for all sufficiently large n,

$$\|P_k - \pi_n\| \geqq 1 - \frac{1}{\alpha^2} - \frac{B}{(e^{2c}-\alpha)^2}.$$

This completes the proof of (1.4), choosing $\alpha = e^{2c}/2$, for example. □

Remark. From the definition of s_1, the set A_α can be interpreted as the event

$$\left|\#\,\mathrm{reds} - \frac{n}{2}\right| \Big/ \sqrt{n} \geqq \alpha.$$

5. Other nearest neighbor walks. A class of problems that can be treated by following the steps above involves a connected graph with vertex set X and an edge set E. Define a metric on X as

$$d(x, y) = \text{length of shortest path from } x \text{ to } y.$$

We want to analyze nearest neighbor walks on this graph. An automorphism of the graph is a 1–1 mapping from X to X which preserves the edge set. Let G be the group of automorphisms of X. Call the graph *2-point homogeneous* if $d(x, y) = d(x', y')$ implies there is a $g \in G$ such that $gx = x'$, $gy = y'$. Taking $x = y$, $x' = y'$ shows G operates transitively on X, so $X \cong G/K$ where $K = \{g \in G: gx_0 = x_0\}$ for some fixed point x_0. Stanton (1984) shows

THEOREM. *For a 2-point homogeneous graph, (G, K) form a Gelfand pair and the spherical functions are orthogonal polynomials.*

This means that in principle the analysis above can be carried out for such examples. Here is a list of some of the examples in Stanton:

Example 1. $X = Z_2^n$, $d(x, y)$ = number of coordinates where x and y differ. Here the random walk becomes nearest neighbor walk on the n cube. This is a well studied problem equivalent to the well-known model of diffusion known as the Ehrenfest urn model. A wonderful discussion of this model is in Kac (1945). Further references are in Letac and Takacs (1979). The straightforward random walk never converges because of parity—after an even number of steps the walk is at a point at an even distance from 0. One simple way to get convergence is to stay fixed with probability $1/(n+1)$ and move to a vertex 1 away with probability $1/(n+1)$. For this process, the analysis can be carried out just as in §§3 and 4 to show $\frac{1}{4}n \log n + cn$ steps suffice and that this many steps are needed.

THEOREM 3. *Let $X = Z_2^n$. Let $P(00 \cdots 0) = P(10 \cdots 0) = \cdots = P(00 \cdots 1) = 1/(n+1)$. Let U be the uniform distribution on X. Suppose $k = \frac{1}{4}(n+1)\log n + c(n+1)$ for $c > 0$. Then*

$$\|P^{*k} - U\|^2 \leqq \tfrac{1}{2}(e^{e^{-4c}} - 1).$$

Conversely, for $k = \frac{1}{4}(n+1)\log n - c(n+1)$, for $c > 0$, the variation distance does not tend to zero as n tends to infinity: $\underline{\lim}\ \|P^{*k} - U\| \geqq (1 - 8e^{-c})$.

Remark. It is curious that the critical rate is precisely the same $\frac{1}{4}n \log n$, for the cube and n sets of a $2n$ set.

Example 2. Let F_q be a finite field with q elements. Let V be a vector space of dimension n over F_q. Let X be the set of k-dimensional subspaces of V, with metric $d(x, y) = k - \dim(x \cap y)$. Here, $G = GL_n(q)$ operates transitively on X. Stanton (1984) gives all the ingredients needed to carry out the analysis.

Example 3. Let X be the set of $(n-r) \times r$ matrices over F_q with metric $d(x, y) = \operatorname{rank}(x - y)$. Here, $GL_{n-r} \times GL_r$ operates transitively on X. Again a complete analysis seems in reach using results given by Stanton.

Example 4. For q odd, let X be the set of skew-symmetric matrices over F_{q^2} with metric $\frac{1}{2}\operatorname{rank}(x - y)$. Here $G = GL_n$ acts on X by $x \to A^T x A$. Again Stanton gives enough information about spherical functions and dimensions to allow a complete analysis.

Stanton also gives results for orthogonal, hermitian, and symplectic matrices over finite fields. He also gives results for a variety of less familiar combinatorial objects. Combinatorialists have also studied such objects: see Biggs (1974, Chaps. 20, 21). Further surveys and examples of Gelfand pairs are given by Heyer (1983) and Sloane (1982).

Finally, there are Gelfand pairs that do not arise from two point homogeneous graphs. An example is the set X of all partitions of $\{1, 2, \cdots, 2n\}$ into n two-element subsets. In graph theoretic language this is the set of all "matchings" of a $2n$ set. The symmetric group S_{2n} acts transitively on X and yields a Gelfand pair. A natural random walk involves picking two elements at random and switching them to form a new partition. This gives an algorithm that converges to a random matching. The spherical functions are "zonal polynomials." It can be shown that $\frac{1}{2}n \log n$ switches suffice.

REFERENCES

D. ALDOUS (1983), *On the time taken by random walks on finite groups to visit every state*, Z. Wahrsch. Verw. Geb., 62, pp. 361–374.

C. ARAGON, D. JOHNSON, L. MEGEOCH AND C. SCHEVON (1984), *Optimization by simulated annealing: An experimental evaluation*, Technical memorandum, Bell Laboratories, Murray Hill, NJ.

N. BIGGS (1974), *Algebraic Graph Theory*, Cambridge Univ. Press, Cambridge.

P. BOUGEROL (1983), *Un mini-cours sur les couples de Gelfand*, Publication Laboratoire de Statistique et Probabilities, Université Paul Sabatier, Toulouse, France, No. 1-83.

P. DIACONIS (1982), *Lecture notes on the use of group representations in probability and statistics*, IMS Lecture Note Series, to appear.

P. DIACONIS AND M. SHAHSHAHANI (1981), *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Geb., 57, pp. 159–179.

J. DIEUDONNÉ (1978), *Treatise on Analysis* VI, Academic Press, New York.

W.J. EWENS (1979), *Mathematical Population Genetics*, Springer-Verlag, Berlin.

W. FELLER, (1968), *An Introduction to Probability and Its Applications*, Vol. 1, 3rd ed., John Wiley, New York.

L. FLATTO, A. ODLYZKO AND D. WALES (1985), *Random shuffles and group representations*, Ann. Prob., 13, pp. 154–178.

H. HEYER (1983), *Convolution semi-groups of probability measures on Gelfand pairs.*, Expos. Math., 1, pp. 3–45.

G.D. JAMES (1978), *Representation of the Symmetric Groups*, Springer Lecture Notes in Mathematics 682, Springer-Verlag, Berlin.

N.L. JOHNSON AND S. KOTZ (1977), *Urn Models and Their Application*, John Wiley, New York.

M. KAC (1945), *Random walk and the theory of Brownian motion*, Amer. Math. Monthly, 54, pp. 369–391.

S. KARLIN AND J. MCGREGOR (1961), *The Hahn polynomials, formulas and an application*, Scripta Math., 26, pp. 33–46.

——— (1975). *Linear growth models with many types and multidimensional Hahn polynomials*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, pp. 261–288.

S. KIRKPATRICK, C. GELATT AND M. VECCHI (1983), *Optimization by simulated annealing*, Science, 220, pp. 671–680.

G. LETAC (1981), *Problèmes Classiques de Probabilite sur un Couple de Gelfand*, in Springer Lecture Notes in Mathematics 861, Springer-Verlag, Berlin.

G. LETAC AND L. TAKACS (1979), *Random walks on the m-dimensional cube*, J. Reine Angew. Math., 310, pp. 187–195.

P.A.P. MORAN (1958), *Random Processes in Genetics*, Proc. Cambridge Phil. Soc., 54, pp. 60–72.

J. PIAGET AND B. INHELDER (1975), *The Origin of the Idea of Chance in Children*, Norton, New York.

J.P. SERRE (1977), *Linear Representations of Finite Groups*, Springer-Verlag, New York.

N.J.A. SLOANE (1982), *Recent bounds for codes, sphere packings, and related problems obtained by linear programming and other methods*, Contemp. Math, 9, pp. 153–185.

D. STANTON (1984), *Orthogonal polynomials and Chevalley groups*, in Special Functions: Group Theoretical Aspects and Applications, R. Askey, T.H. Koornwinder and W. Schempp, eds., Reidel, Dordrecht, pp. 87–128.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

REAL SINGULARITIES OF SINGULAR STURM–LIOUVILLE EXPANSIONS*

GILBERT G. WALTER[†] AND AHMED J. ZAYED[‡]

Abstract. Elliptic equations in polar coordinates lead to singular Sturm–Liouville problems on $(0, \infty)$ with equations of the form $y''+(\lambda-q)y=0$. Let $\phi(x,\lambda)$ be solutions to the problem satisfying the condition $\phi(0,\lambda)=\sin\beta$, $\phi'(0,\lambda)=-\cos\beta$. The associated generalized Fourier transform $F(\lambda)=\int_0^\infty f(x)\phi(x,\lambda)\,dx$ is extended to cases where $F(\lambda)$ is a function of polynomial growth. This enables us to study the location of singularities of the analytic representation $\hat{f}$ of the generalized function $f(x)=\int_{-\infty}^\infty F(\lambda)\phi(x,\lambda)\,d\rho(\lambda)$. We do so by comparing them to the location of the singularities of the analytic representation $\hat{g}$ of the tempered distribution g which is the Fourier transform of $F(s^2)$.

Key words. generalized functions, analytic representations, Fourier transforms

AMS(MOS) subject classifications. Primary 34B25; secondary 35B50

1. Introduction. In this paper we extend the results of [7] on singularities of Sturm–Liouville expansions to include cases in which the location of the singular points lies on the real axis. This includes many cases of interest in which the function to be expanded is not holomorphic on all of $(0, \infty)$.

Indeed most convergence theorems for Sturm–Liouville expansions involve only real derivatives of the functions to be expanded, or as in the case of L^1 or L^2, no derivatives at all. Nonetheless they may be locally holomorphic and it is of interest in some applications to find the location of those real singular points [4].

In order to develop a theory encompassing such points we first introduce a space of generalized functions which will include most ordinary functions of interest. Each element in this space will have a convergent Sturm–Liouville expansion and an analytic representation obtained from a dual system of eigenfunctions. The singular points of these analytic representations are then compared to those of an associated Laplace transform of a tempered distribution [1]. The results obtained are similar to those in [7] whose terminology and results we shall also use.

A similar theory had been developed in [6] for regular Sturm–Liouville series. This in turn was an extension of the results found in [3] to the case of real singularities.

2. Preliminaries. We consider singular Sturm–Liouville problems of the form

$$y''+(\lambda-q(x))y=0, \tag{2.1}$$

with boundary conditions

$$\begin{aligned} &y(0)\cos\alpha+y'(0)\sin\alpha=0,\\ &|y(\infty)|<\infty. \end{aligned} \tag{2.2}$$

*Received by the editors August 7, 1984, and in final revised form August 20, 1985.

[†] Department of Mathematics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

[‡] Department of Mathematics, California Polytechnic State University, San Luis Obispo, California 93407.

Let $\phi(x,\lambda)$ be a solution of (2.1) such that

$$\phi(0,\lambda)=\sin\alpha, \qquad \phi'(0,\lambda)=-\cos\alpha.$$

We shall suppose that $q(x)\in L^1(0,\infty)$ and is holomorphic in the right half plane. Hence the negative portion of the spectrum of the problem (2.1), (2.2) is discrete and bounded below, and can be ignored for most of what we shall do.

We shall be interested in Sturm–Liouville expansions of the form

$$f(x)\sim\int_{-\infty}^{\infty}F(\lambda)\phi(x,\lambda)\,d\rho(\lambda) \tag{2.3}$$

where $F(\lambda)$ is a continuous function of polynomial growth and where $\rho(\lambda)$ is the spectral measure. In this case the integral in (2.3) does not exist in the usual sense but must be interpreted as a generalized inverse Fourier transform of a generalized function. Similarly $f(x)$ will not necessarily be an ordinary function but rather a generalized function belonging to a particular space.

2.1. A space of generalized functions. Let A_0 denote $L^2(0,\infty)$ and let A_1 be the space of those functions $\phi\in L^2(0,\infty)$ whose derivative is absolutely continuous such that

$$\phi''-q\phi\in L^2(0,\infty), \qquad \phi(0)\cos\alpha+\phi'(0)\sin\alpha=0$$

and successively,

$$A_{n+1}=\{\phi\in L^2(0,\infty)|\phi^{(2n+1)} \text{ is absolutely continuous,}$$
$$(D^2-q)\phi\in A_n,\ \phi(0)\cos\alpha+\phi'(0)\sin\alpha=0\}.$$

Then

$$A=\bigcap_{n=0}^{\infty}A_n$$

is a linear space contained in $L^2(0,\infty)$ topologized by the countable family of seminorms given by

$$\rho_k(\phi)=\left\|(D^2-q)^k\phi\right\|_2.$$

PROPOSITION 2.1. *Let ψ be in A and let $\Psi(\lambda)$ be its generalized Fourier transform*

$$\Psi(\lambda)=\underset{N\to\infty}{\text{l.i.m.}}\int_0^N\psi(x)\phi(x,\lambda)\,dx; \tag{2.4}$$

then the eigenfunction expansion of ψ

$$\int_{-\infty}^{\infty}\Psi(\lambda)\phi(x,\lambda)\,d\rho(\lambda) \tag{2.5}$$

converges to ψ in the sense of A. Furthermore

$$\lambda^p\Psi(\lambda)\in L^2(\rho), \qquad p>0.$$

Proof. Let ψ_N be given by

$$\psi_N(x)=\int_{-N}^{N}\Psi(\lambda)\phi(x,\lambda)\,d\rho(\lambda).$$

Then $\psi_N \in L^2(0,\infty)$ ([5, p. 131]) as is $(D^2-q)^k\psi_N$ for each integer $k \geqq 0$, since $\lambda^k\Psi(\lambda)\chi_{[-N,N]}$, (where $\chi_{[-N,N]}$ is the characteristic function of $[-N,N]$) is also in $L^2(\rho)$. By the expansion theorem for Sturm–Liouville problems ([5, p. 130])

$$(D^2-q)^k\psi_N \to (D^2-q)^k\psi \qquad (L^2).$$

Hence $\psi_N \to \psi$ in the sense of A. The second conclusion follows similarly.

We now consider the dual space A' of A. Clearly $L^2(0,\infty) \subset A'$ and A contains $S(0,\infty)$, the space of all C^∞-functions with support in a closed half line in $(0,\infty)$ and which with their derivatives are rapidly decreasing at ∞. Since convergence in S implies convergence in A, S' contains A'. That is, the elements of A' are tempered distributions on $(0,\infty)$. Similarly A' contains E'_+ the space of distributions of compact support on $(0,\infty)$.

Proposition 2.2. *Let $f \in E'_+$ and let $F(\lambda)$ be its generalized Fourier transform*

$$F(\lambda) = \langle f, \phi(\cdot,\lambda)\rangle; \tag{2.6}$$

then the restriction of F to $(0,\infty)$ is a continuous function of polynomial growth.

Proof. Since $f \in A'$, it must belong to the dual space A'_k of one of the A_k's ([2, p. 11]). Hence $f=(D^2-q)^kG$ for some $G \in L^2$ since $\langle f,\varphi\rangle = \langle G,(D^2-q)^kG\rangle$ and $G \in A'_0 = L^2$. Since f has compact support, G satisfies

$$(D^2-q)^kG=0$$

in $(0,a)\cup(b,\infty)$ for some $0<a<b$. Let χ be the characteristic function of $(a/2,\ 2b)$. Then we have

$$\begin{aligned} f &= (D^2-q)^kG = (D^2-q)^k[G\chi] + (D^2-q)^k[G(1-\chi)] \\ &= (D^2-q)^kG_1 + (D^2-q)^{k-1}[((D^2-q)G)(1-\chi) + 2DGD(1-\chi) + GD^2(1-\chi)] \\ &= (D^2-q)^kG_1 + \sum_{i=0}^{2k-1} \left(c_i\delta_{a/2}^{(i)} + d_i\delta_{2b}^{(i)}\right)(-1)^i, \end{aligned}$$

where $G_1 = G\chi$ has compact support and $\delta_\alpha^{(i)}(x) = \delta^{(i)}(x-\alpha)$. Therefore we have

$$\begin{aligned} F(\lambda) &= \langle (D^2-q)^kG_1,\ \phi(\cdot,\lambda)\rangle + \sum_{i=0}^{2k-1} c_i\phi^{(i)}\left(\frac{a}{2},\ \lambda\right) + d_i\phi^{(i)}(2b,\lambda) \\ &= \langle G_1,\ (D^2-q)^k\phi(\cdot,\lambda)\rangle + O(\lambda^{k-1/2}) \\ &= (-\lambda)^k\langle G_1,\ \phi(\cdot,\lambda)\rangle + O(\lambda^{k-1/2}) \end{aligned}$$

by the asymptotic formula for ϕ [5, p. 206]. Thus the conclusion follows.

For f in A' we cannot use (2.6) as the definition of the generalized Fourier transform since $\phi(x,\lambda) \notin A$. However we can use another property which is consistent with Parseval's relation.

Definition 2.1. Let $f=(D^2-q)^kG$ be an element of A' where $G \in L^2(0,\infty)$. Then the generalized Fourier transform F of f is given by

$$F(\lambda) = (-\lambda)^k \operatorname*{l.i.m.}_{M\to\infty} \int_0^M G(x)\phi(x,\lambda)\,dx. \tag{2.7}$$

If $G \in L^1(0,\infty)$ as well, then f is called admissible.

PROPOSITION 2.3. *Let $f \in A'$ and let $F_1(\lambda)$ be the restriction of its generalized Fourier transform to positive values of λ. Then*

(i) *for admissible f $F_1(\lambda)$ is a continuous function of polynomial growth*;

(ii) *the eigenfunction expansion of f converges to f in the sense of A'.*

Proof. Since $\phi(x,\lambda)$ is bounded for $x \in (0,\infty)$, $F(\lambda)$ is at most of polynomial growth by (2.7). The continuity follows from the expression ([5, p. 206])

$$\phi(x,\lambda) = \sin\alpha\cos\sqrt{\lambda}\,x - \cos\alpha\frac{\sin\sqrt{\lambda}\,x}{\sqrt{\lambda}} + \frac{1}{\sqrt{\lambda}}\int_0^x \sin\sqrt{\lambda}\,(x-t)q(t)\phi(t,\lambda)\,dt. \tag{2.8}$$

Now we let $F(\lambda)$ be a continuous function of polynomial growth; then f_N given by

$$f_N(t) = \int_{-N}^{N} F(\lambda)\phi(t,\lambda)\,d\rho(\lambda) = \int_{-\infty}^{\infty} F_N(\lambda)\phi(t,\lambda)\,d\rho(\lambda)$$

belongs to $L^2(0,\infty)$ and hence A'. Furthermore

$$\langle f_N, \phi\rangle = \int_{-\infty}^{\infty} F_N(\lambda)\Phi(\lambda)\,d\rho(\lambda) \to \int_{-\infty}^{\infty} F(\lambda)\Phi(\lambda)\,d\rho(\lambda) \tag{2.9}$$

where the integral converges by Proposition 2.1. But this is just the eigenfunction expansion of f if $F(\lambda)$ is as in (2.7).

2.2. Analytic representations. The analytic representation of a function $h \in L^2(-\infty,\infty)$ is given by

$$\hat{h}(z) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{h(x)}{x-z}\,dx, \qquad \operatorname{Im} z \neq 0. \tag{2.10}$$

This can be extended to S' by using the inverse Fourier transform $\mathscr{F}^{-1}$

$$\hat{h}(z) = \begin{cases} \displaystyle\int_0^{\infty} e^{izs}(\mathscr{F}^{-1}h)(s)\,ds, & \operatorname{Im} z > 0, \\ \displaystyle\int_{-\infty}^{0} e^{izs}(\mathscr{F}^{-1}h)(s)\,ds, & \operatorname{Im} z < 0. \end{cases}$$

(See [1].) In our case we have, for $\lambda = s^2$,

$$\hat{g}(z) = \begin{cases} \displaystyle\int_0^{\infty} e^{izs}F(s^2)\,ds, & \operatorname{Im} > 0, \\ 0, & \operatorname{Im} z < 0. \end{cases} \tag{2.11}$$

This $\hat{g}$ is the analytic representation of a tempered distribution $g \in S'$. It is related to the element $f \in A'$ through its Fourier transform F which is the same as the generalized Fourier transform of f, i.e. $\mathscr{F}(g) = F(s^2)$ and $F(s^2) = \langle f, \phi(\cdot, s^2)\rangle$.

In order to construct an analytic representation of f we use another solution to (2.1), namely

$$\psi^{\pm}(z,\lambda) = \phi(z,\lambda)\int_{\pm i\infty}^{z} \frac{d(\lambda)}{\phi^2(z',\lambda)}\,dz'$$

where $+i\infty$ is used for $\operatorname{Im} z>0$, $-i\infty$ for $\operatorname{Im} z<0$, and $d(\lambda)$ is a constant to be determined. We shall assume that the behavior of $\phi(z,\lambda)$ is similar to its upper bound off the real axis.

Assumption 1.

$$|\phi(z,s^2)|\sim e^{|\operatorname{Im} zs|}, \quad s \text{ real} \quad (\text{see [5]}).$$

Then $|\psi^+(z,s^2)|\leqq Ae^{-s\operatorname{Im} z}$, s real, $\operatorname{Im} z>0$. Since q is assumed holomorphic in the right half plane, $\psi^\pm$ is also holomorphic there. The function $\psi^+(x)-\psi^-(x)$ is also a solution to (2.1) on the real axis. It is related to ϕ by

$$\psi^+(x,\lambda)-\psi^-(x,\lambda)=\phi(x,\lambda)\left\{\int_{i\infty}^{x}\frac{c(\lambda)}{\phi^2(x,z)}\,dz-\int_{-i\infty}^{x}\frac{c(\lambda)}{\phi^2(x,z)}\,dz\right\}$$
$$=\phi(x,\lambda)$$

provided

$$c^{-1}(\lambda)=\int_{+i\infty}^{-i\infty}\frac{1}{\phi^2(z,\lambda)}\,dz.$$

This is independent of the contour since the Wronskian of $\psi^+-\psi^-$ with ϕ is zero. Thus we may take $\psi^+(\psi^-)$ to be the analytic representation of $\phi(x,\lambda)$ in the upper (lower) half plane respectively.

The usual analytic representation of ϕ given by (2.10) exists as well provided the integral is l.i.m.

From this we get an analytic representation for $f\in A'$ given by

$$\hat{f}^{\pm}(z)=\int_{-\infty}^{\infty}F(\lambda)\psi^{\pm}(z,\lambda)\,d\rho(\lambda)$$

which converges for $|\operatorname{Im} z|$ sufficiently large.

3. Associated kernels. It was shown in [7] that the kernels

$$K(t,z)=\int_0^\infty\phi(t,s^2)e^{-isz}\,d\rho(s^2),\qquad \operatorname{Im} z<0,\quad 0<t \tag{3.1a}$$

and

$$L(t,z)=\int_0^\infty\phi(t,s^2)e^{isz}\,ds,\qquad \operatorname{Im} z>0,\quad 0<t \tag{3.1b}$$

are holomorphic in the lower (resp. upper) half plane and may be continued to the real axis except possibly at the point $z=\pm t$. The analytic representations of each, $\hat{K}$ and $\hat{L}$, may be shown to be singular at most at the same values as well. Indeed $K(t,z)\in L^2(0,\infty)$ for $\operatorname{Im} z<0$ as a function of t and

$$\hat{K}(w,z)=\frac{1}{2\pi i}\int_0^\infty\frac{K(t,z)}{t-w}\,dt \tag{3.2}$$

is holomorphic for $\operatorname{Im} z<0$, $\operatorname{Im} w\neq 0$. It may be continued for w fixed in the upper half plane to any value of z except possibly $z=\pm w$ by deforming the contour of integration.

The same is true for $\hat{L}(w,z)$. Hence both are holomorphic in C^2 except possibly on $w=0$ and $z=\pm w$.

Another analytic representation of these kernels is obtained by using $\psi^{\pm}$ in place of ϕ. We have

$$\hat{K}^{\pm}(w,z)=\int_0^{\infty}\psi^{\pm}(w,s^2)e^{-isz}d\rho(s^2)$$

which differs from $\hat{K}$ by a function holomorphic in $\{(w,z)|\operatorname{Re} w>0\}$ and similarly for $\hat{L}^{\pm}(w,z)$.

4. Relation between ***f*** **and** ***g*** **and their analytic representations.** For the same coefficient function $F(\lambda)$ we have on the one hand, formally,

$$f(t)=\int_{-\infty}^{\infty}\phi(t,\lambda)F(\lambda)\,d\rho(\lambda) \tag{4.1}$$

where $f\in A'$, and on the other

$$g(x)=\int_0^{\infty}e^{ixs}F(s^2)\,ds \tag{4.2}$$

where $g\in S'$. Both integrals must be interpreted as the limits in an appropriate sense, (namely A' and S') of the functions obtained by truncating F to its value on bounded intervals. But $F(\lambda)$ may be obtained from f by using the fact that $f=((D^2-q)^2+1)^pG$ where $G\in L^2$ is given by

$$G(t)=\int_{-\infty}^{\infty}\phi(t,\lambda)\frac{F(\lambda)}{(\lambda^2+1)^p}\,d\rho(\lambda).$$

If p is chosen sufficiently large, $F(\lambda)/(\lambda^2+1)^p\in L^2(\rho)$ and hence $G(t)\in L^2$ and the usual inversion theorem holds

$$\frac{F(\lambda)}{(\lambda^2+1)^p}=\underset{N\to\infty}{\text{l.i.m.}}\int_0^N\phi(t,\lambda)G(t)\,dt. \tag{4.3}$$

The analytic representation of $g(x)$ is also given by (4.2) with x replaced by the complex variable z. We now replace F by the expression (4.3)

$$\hat{g}(z)=\int_0^{\infty}e^{isz}(s^4+1)^p\int_0^{\infty}\phi(t,s^2)G(t)\,dt\,ds.$$

For $\operatorname{Im} z>0$, $e^{isz}(s^4+1)^p\in L^2$ and hence this integral exists. It may also be given by

$$\begin{aligned}\hat{g}(z)&=(D_z^4+1)^p\int_0^{\infty}\int_0^{\infty}e^{isz}\phi(t,s^2)\,ds\,G(t)\,dt \\ &=(D_z^4+1)^p\int_0^{\infty}L(t,z)G(t)\,dt \\ &=(D_z^4+1)^p\int_0^{\infty}L(t,z)(\hat{G}(t+i0)-\hat{G}(t-i0))\,dt \\ &=(D_z^4+1)^p\left(\int_{c^+}-\int_{c^-}\right)L(w,z)\hat{G}(w)\,dw, \qquad \operatorname{Im} z>0\end{aligned} \tag{4.4}$$

where $\hat{G}$ is the usual analytic representation of G and c^+, c^- are contours from 0 to ∞ lying respectively in the upper and lower half plane.

To go in the other direction we first use the fact that

$$\hat{f}(z)=\int_0^{\infty}\psi^{\pm}(z,s^2)F(s^2)\,d\rho(s^2). \tag{4.5}$$

Since

$$F(s^2)=\frac{1}{2\pi}\int_{-\infty}^{\infty}g(x)e^{-ixs}\,dx$$

we have by interchanging the order of integration

$$\begin{aligned}\hat{f}(z)&=\frac{1}{2\pi}\int_{-\infty}^{\infty}g(x)\int_0^{\infty}\psi^{\pm}(z,s^2)e^{-ixs}\,d\rho(s^2)\,dx \\ &=\frac{1}{2\pi}\langle g,\,\hat{K}^{\pm}(z,\cdot)\rangle \\ &=\frac{1}{2\pi}\int_{c^+}\hat{K}^{\pm}(z,w)\hat{g}(w)\,dw\end{aligned} \tag{4.6}$$

where c^+ is a contour in the upper half plane.

5. The singularity theorem. The main result can now be given.

THEOREM. *Let $f\in A'$ and admissible, let $F(\lambda)$ be the generalized Fourier transform of f; then the analytic representation of f is singular at $t=\alpha>0$ if and only if $\hat{g}(z)$ given by*

$$\hat{g}(z)=\int_0^{\infty}e^{isz}F(s^2)\,ds$$

is singular at α or $-\alpha$.

Proof. For $f\in A'$ we have

$$f(t)=\underset{N\to\infty}{\text{l.i.m.}}\int_{-N}^{N}F(\lambda)\phi(t,\lambda)\,d\rho(\lambda)$$

where the limit is in the sense of A'. Since the negative part of the spectrum is discrete and bounded below, we may express f as

$$\begin{aligned}f(t)&=\int_{-\infty}^{0}F(\lambda)\phi(t,\lambda)\,d\rho(\lambda)+\int_0^{\infty}F(\lambda)\phi(t,\lambda)\,d\rho(\lambda) \\ &=m(t)+f_1(t)\end{aligned} \tag{5.1}$$

where m is holomorphic for $\operatorname{Re}t>0$ and $f_1\in A'$.

Clearly $\hat{f}(z)$ and $\hat{f}_1(z)$ the analytic representations of f and f_1 respectively, are singular at the same points in $\operatorname{Re}z>0$.

Let us suppose that $\hat{f}$ has an isolated singularity at $z=\alpha$ on the positive real axis. Then either $\hat{G}(\alpha+i0)$ or $\hat{G}(\alpha-i0)$ of (4.4) is singular there as well, say $\hat{G}(\alpha+i0)$. Then

$$\int_{c^+}L(w,z)\hat{G}(w)\,dw \tag{5.2}$$

is holomorphic for $\operatorname{Im}z>0$. It may be continued to Real z for all values except the singularities of $\hat{G}$ or their negatives (since $L(w,z)$ has singularities at most at $z=\pm w$) by deforming the contour c^+. The same is true for c^-. Hence the only possible singular point of $\hat{g}(z)$ is at $z=\pm\alpha$. Since $\hat{g}(z)=0$ for $\operatorname{Im}z<0$, no possible singularities arise there.

Now suppose $\hat{g}(w)$ has an isolated singularity at $\beta\neq 0$. Then by (4.6) for $\operatorname{Im} z<0$, $\hat{f}(z)$ may be continued to the entire real axis since $\hat{K}^{-}(z,w)$ has singularities only at $w=\pm z$, and the contour lies above the x-axis. For $\operatorname{Im} z>0$, $\hat{f}(z)$ may be continued to the real axis by again deforming the contour except at whichever of β or $-\beta$ is positive.

Hence if $\hat{f}$ is singular at $t=\alpha$, $\hat{g}$ must be singular at least at $-\alpha$ and α since if it were not, $\hat{f}$ would not be singular at $t=\alpha$

Example 1. $q(x)=0$.

a) $\alpha=\pi/2$, $\phi(x,s)=\cos xs$, $d\rho(\lambda)=ds$.

For $F(s^2)=e^{ias}s^{\nu-1}$, $a,s>0$, $\operatorname{Re}\nu>0$, we have $f\in A'$ given by

$$f(x)=\int_0^\infty e^{ias}s^{\nu-1}\cos xs\,ds$$

where the integral converges in the sense of A'. The integral converges locally as well to the function

$$f(x)=\frac{\Gamma(\nu)}{2(-1)^\nu}\{(x-a)^{-\nu}+(x+a)^{-\nu}\}.$$

The analytic representation $\hat{f}$ of f has a singularity on $[0,\infty)$ at most at $x=a$. Since $\psi^{\pm}(z,s)$ can be calculated explicitly in this case,

$$\psi^{\pm}(z,s)=\pm\frac{e^{\pm isz}}{2},$$

we can also find $\hat{f}$ explicitly:

$$\hat{f}^{\pm}(z)=\pm\frac{\Gamma(\nu)}{2(-1)^\nu}(a\pm z)^{-\nu}$$

with + or − corresponding to $\operatorname{Im} z>0$ or $\operatorname{Im} z<0$. The corresponding $\hat{g}$ is

$$\hat{g}(z)=\int_0^\infty e^{ias}s^{\nu-1}e^{isz}\,ds=\begin{cases}\dfrac{\Gamma(\nu)}{(-1)^\nu}(z+a)^{-\nu}, & \operatorname{Im} z>0,\\ 0, & \operatorname{Im} z<0,\end{cases}$$

which agrees with the theorem of §5.

b) $0\neq\alpha\neq\pi/2$, $\phi(x,s)=\sin\alpha\cos xs-(\cos\alpha/s)\sin xs$, $d\rho(\lambda)=2s\rho'(s)\,ds=(1/\pi)\{2s^2/(\cos^2\alpha+s^2\sin^2\alpha)\}\,ds$.

For $F(s^2)=\pi(\cos^2\alpha+s^2\sin^2\alpha)s^\mu$, $\operatorname{Re}\mu>-2$, we have $f\in A'$ given locally by

$$\begin{aligned}f(x)&=\frac{1}{\pi}\int_0^\infty F(s^2)\phi(x,s)2s\rho'(s)\,ds\\&=2\int_0^\infty \sin\alpha\cos xs\,s^{\mu+2}\,ds\\&\quad-2\cos\alpha\int_0^\infty \sin xs\,s^{\mu+1}\,ds\\&=2\sin\alpha\Gamma(\nu+3)\cos\frac{((\nu+3)\pi)}{2}x^{-\mu-3}\\&\quad-2\cos\alpha\Gamma(\nu+2)\sin\frac{((\nu+2)\pi)}{2}x^{-\mu-2}.\end{aligned}$$

Thus $\hat{f}^{\pm}(z)$ has singularities at most $z=0$. Similarly $\hat{g}(z)$ can be found to be

$$\hat{g}(z)=\cos^2\alpha\frac{\Gamma(\mu+1)}{(-1)^{\mu+1}}z^{-\mu-1}+\sin^2\alpha\frac{\Gamma(\mu+3)}{(-1)^{\mu+3}}z^{-\mu-3}$$

for $\operatorname{Im} z>0$.

Example 2. $q(x)=(\nu^2-1/4)/x^2$, $\nu>1$, $\phi(x,s)=\sqrt{x}\,J_\nu(xs)$, $d\rho(\lambda)=s\,ds$.

a) $F(s^2)=s^{\mu-2}$, $\operatorname{Re}(\mu+\nu)-1>0$, $\mu>1$.

Then $f\in A'$ is given locally by

$$f(x)=\int_0^\infty s^{\mu-1}\sqrt{x}\,J_\nu(xs)\,ds=\frac{\Gamma((\mu+\nu)/2)2^{\mu-1}}{\Gamma((\nu-\mu)/2+1)x^{\mu-1/2}},$$

while

$$\hat{g}(z)=\begin{cases}\dfrac{\Gamma(\mu-1)}{(-1)^{\mu-1}}z^{1-\mu}, & \operatorname{Im} z>0,\\ 0, & \operatorname{Im} z<0.\end{cases}$$

b) $F(s^2)=s^{\nu-1}\sin as$.

Then

$$f(x)=\int_0^\infty s^{\nu-1/2}\sin as\sqrt{sx}\,J_\nu(sx)\,ds$$

$$=\frac{\sqrt{\pi}\,2^\nu x^{\nu+1/2}}{\Gamma(1/2\nu)(a^2-x^2)^{\nu+1/2}}\chi_{[0,a)}(x)$$

where $\chi_{[0,a)}$ is the characteristic function of $[0,a)$. Now

$$\hat{g}(z)=\begin{cases}\dfrac{|z+a|^{-\nu}-|z-a|^{-\nu}\operatorname{sgn}(z-a)}{4\Gamma(1-\nu)\cos(\nu\pi/2)}+i\dfrac{|z-a|^{-\nu}|z+a|^{-\nu}}{4\Gamma(1-\nu)\sin(\nu\pi/2)}, & \operatorname{Im} z>0,\\ 0, & \operatorname{Im} z<0.\end{cases}$$

REFERENCES

[1] H. Bremermann, *Distributions, Complex Variables and Fourier Transforms*, Addison-Wesley, Reading, MA, 1965.

[2] A. Friedman, *Generalized Functions and Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[3] R. Gilbert and H. Howard, *On the singularities of Sturm-Liouville expansions*, Proc. Symposium on Analytic Methods in Mathematical Physics, Indiana Univ. Press, Bloomington, IN, 1969.

[4] R. Gilbert and S. Shieh, *A new method in the theory of potential scattering*, J. Math. Phys., 7 (1966), pp. 431–433.

[5] B. Levitan and I. Sargsjan, *Introduction to spectral theory*, Math. Monographs Vol. 39, American Mathematical Society, Providence, RI, 1975.

[6] G. Walter, *Singular points of Sturm-Liouville series*, this Journal, 2 (1971), pp. 393–401.

[7] A. Zayed and G. Walter, *On the singularities of a singular Sturm-Liouville expansion and an associated class of Elliptic* P.D.E.'*s*, this Journal, 16 (1985), pp. 725–740.

[8] A. Zemanian, *Generalized Integral Transforms*, John Wiley, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

018

q-POLLACZEK POLYNOMIALS AND A CONJECTURE OF ANDREWS AND ASKEY*

W. A. AL-SALAM[†] AND T. S. CHIHARA[‡]

Abstract. The q-Pollaczek polynomials are orthogonal polynomials having a generating function of the form $A(t)\prod_{k=0}^{\infty}F(x,tq^k)=\sum_{n=0}^{\infty}P_n(x)t^n$, where $F(x,t)=[1-xH(t)]/[1-xK(t)]$ and $A(t)$, $H(t)$ and $K(t)$ are formal power series with $H(0)=K(0)=0$, $A(0)H'(0)K'(0)\neq 0$. We determine all orthogonal polynomials having generating functions of this form. We find that in addition to the q-Pollaczek polynomials, there are two other sets that are closely related to the q-Pollaczek polynomials.

Key words. q-Pollaczek polynomials, orthogonal polynomials

AMS(MOS) subject classification. Primary 33A65

1. Introduction. In the characterization of a class of orthogonal polynomials having a certain "convolution structure" [2], we were led to generating functions of the form

$$\prod_{k=0}^{\infty}\frac{1-atq^k+bt^2q^{2k}}{1-xtq^k+ct^2q^{2k}}=\sum_{n=0}^{\infty}P_n(x)t^n \tag{1.1}$$

in which the $P_n(x)$ are orthogonal polynomials. These polynomials are special cases of certain orthogonal ${}_4\phi_3$ polynomials ("q-Askey–Wilson polynomials") studied by Askey and Wilson and the orthogonality relations for $|q|<1$ can be obtained from those given in [6]. These relations are rederived by Askey and Ismail in [5] and the orthogonality relations are also given for $|q|>1$ in those cases where the associated Hamburger moment problem is determined. Andrews and Askey [4] have recently conjectured that the only generating functions of the form

$$A(t)\prod_{k=0}^{\infty}\left[1-xK(tq^k)\right]^{-1}=\sum_{n=0}^{\infty}P_n(x)t^n \tag{1.2}$$

in which $\{P_n(x)\}$ is a sequence of orthogonal polynomials are (after change of variables and renormalization) the ones we found in [2].

Independently, in a private communication, Askey suggested to one of us that it would be interesting to characterize those orthogonal polynomials that have generating functions of the form

$$A(t)\prod_{k=0}^{\infty}\left[1-xH(tq^k)\right]=\sum_{n=0}^{\infty}P_n(x)t^n. \tag{1.3}$$

* Received by the editors July 29, 1985; accepted for publication October 1, 1985.

[†] Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A 2975.

[‡] Department of Mathematics, Purdue University–Calumet, Hammond, Indiana 46323. The research of this author was supported by the National Science Foundation under grant DMS 8400701.

Al-Salam was able to find all orthogonal polynomials of the latter class and this suggested to him the further characterization problem of finding all orthogonal polynomials having generating functions of the form

$$(1.4)\qquad A(t)\prod_{m=0}^{\infty}\frac{1-\delta xH(q^m t)}{1-\theta xK(q^m t)}=\sum_{n=0}^{\infty}P_n(x)t^n.$$

When apprised of the latter, Askey noted that the q-Pollaczek polynomials have a generating function of this type. He pointed out that the q-Pollaczek polynomials can be defined (after a scale change of the independent variable: $x\to x/2$) by the recurrence relation

$$(1.5)\qquad (1-q^{n+1})P_{n+1}(x)=[(1-aq^n)x+bq^n]P_n(x)-(1-cq^{n-1})P_{n-1}(x),$$

$P_n(x)=P_n(x;a,b,c)$. They have the generating function

$$(1.6)\qquad F(x;q;a,b,c;t)=\prod_{k=0}^{\infty}\frac{1-(ax+b)tq^k+ct^2q^{2k}}{1-xtq^k+t^2q^{2k}}=\sum_{n=0}^{\infty}P_n(x)t^n.$$

This can be written in the form (1.4) with

$$A(t)=\prod_{k=0}^{\infty}\frac{1-btq^k+ct^2q^{2k}}{1+t^2q^{2k}},$$
$$H(t)=(1-bt+ct^2)^{-1},$$
$$K(t)=(1+t^2)^{-1}.$$

Askey suggested that these were essentially the only orthogonal polynomials having such generating functions. These polynomials were initially studied by Askey and Ismail and the orthogonality measures have now been found in most cases. The special case of the symmetric polynomials ($b=0$) is discussed thoroughly for $0<q<1$ by Askey and Ismail in their memoir [5]. For the general case, weight functions have been obtained for all cases where the spectrum is continuous by Charris and Ismail [8]. These authors also discuss a sieved version of the Pollaczek polynomials which can be obtained from the q-Pollaczek polynomials by letting $q\to$ a root of unity.

In this paper, we will investigate Askey's suggestion by looking for all cases where (1.4) generates orthogonal polynomials. We will show that the initial conjecture of Andrews and Askey concerning (1.2) is correct and that in the general case (1.4) with $\delta\theta\neq 0$, Askey's conjecture is nearly right. However there are two exceptional cases closely related to the q-Pollaczek polynomials. We mention that the different but related generating function

$$A(t)\prod_{n=0}^{\infty}[1-xK(t)q^n]^{-1}=\sum_{n=0}^{\infty}P_n(x)t^n$$

was studied by Ismail [11] who found all cases where the $P_n(x)$ are orthogonal polynomials.

2. General necessary conditions. We study the polynomial generating function

$$(2.1)\qquad G(x,t)=A(t)\prod_{k=0}^{\infty}\frac{1-\delta xH(tq^k)}{1-\theta xK(tq^k)}=\sum_{n=0}^{\infty}Q_n(x)t^n$$

where $0<|q|<1$, $|\delta|+|\theta|>0$, and

$$A(t)=\sum_{m=0}^{\infty} a_m t^m, \tag{2.2}$$

$$H(t)=\sum_{m=1}^{\infty} h_m t^m,$$

$$K(t)=\sum_{m=1}^{\infty} k_m t^m$$

and we assume that $\{Q_n(x)\}$ is a sequence of orthogonal polynomials. In particular, $Q_n(x)$ must be of precise degree n and there is no loss of generality if we take

$$a_0=1, \qquad h_1=k_1=1, \tag{2.3}$$

and for convenience we also write

$$h_0=k_0=0. \tag{2.4}$$

Now considering (2.1), replace x by x^{-1}, t by xt. Upon letting $x\to 0$, we obtain

$$A(0)\prod_{m=0}^{\infty}\frac{1-\delta q^m t}{1-\theta q^m t}\sum_{n=0}^{\infty} b_n t^n$$

where b_n denotes the leading coefficient of $Q_n(x)$. This yields

$$b_n=\frac{(\delta/\theta;q)_n\theta^n}{(q;q)_n}. \tag{2.5}$$

Here we use the usual notation,

$$(a;q)_0=1, \qquad (a;q)_n=(1-a)(1-aq)\cdots(1-aq^{n-1}) \quad \text{for } n>0.$$

It then follows that the three-term recurrence relation for these polynomials has the form

$$(1-q^{n+1})Q_{n+1}(x)=[(\theta-\delta q^n)x+\beta_n]Q_n(x)-\gamma_n Q_{n-1}(x). \tag{2.6}$$

From (2.1), we have that $Q_n(0)=a_n$ so

$$(1-q^{n+1})a_{n+1}=\beta_n a_n-\gamma_n a_{n-1}, \qquad n\geqq 1. \tag{2.7}$$

Next we set

$$\mathscr{A}(t)=\frac{A(qt)}{A(t)}=\sum_{n=0}^{\infty}\alpha_n t^n.$$

We then have

$$G(x,qt)=\mathscr{A}(t)\frac{1-\theta xK(t)}{1-\delta xH(t)}G(x,t)$$

from which we obtain

$$[1-\theta xK(t)]\sum_{m=0}^{\infty}\alpha_m t^m\sum_{n=0}^{\infty}Q_n(x)t^n=[1-\delta xH(t)]\sum_{n=0}^{\infty}Q_n(x)q^nt^n.$$

This can be expanded to

$$\sum_{n=0}^{\infty}\sum_{j=0}^{n}\alpha_jQ_{n-j}(x)t^n-\sum_{n=0}^{\infty}Q_n(x)q^nt^n$$
$$=x\left[\theta K(t)\sum_{n=0}^{\infty}\sum_{j=0}^{n}\alpha_jQ_{n-j}(x)t^n-\delta H(t)\sum_{n=0}^{\infty}Q_n(x)q^nt^n\right]$$

and this yields for $n>0$

$$(2.8)\quad q^nQ_n(x)-\sum_{j=0}^{n}\alpha_jQ_{n-j}(x)$$
$$=x\left[\delta\sum_{m=0}^{n-1}h_{m+1}Q_{n-m-1}(x)q^{n-m-1}-\theta\sum_{m=0}^{n-1}k_{m+1}\sum_{j=0}^{n-m-1}\alpha_jQ_{n-m-j-1}(x)\right].$$

We use the recurrence formula (2.6) to eliminate $\delta xq^sQ_s(x)$ from the right side of (2.8). The result is

$$q^nQ_n(x)-\sum_{j=0}^{n}\alpha_jQ_{n-j}(x)$$
$$=\sum_{i=0}^{n-1}h_{i+1}\left[\beta_{n-i-1}Q_{n-i-1}(x)-\gamma_{n-i-1}Q_{n-i-2}(x)-(1-q^{n-i})Q_{n-i}(x)\right]$$
$$-\theta x\sum_{r=0}^{n-1}\left[k_{r+1}\sum_{j=0}^{n-r-1}\alpha_jQ_{n-r-j-1}(x)-h_{r+1}Q_{n-r-1}(x)\right]$$

which in turn yields

$$(2.9)\quad \sum_{i=0}^{n}\alpha_iQ_{n-i}(x)=\theta x\sum_{i=0}^{n-1}F_{n,i}(x)-\beta_{n-1}Q_{n-1}(x)+(1-q^{n-1})h_2Q_{n-1}(x)$$
$$-\sum_{j=0}^{n}\left[h_j\beta_{n-j}-h_{j-1}\gamma_{n-j+1}-h_{j+1}(1-q^{n-j})\right]Q_{n-j}(x)$$

where

$$F_{n,i}(x)=k_{i+1}\sum_{j=0}^{n-i-1}\alpha_jQ_{n-i-j-1}(x)-h_{i+1}Q_{n-i-1}(x).$$

Once again using (2.8), we eliminate the $xQ_s(x)$ terms from $xF_{n,i}(x)$ and obtain

(2.10)

$$\begin{aligned} xF_{n,i}(x) &= \frac{k_{i+1}-h_{i+1}}{\theta-\delta q^{n-i-1}}(1-q^{n-i})Q_{n-i}(x) \\ &+ \left[\frac{\alpha_1 k_{i+1}(1-q^{n-i-1})}{\theta-\delta q^{n-i-2}} - \frac{k_{i+1}-h_{i+1}}{\theta-\delta q^{n-i-1}}\beta_{n-i-1}\right]Q_{n-i-1}(x) \\ &+ \left[\frac{\alpha_2 k_{i+1}(1-q^{n-i-2})}{\theta-\delta q^{n-i-3}} - \frac{\alpha_1 k_{i+1}\beta_{n-i-2}}{\theta-\delta q^{n-i-2}} + \frac{k_{i+1}-h_{i+1}}{\theta-\delta q^{n-i-1}}\gamma_{n-i-1}\right]Q_{n-i-2}(x) \\ &+ k_{i+1}\sum_{s=3}^{n-i} D_s(n,i)Q_{n-i-s}(x) \end{aligned}$$

where

$$D_s(n,i) = \frac{\alpha_s(1-q^{n-i-s})}{\theta-\delta q^{n-i-s-1}} - \frac{\alpha_{s-1}\gamma_{n-i-s}}{\theta-\delta q^{n-i-+s}} + \frac{\alpha_{s-2}\gamma_{n-i-s+1}}{\theta-\delta q^{n-i-s+1}}.$$

We next compare the coefficients of $Q_{n-1}(x)$ in (2.9) with the aid of (2.10) and find

$$\alpha_1 = \frac{\theta(\alpha_1+k_2-h_2)(1-q^{n-1})}{\theta-\delta q^{n-2}} - \beta_{n-1} + (1-q^{n-1})h_2, \qquad n \geqq 1,$$

so that

$$\beta_n = \frac{\theta(\alpha_1+k_2)-\delta h_2 q^{n-1}}{\theta-\delta q^{n-1}}(1-q^n) - \alpha_1, \qquad n \geqq 0. \tag{2.11}$$

Similarly, comparison of the coefficients of $Q_{n-2}(x)$ yields

$$\begin{aligned} \gamma_n &= \frac{\theta(\alpha_1+k_2)-\delta h_2 q^{n-1}}{\theta-\delta q^{n-1}}\beta_{n-1} \\ &- \frac{\theta(\alpha_2+\alpha_1 k_2+k_3)-\delta h_3 q^{n-2}}{\theta-\delta q^{n-2}}(1-q^{n-1}) + \alpha_2, \qquad n \geqq 1. \end{aligned} \tag{2.12}$$

As derived, (2.12) does not necessarily hold for $n=1$ but (2.7) can be used to obtain

$$\gamma_1 = -\frac{\alpha_1(\beta_1+\alpha_1)}{(1-q)} + \alpha_2$$

and if (2.11) is used to eliminate β_1, this reduces to (2.12) for $n=1$. Finally, we equate coefficients of $Q_{n-s}(x)$ and obtain

$$\begin{aligned} \alpha_s &= \left\{\theta\sum_{i=0}^{s}\alpha_i k_{s-i+1} - \delta h_{s+1}q^{n-s-1}\right\}\frac{1-q^{n-s}}{\theta-\delta q^{n-s-1}} \\ &- \left\{\theta\sum_{i=0}^{s-1}\alpha_i k_{s-i} - \delta h_s q^{n-s}\right\}\frac{\beta_{n-s}}{\theta-\delta q^{n-s}} \\ &+ \left\{\theta\sum_{i=0}^{s-2}\alpha_i k_{s-i-1} - \delta h_{s-1}q^{n-s+1}\right\}\frac{\gamma_{n-s+1}}{\theta-\delta q^{n-s+1}}. \end{aligned} \tag{2.13}$$

One can verify directly that (2.13) remains valid for $n \geqq 1$ provided we interpret an empty sum as 0.

3. The special cases where $\delta\theta = 0$. We first take up the case $\delta = 0$. We can then set $\theta = 1$ without loss of generality so that (2.11) and (2.12) can be written

$$\beta_n = \beta - bq^n, \qquad \gamma_n = \gamma - cq^{n-1} \tag{3.1}$$

where

$$\beta = k_2, \quad \gamma = k_2^2 - k_3, \quad b = \alpha_1 + k_2, \quad c = \alpha_1^2 + \alpha_1 k_2 - \alpha_2 - k_3.$$

Setting $Q_n^*(x) = (q;q)_n Q_n(x-\beta)$, we can write the recurrence formula (2.6) as

$$Q_{n+1}^*(x) = (x - bq^n) Q_n^*(x) - (\gamma - cq^{n-1}) Q_{n-1}^*(x). \tag{3.2}$$

Thus the $Q_n^*(x)$ are the polynomials found by us in [2], the ones whose orthogonality relations were first found by Askey and Wilson (see the introduction). If $\gamma \neq 0$, these polynomials are special cases of the q-Pollaczek polynomials. If $\gamma = 0$, these polynomials are the Al-Salam and Carlitz q-polynomials $U_n^{(a)}(x)$ (see [1], [8, p. 196]). The generating function given in [2] is

$$\phi(x,t) = \prod_{k=0}^{\infty} \frac{1 - btq^k + ct^2q^{2k}}{1 - xtq^k + t^2q^{2k}} = \sum_{n=0}^{\infty} Q_n(x-\beta)t^n \tag{3.3}$$

which is one of the right form.

It remains to show that the product representation in (3.3) is the only one of the form (2.1). We let $n \to \infty$ in (2.13) and obtain

$$\alpha_s = \sum_{i=0}^{s} \alpha_i k_{s-i+1} - \beta \sum_{i=0}^{s-1} \alpha_i k_{s-i} + \gamma \sum_{i=0}^{s-2} \alpha_i k_{s-i-1},$$

$$\sum_{i=0}^{s-1} \alpha_i (k_{s-i+1} + \beta k_{s-i} - \gamma k_{s-i-1}) = 0, \qquad s \geqq 1.$$

Since $\alpha_0 \neq 0$, it follows that

$$k_{s+1} - \beta k_s + \gamma k_{s-1} = 0, \qquad s \geqq 1.$$

Together with the initial conditions $k_0 = 0$, $k_1 = 1$, this yields

$$K(t) = \frac{t}{1 - \beta t + \gamma t^2}. \tag{3.4}$$

Thus

$$1 - xK(t) = \frac{1 - (x+\beta)t + \gamma t^2}{1 - \beta t + \gamma t^2}$$

and this shows that the generating function (2.1) must be (3.3).

Next we let $\theta = 0$ and take $\delta = 1$. Now (2.11) and (2.12) reduce to

$$\beta_n = \beta - bq^n, \qquad \gamma_n = \gamma - cq^{n-1} \tag{3.5}$$

where

$$b=h_2,\quad c=h_2^2-h_3,\quad \beta=h_2-\alpha_1,\quad \gamma=h_2^2-h_3-\alpha_1h_2+\alpha_2.$$

Now (2.13) reads

$$\alpha_s=-(1-q^{n-s})[h_{s+1}-bh_s+ch_{s-1}]-\alpha_1(h_s-bh_{s-1})-\alpha_2h_{s-1},$$

from which we conclude

$$(3.6)\qquad h_{s+1}-bh_s+ch_{s-1}=0,\qquad s\geqq 1,$$

$$(3.7)\qquad \alpha_s=-\alpha_1h_s+(b\alpha_1-\alpha_2)h_{s-1},\qquad s\geqq 1.$$

For $s=1$ and 2, (3.6) yields $\alpha_1=\alpha_2=0$; hence $\alpha_s=0$ for all $s\geqq 1$. Thus $A(t)=1$ and from (3.7) we find,

$$(3.8)\qquad H(t)=\frac{t}{1-bt+ct^2}.$$

The recurrence formula (2.6) becomes

$$(3.9)\qquad (1-q^{n+1})Q_{n+1}(x)=[-xq^n+\beta-bq^n]Q_n(x)-(\gamma-cq^{n-1})Q_{n-1}(x).$$

Comparing (3.9) with the recurrence formula (3.68) in Askey and Ismail's memoir [5], we see that (using their notation)

$$(3.10)\qquad Q_n(x)=v_n(x+b;q;\beta,\gamma,c).$$

For $\gamma\neq 0$ these polynomials are special cases of the q-Pollaczek polynomials. For $\gamma=0$, they can be transformed to the Al-Salam and Carlitz polynomials $V_n^{(a)}(x)$ ([1], [8, p. 196]). The generating function (2.1) determined by (3.8) is the one given by Askey and Ismail.

4. The case $\delta\theta\neq 0$.

Turning now to the case $\delta\theta\neq 0$, we take $\theta=1$ without loss of generality.

Referring to (2.11) and (2.12), we first note the limits

$$(4.1)\qquad \beta\equiv\lim_{n\to\infty}\beta_n=k_2,$$

$$(4.2)\qquad \gamma\equiv\lim_{n\to\infty}\gamma_n=k_2^2-k_3.$$

We let $n\to\infty$ in (2.13) and get exactly as we did in §3

$$(4.3)\qquad K(t)=\frac{t}{1-\beta t+\gamma t^2}.$$

Next we note that if we write $x=q^n$, (2.13) can be rewritten as

$$(4.4)\qquad \begin{aligned}\alpha_s=&\left(A_s-\delta q^{-s-1}h_{s+1}x\right)\frac{1-xq^{-s}}{1-\delta q^{-s-1}x}\\&-\left(B_s-\delta q^{-s}h_sx\right)\frac{\beta(x)}{1-\delta q^{-s}x}\\&+\left(C_s-\delta q^{-s+1}h_{s-1}x\right)\frac{\gamma(x)}{1-\delta q^{-s+1}x}\end{aligned}$$

where A_s, B_s, and C_s are independent of x, $\beta(x)$ is obtained from (2.11):

$$\beta(x)=\frac{(\alpha_1+k_2-\delta h_2 q^{-s-1}x)(1-q^{-s}x)}{1-\delta q^{-s-1}x}-\alpha_1 \tag{4.5}$$

and $\gamma(x)$ is similarly obtained from (2.12). Since $|q|\neq 0,1$, (4.4) is valid for infinitely many values of x and hence it must be an identity in x. We therefore multiply both sides of (4.4) by x^{-1} and let $x\to\infty$. We have

$$\lim_{x\to\infty}\frac{\beta(x)}{1-\delta q^{-s}x}=\frac{h_2}{\delta},\qquad \lim_{x\to\infty}\frac{\gamma(x)}{1-\delta q^{-s+1}x}=\frac{h_2^2-h_3}{q^2}$$

so (4.4) gives

$$0=h_{s+1}-h_2h_s+(h_2^2-h_3)h_{s-1},\qquad s\geqq 1.$$

Letting

$$b=h_2,\qquad c=h_2^2-h_3, \tag{4.6}$$

we conclude

$$H(t)=\frac{t}{1-bt+ct^2}. \tag{4.7}$$

The generating function (2.1) thus has the form

$$G(x,t)=D(t)\prod_{k=0}^{\infty}\frac{1-(\delta x+b)tq^k+ct^2q^{2k}}{1-(x+\beta)tq^k+\gamma t^2q^{2k}} \tag{4.8}$$

where

$$D(t)=A(t)\prod_{k=0}^{\infty}\frac{1-\beta tq^k+\gamma t^2q^{2i}k}{1-btq^k+ct^2q^{2k}}=\sum_{n=0}^{\infty}d_nt^n.$$

At this point we can see that there will be no loss of generality if we assume henceforth that $\beta=0$ and $\gamma=1$ since this amounts to a translation in x and a scale change in t. With this convention adopted, comparison of (4.8) with (1.6) shows that we have

$$Q_n(x)=\sum_{k=0}^{n}d_kP_{n-k}(x;q;a,b,c)\qquad(a=\delta,d_0=1). \tag{4.9}$$

We will complete the characterization problem by determining all sequences $\{d_n\}$ such that the $Q_n(x)$ determined by (4.9) will be orthogonal polynomials. To this end, we recall the recurrence formula (2.6) which now reads

$$(1-q^{n+1})Q_{n+1}(x)=[(1-aq^n)x+\beta_n]Q_n(x)-\gamma_nQ_{n-1}(x). \tag{4.10}$$

Now we use (4.9) to eliminate $Q_k(x)$ from (4.10). We also use the recurrence relation (1.5) to linearize the $xP_k(x)$ terms. The result is

$$\begin{aligned}&(1-q^{n+1})\sum_{k=0}^{n+1}d_kP_{n+1-k}(x)\\&\quad=\beta_n\sum_{k=0}^{n}d_kP_{n-k}(x)-\gamma_n\sum_{k=0}^{n-1}d_kP_{n-k-1}(x)\\&\qquad+(1-aq^n)\sum_{k=0}^{n}\frac{d_k}{1-aq^{n-k}}\big[(1-q^{n-k+1})P_{n+1-k}(x)-bq^{n-k}P_{n-k}(x)\\&\qquad\qquad+(1-cq^{n-k-1})P_{n-k-1}(x)\big].\end{aligned}$$

Note that (4.10) shows that a is restricted by the condition that

$$1-aq^n\neq 0, \qquad n\geqq 0. \tag{4.11}$$

Equating coefficients of $P_{n-k}(x)$ now yields

$$(1-q^{n+1})d_{k+1}=d_k\beta_n-d_{k-1}\gamma_n+\frac{(1-aq^n)(1-q^{n-k})}{1-aq^{n-k-1}}d_{k+1}$$
$$-\frac{bq^{n-k}(1-aq^n)}{1-aq^{n-k}}d_k+\frac{(1-aq^n)(1-cq^{n-k})}{1-aq^{n-k+1}}d_{k-1}$$

and this can be written

$$\frac{q^{n-k-1}(1-q^{k+1})(q-a)}{1-aq^{n-k-1}}d_{k+1}=\left\{\beta_n-\frac{bq^{n-k}(1-aq^n)}{1-aq^{n-k}}\right\}d_k$$
$$+\left\{\frac{(1-aq^n)(1-cq^{n-k})}{1-aq^{n-k+1}}-\gamma_n\right\}d_{k-1}. \tag{4.12}$$

Setting $k=0$ and $k=1$ yields, respectively,

$$\beta_n=bq^n+\frac{d_1q^{n-1}(1-q)(q-a)}{1-aq^{n-1}}, \tag{4.13}$$

(4.14)

$$\gamma_n=1-cq^{n-1}-\frac{d_1b(1-q)q^{n-1}}{1-aq^{n-1}}+\frac{d_1^2(1-q)(q-a)q^{n-1}}{1-aq^{n-1}}-\frac{d_2(1-q^2)(q-a)q^{n-2}}{1-aq^{n-2}}.$$

Because of (4.10), (4.13) and (4.14) are valid for $n\geqq 0$ and $n\geqq 1$, respectively, except in the special case $a=q$. In the latter case, (4.13) need not hold for $n=0$ and (4.14) need not hold for $n=1$.

Using (4.13) and (4.14) in (4.10) now gives

(4.15)

$$\frac{(1-q^{k+1})(q-a)}{1-aq^{n-k-1}}d_{k+1}+\left\{\frac{bq(1-q^k)}{1-aq^{n-k}}-\frac{d_1(1-q)(q-a)q^k}{1-aq^{n-1}}\right\}d_k$$
$$+\left\{\frac{q(c-aq)(1-q^{k-1})}{1-aq^{n-k+1}}-\frac{d_1b(1-q)q^k}{1-aq^{n-1}}\right.$$
$$\left.+\frac{d_1^2(1-q)(q-a)q^k}{1-aq^{n-1}}-\frac{d_2(1-q^2)(q-a)q^{k-1}}{1-aq^{n-2}}\right\}d_{k-1}=0.$$

5. The case $\delta\theta\neq 0$ continued. Referring to (4.15), we note as before that we can formally replace q^n by x and observe that the result must be an identity involving a rational function with poles at $x=a^{-1}q^s$, $s=1,2,k-1,k,k+1$. Taking $k\geqq 2$, the residue at $x=a^{-1}q^{k+1}$ gives us the condition

$$(q-a)d_k=0, \qquad k\geqq 3. \tag{5.1}$$

Let us first consider the case $a \neq q$. In this case, $d_m = 0$ for $m \geqq 3$. Now take $k=2$ and look at the residue at $x = a^{-1}q^2$. We find

$$d_2[a-(q-a)d_1]=0.$$

If $d_2 \neq 0$, then

$$d_1 = \frac{a}{q-a}. \tag{5.2}$$

Then let $k=3$ and consider the pole $x=a^{-1}q^2$. We find

$$[q(1-q^2(c-aq)-d_2q^2(1-q^2)(q-a))]\,d_2=0 \tag{5.3}$$

so that

$$d_2 = \frac{c-aq}{q(q-a)}. \tag{5.4}$$

Thus for the case $a \neq q$, $c \neq aq$, we will have

$$Q_n(x) = P_n(x) + \frac{a}{q-a}P_{n-1}(x) + \frac{c-aq}{q(q-a)}P_{n-2}(x) \tag{5.5}$$

and the corresponding recurrence formula will read

$$(1-q^{n+1})Q_{n+1}(x) = \left\{(1-aq^n)x + \frac{b(1-aq^n)q^{n-1}}{1-aq^{n-1}}\right\}Q_n(x) - \frac{(1-cq^{n-3})(1-aq^n)}{1-aq^{n-2}}Q_{n-1}(x). \tag{5.6}$$

If we set

$$Q_n^{\#}(x) + \frac{1-aq^{-1}}{1-aq^{n-1}}Q_n(x),$$

it is easily seen from (5.6) that $Q_n^{\#}(x)$ satisfies the Pollaczek recurrence (1.5) with parameters aq^{-1}, bq^{-1}, cq^{-2}. Thus

$$Q_n(x) = \frac{1-aq^{n-1}}{1-aq^{-1}}P_n(x;q;aq^{-1},bq^{-1},cq^{-2}), \qquad a \neq q, \quad c \neq aq. \tag{5.7}$$

Still keeping $a \neq q$, suppose next that $d_2 = 0$ so that

$$Q_n(x) = P_n(x) + d_1 P_{n-1}(x). \tag{5.8}$$

We then take $k=2$ in (4.15) and consider $x=a^{-1}q$. This time we get

$$q(q-a)d_1^2 - bqd_1 + c - aq = 0 \tag{5.9}$$

as the condition determining the value of d_1. The coefficients for the recurrence formula can be written

$$\begin{aligned} \beta_n &= bq^n + \frac{d_1(1-q)(q-a)q^{n-1}}{1-aq^{n-1}}, \\ \gamma_n &= \frac{(1-cq^{n-2})(1-aq^n)}{1-aq^{n-1}}. \end{aligned} \tag{5.10}$$

We will identify these polynomials in terms of their orthogonality relations in the next section. Here we turn to the remaining cases which occur when $a=q$. When $a=q$, (4.15) reduces to

$$\text{(5.11)}\qquad \frac{bq(1-q^k)}{1-q^{n-k+1}}d_k+\left\{\frac{q(c-q^2)(1-q^{k-1})}{1-q^{n-k+2}}-\frac{d_1b(1-q)q^k}{1-q^n}\right\}d_{k-1}=0.$$

If $b=0$, (5.11) yields the condition

$$(c-q^2)d_n=0,\qquad n\geqq 1.$$

Thus if $c\neq q^2$, we have $d_k=0$ for $k\geqq 1$. Therefore when $a=q$, $b=0$, $c\neq q^2$,

$$\text{(5.12)}\qquad Q_n(x)=P_n(x;q;q,0,c).$$

If, however, $c=q^2$, the recursion coefficients (4.13), (4.14) become

$$\beta_n=0,\quad n\geqq 1,\qquad \gamma_n=1-q^{n+1},\quad n\geqq 2.$$

Thus the recurrence relation reduces to

$$\text{(5.13)}\qquad \begin{aligned} &Q_{n+1}(x)=xQ_n(x)-Q_{n-1}(x),\qquad n\geqq 2.\\ &Q_2(x)=xQ_1(x)-\gamma Q_0(x),\\ &Q_1(x)=x+\beta,\qquad Q_0(x)=1.\end{aligned}$$

It follows that for the case $a=q$, $b=0$, $c=q^2$, the generating function is

$$\text{(5.14)}\qquad \frac{1+\beta t+\gamma t^2}{1-xt+t^2}=\sum_{n=0}^{\infty}Q_n(x)t^n$$

where β and $\gamma>0$ are arbitrary. The polynomials are

$$\text{(5.15)}\qquad Q_n(x)=U_n(x/2)+\beta U_{n-1}(x/2)+\gamma U_{n-2}(x/2),$$

where $U_n(x)=P_n(2x;q;q,0,q^2)$ are the Chebyshev polynomials of the second kind (with $U_{-2}(x)=U_{-1}(x)=0$). The orthogonality of these polynomials is included in studies of a more general situation by Geronimus (see [10, p. 52]). The distribution functions (including any isolated mass points that occur) are given explicitly for all cases by Allaway [3]. (Note that $Q_n^*(x)=\gamma^{-1}Q_n(x)$ satisfies (5.13) for $n>1$.) They can also be found in [6] where Askey and Wilson have given a further extension of Allaway's work.

The remaining cases occur when $a=q$, $b\neq 0$. To handle this case, first take $k\geqq 2$ in (5.11). The residue at $x=q^{k-1}$ then tells us that $d_k=0$ for $k\geqq 2$. When $k=2$, (5.11) then yields the condition

$$(c-q^2-bqd_1)d_1=0.$$

If $d_1=0$, we have $Q_n(x)=P_n(x;q;q,b,c)$ $(b\neq 0)$ so take $d_1\neq 0$. Then

$$\text{(5.16)}\qquad d_1=\frac{c-q^2}{bq},\qquad c\neq q^2.$$

The recursion coefficients now become

$$\text{(5.17)}\qquad \beta_n=bq^n,\qquad \gamma_n=\frac{(1-cq^{n-2})(1-q^{n+1})}{1-q^n},\qquad n\geqq 1.$$

According to the remarks made following (4.14), the formula for γ_n is valid for $n \geqq 2$ but it can be verified directly that (5.17) gives γ_1 correctly also. Also, it is easy to verify directly that

$$\beta_0 = b + d_1(1-q). \tag{5.18}$$

When these formulas are compared with (5.9) and (5.10), we see that this case is included in (5.8) as the limiting case $a=q$, $b \neq 0$, $c \neq q^2$. Thus (5.8) with d_1 determined by (5.9) remains the only case for which orthogonality relations are not known.

6. Orthogonality relations. We will now find orthogonality relations for the polynomials given by (5.8) and (5.9) with $P_n(x) = P_n(x; q; a, b, c)$ and with $a=q$ now allowed. We set

$$\sigma = d_1 + \frac{1}{d_1}, \tag{6.1}$$

translate: $x \to x - \sigma$, and then consider the monic polynomials

$$\hat{Q}_n(x) = \frac{(q;q)_n}{(a;q)_n} Q_n(x-\sigma). \tag{6.2}$$

With reference to (5.10), we find these polynomials satisfy

$$\hat{Q}_n(x) = (x - c_n)\hat{Q}_{n-1}(x) - \lambda_n \hat{Q}_{n-2}(x), \qquad n \geqq 1, \tag{6.3}$$

$$c_n = \sigma - \frac{bq^{n-1}}{1-aq^{n-1}} - \frac{d_1(1-q)(q-a)q^{n-2}}{(1-aq^{n-2})(1-aq^{n-1})}, \tag{6.4}$$

$$\lambda_{n+1} = \frac{(1-cq^{n-2})(1-q^n)}{(1-aq^{n-1})^2}. \tag{6.5}$$

Note that when $a=q$, $c_1 = d_1^{-1} - b/(1-q)$.

Now define

$$\Gamma_{2n+1} = \frac{(1-q^n)d_1}{1-aq^{n-1}}, \quad \Gamma_{2n+2} = \frac{1-cq^{n-1}}{(1-aq^n)d_1}, \quad n \geqq 0. \tag{6.6}$$

With the aid of (5.9), one can now verify directly (but tediously) that

$$c_n = \Gamma_{2n-1} + \Gamma_{2n}, \quad \lambda_{n+1} = \Gamma_{2n}\Gamma_{2n+1}, \qquad n \geqq 1. \tag{6.7}$$

On the other hand, the corresponding monic q-Pollaczek polynomials,

$$\hat{P}_n(x) = [(q;q)_n/(a;q)_n]\, P_n(x-\sigma; q; a, b, c),$$

satisfy

$$\hat{P}_n(x) = (x - d_n)\hat{P}_{n-1}(x) - \nu_n \hat{P}_{n-2}(x), \qquad n \geqq 1, \tag{6.8}$$

$$d_n = \sigma - \frac{bq^{n-1}}{1-aq^{n-1}}, \qquad \nu_{n+1} = \frac{(1-cq^{n-1})(1-q^n)}{(1-aq^{n-1})(1-aq^n)}. \tag{6.9}$$

Now one can verify that

$$(6.10) \qquad d_n = \Gamma_{2n} + \Gamma_{2n+1}, \quad \nu_{n+1} = \Gamma_{2n+1}\Gamma_{2n+2}, \qquad n \geqq 1.$$

It follows from (6.6) and (6.10) that the $\hat{P}_n(x)$ are kernel polynomials (with K-parameter 0) corresponding to $\hat{Q}_n(x)$. That is, if $\{\hat{Q}_n(x)\}$ is an orthogonal polynomial sequence (OPS) with respect to a measure $d\mu(x)$, then the existence of Γ_n such that (6.6) and (6.10) hold is necessary and sufficient for $\{P_n(x)\}$ to be the OPS with respect to $x d\mu(x)$ (see [9, p. 46]). Therefore, if $d\varphi(x)$ denotes the measure with respect to which the q-Pollaczek polynomials are orthogonal, then the polynomials given by (5.8) are orthogonal with respect to the measure given by

$$(6.11) \qquad d\psi(x) = (x+\sigma)^{-1} d\varphi(x+\sigma),$$

together with a possible positive mass at $x = -\sigma$. All of this is subject to restricting the parameters a, b, c and d_1 so that the Γ_n are positive for $n \geqq 2$. If they are not all positive but are all nonzero, then corresponding relations still hold formally with integrals replaced by moment functionals (see [9]). We will here assume we have the positive definite case:

$$\begin{aligned} d_1 > 0 \text{ and either} \quad &\text{(i)} \quad 0<q<1, \quad a<1, \quad c<q; \\ \text{or} \quad &\text{(ii)} \quad -1<q<0, \quad q^{-1}<a<1, \quad 0<c<1. \end{aligned}$$

We will complete the analysis by determining the cases where $d\psi$ has positive mass at $x = -\sigma$.

First note that we have not lost any generality by the assumption that $|q|<1$ since the case $|q|>1$ can be handled by setting $q = p^{-1}$, $|p|<1$. The resulting recurrence formula written in terms of p can be reduced after a linear change of variable to the original form and the generating function would remain of the same type. Now in any case, the coefficients in the recurrence formula are bounded so the corresponding Hamburger moment problem is determined. We consider the corresponding orthonormal polynomials

$$(6.12) \qquad \hat{q}_n(x) = (\lambda_1\lambda_2 \cdots \lambda_{n+1})^{-1/2} \hat{Q}_n(x),$$

where λ_1 denotes the total mass of the measure $d\psi$. Now $[\hat{q}_n(0)]^2 = (\Gamma_2 \cdots \Gamma_{2n})/(\lambda_1\Gamma_3 \cdots \Gamma_{2n+1})$ (see [9, p. 49]). Hence we have

$$(6.13) \qquad [\hat{q}_n(0)]^2 = \frac{(cq^{-1};q)_n}{\lambda_1 (q;q)_n d_1^{2n}} \sim d_1^{-2n}.$$

Now it is a classical result from the problem of moments [12] that when the Hamburger moment problem is determined, the mass at x is $\rho(x) = \{\Sigma[\hat{q}_n(x)]^2\}^{-1}$. Thus when $d_1 > 1$, at the point $x = -\sigma$, $d\psi(x)$ will have the mass $\rho(0) = \lambda_1 J$, where

$$(6.14) \qquad J = \{{}_1\phi_0(cq^{-1}; -; r)\}^{-1} = \prod_{n=0}^{\infty} \frac{1-rq^n}{(1-rcq^{n-1})}, \qquad r = d_1^{-1}$$

and ψ will be continuous at σ in all other cases.

Finally, we note that the total mass of $d\psi(x)$ is given by

$$(6.15) \qquad \lambda_1 = (1-J)^{-1} \int_{-\infty}^{\infty} \frac{d\varphi(x+\sigma)}{x+\sigma}.$$

In those cases where the q-Pollaczek polynomials have a continuous spectrum, the measure $d\varphi(x)$ is given by Charris and Ismail [8, §6]. They denote the q-Pollaczek polynomial by $F_n(x;u,v,w;q)$ where

$$F_n(x;u,v,w;q)=P_n(2x;q;uw,2v,w^2). \tag{6.16}$$

7. Summary and remarks. For the special case $\delta=0$ of (2.1), the Andrews–Askey conjecture is verified: the only OPS generated are the so-called Al-Salam–Chihara polynomials (apart from trivial changes of variable and renormalizations). When $\theta=0$, the only OPS generated are the "q-duals" $v_n(x)$ which correspond to the formal replacement of q by q^{-1} in the first set.

In the general case $\delta\theta\neq 0$, the OPS generated are the q-Pollaczek polynomials together with two closely related classes. The first consists of the most general orthogonal solutions of the Chebyshev recurrence (coefficients independent of n), and the second is the class of "inverse kernel" polynomials discussed in §6. An interesting near exception is provided by the polynomials given by (5.5). According to (5.7), these polynomials are also q-Pollaczek polynomials which satisfy the interesting identity (5.5). Additionally these results say that if $\alpha\neq 1$, $\gamma\neq\alpha$, then $P_n(x;q;\alpha,\beta,\gamma)$ and $(1-\alpha q^n)P_n(x;q;\alpha,\beta,\gamma)$ are each generated by different generating functions having the general form (2.1).

The relations involving Γ_n obtained in §6 are quite general and provide a number of formal identities involving q-Pollaczek polynomials. In particular, we obtain the symmetric OPS $\{S_n(x)\}$ defined by

$$S_{2n}(x)=\hat{Q}_n(x^2),\qquad S_{2n+1}(x)=x\hat{P}_n(x^2). \tag{7.1}$$

In the positive-definite case, these are the polynomials orthogonal with respect to the symmetrization of the measure $d\psi(x)$ (6.11).

Finally, we observe that the q-Pollaczek polynomials enjoy a convolution property. Using here the normalization of Askey and Ismail [5, p. 17], consider the generating function for the Al-Salam–Chihara polynomials $Q_n(x)=Q_n(x;q;a,b,c)$:

$$Q(xq;a,b,cc;t)=\prod_{k=0}^{\infty}\frac{1-atq^k+bt^2q^{2k}}{1-xtq^k+ct^2q^{2k}}=\sum_{n=0}^{\infty}Q_n(x)t^n$$

and for their q-duals, $v_n(x)=v_n(x;q;\alpha,\beta,\gamma)$:

$$V(x;q;\alpha,\beta,\gamma;t)=\prod_{k=0}^{\infty}\frac{1-xtq^k+\gamma t^2q^{2k}}{1-\alpha tq^k+\beta t^2q^{2k}}=\sum_{n=0}^{\infty}v_n(x)t^n.$$

Referring to (1.6), we see that

$$Q(x;q;\alpha,\beta,1;t)V(ax+b;q;\alpha,\beta,c;t)=G(x;q;a,b,c;t).$$

Thus we have the convolution formula

$$P_n(x;q;a,b,c)=\sum_{k=0}^{n}Q_k(x;q;\alpha,\beta,1)v_{n-k}(ax+b;q;\alpha,\beta,c). \tag{7.2}$$

Note that $P_n(x)$ is independent of α and β.

Thus the polynomials of one orthogonal sequence are expressed as the convolution of polynomials from two other orthogonal sequences in the same independent variable. Other examples of this phenomenon are given by setting $y=x$ in all examples found by us in our characterization problem [2] while another class of examples is furnished by the associated Legendre polynomials which are expressible as a convolution of ordinary Legendre polynomials [7]. These include the Chebyshev polynomials of the second kind as a limiting case. All of this of course suggests a characterization problem. Indeed, it was this characterization problem we originally considered before we followed Polya's dictum and looked for an "easier problem we couldn't solve" (which eventually led to [2]). We suspect that the solution of the more difficult characterization problem with only one independent variable would lead to the discovery of new, important classes of orthogonal polynomials.

REFERENCES

[1] W. A. Al-Salam and L. Carlitz, *Some orthogonal q-polynomials*, Math. Nachr., 30 (1965), pp. 47–61.

[2] W. A. Al-Salam and T. S. Chihara, *Convolutions of orthogonal polynomials*, this Journal, 7 (1976), pp. 16–28.

[3] W. R. Allaway, *The identification of a class of orthogonal polynomials*, dissertation, Univ. of Alberta, 1972.

[4] G. E. Andrews and R. Askey, *Classical Orthogonal Polynomials*, to appear.

[5] R. Askey and M. E. H. Ismail, *Recurrence relations, continued fractions and orthogonal polynomials*, Memoirs Amer. Math. Soc., vol. 49, No. 300, 1984.

[6] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Memoirs Amer. Math. Soc., vol. 54, No. 319, 1985.

[7] P. Barrucand and D. Dickinson, *On the associated Legendre polynomials*, in Orthogonal Expansions and Their Continual Analogues, D. T. Haimo, ed., Southern Illinois Univ. Press, Edwardsville, IL, 1968, pp. 43–50.

[8] J. A. Charris and M. E. H. Ismail, *On sieved orthogonal polynomials* V: *Sieved Pollaczek polynomials*, to appear.

[9] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[10] Ja. L. Geronimus, *Orthogonal Polynomials*, Amer. Math. Soc. Translations, Series 2, vol. 108, American Mathematical Society, Providence, RI, 1977.

[11] M. E. H. Ismail, *Orthogonal polynomials in a certain class of polynomials*, Bull. Institut. Politechnic Din IASI, Sect. 1, v. XX (XXIV) (1974), pp. 45–50.

[12] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Mathematical Surveys, No. 1, American Mathematical Society, Providence, RI, 1943, 1963.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

019

ON THE GENERALIZED CHEBYSHEV POLYNOMIALS*

MOURAD E. H. ISMAIL† AND FUAD S. MULLA‡

Abstract. We study the spectrum of the Jacobi matrix $(\delta_{m,n+1}+\delta_{m,n-1}+aq^n\delta_{m,n})$, $m,n=0,1,\cdots$ and the corresponding orthogonal polynomials. The spectral measure is computed when $q\in(-1,1)$ and sufficient conditions are given to guarantee the absolute continuity of the spectral measure. When $q>1$ or <-1 the measure is purely discrete. The case $q=-1$ leads to a set of polynomials orthogonal on the union of two disjoint intervals. When $q=1$, the polynomials are essentially the Chebyshev polynomials $\{U_n(x)\}$.

Key words. Chebyshev polynomials, asymptotics, continued fractions

AMS(MOS) subject classifications. Primary 33A65, 42C05

1. Introduction. The Chebyshev polynomials $U_n(x):=\sin[(n+1)\theta]/\sin\theta$, $x=\cos\theta$, satisfy the three term recursion formula

$$xp_n(x)=\frac{1}{2}p_{n+1}(x)+\frac{1}{2}p_{n-1}(x),\qquad n>0,\tag{1.1}$$

and the initial conditions $p_0(x)=1$, $p_1(x)=2x$. The associated Jacobi matrix is

$$A=\frac{1}{2}(\delta_{m,n+1}+\delta_{m+1,n}),\qquad m,n=0,1,2,\cdots.\tag{1.2}$$

The spectrum of A is $[-1,1]$. Avron and Simon [5] considered the doubly infinite Jacobi matrix

$$B=\frac{1}{2}(\delta_{m,n+1}+\delta_{m+1,n})+(\lambda\cos(2\pi n\alpha+\varepsilon)\delta_{m,n}),\qquad m,n=0,\pm1,\cdots,\tag{1.3}$$

where α is an irrational number belonging to $(0,1)$ and $\varepsilon>0$. They conjectured that the spectrum of B is a Cantor set when $\lambda>1$. The spectrum of the semi-infinite Jacobi matrix

$$B_+=\frac{1}{2}(\delta_{m,n+1}+\delta_{m+1,n})+(\lambda\cos(2\pi n\alpha+\varepsilon)\delta_{m,n}),\qquad m,n=0,1,\cdots,\tag{1.4}$$

where α and ε are as before, is not known but is also very likely to be a Cantor set when $\lambda>1$. The orthogonal polynomials associated with B_+ are generated by

$$\begin{aligned}&2xp_n(x)=p_{n+1}(x)+p_{n-1}(x)+2\lambda\cos(2\pi n\alpha+\varepsilon)p_n(x),\qquad n>0,\\&p_0(x)=1,\qquad p_1(x)=2x-2\lambda\cos\varepsilon,\end{aligned}\tag{1.5}$$

and are very interesting. The limiting distribution of their zeros coincides with the spectrum of B_+. Even the special case $\varepsilon=0$ is very interesting. When $\varepsilon=0$ the term $2\lambda\cos(2\pi\alpha n)$ is $\lambda(q^n+q^{-n})$, where q now lies on the unit circle and $\arg q$ is not a rational multiple of π. This problem, or rather our inability to solve it, motivated us to

*Received by the editors February 14, 1983, and in final revised form June 12, 1985.

†Department of Mathematics, Arizona State University, Tempe, Arizona 85287. This work was done while this author was visiting Kuwait University in 1982.

‡Department of Mathematics, Kuwait University, P. O. Box 5969, Kuwait.

study the polynomials $\theta_n^{(a)}(x;\ q)$ generated by

$$(1.6)\qquad 2x\theta_n^{(a)}(x;\ q)=\theta_{n+1}^{(a)}(x;\ q)+\theta_{n-1}^{(a)}(x;\ q)+aq^n\theta_n^{(a)}(x;\ q),\qquad n>0$$

with

$$(1.7)\qquad \theta_0^{(a)}(x;\ q)=1,\qquad \theta_1^{(a)}(x;\ q)=2x-a.$$

When $a=0$ the θ_n's reduce to the Chebyshev polynomials $U_n(\cos\theta)=\sin[(n+1)\theta]/\sin\theta$. Since the θ_n's generalize the Chebyshev polynomials it is appropriate to call them the generalized Chebyshev polynomials. The associated Jacobi matrix is

$$(1.8)\qquad J=A+D,$$

where D is the diagonal matrix

$$(1.9)\qquad D=\frac{1}{2}(aq^n\delta_{m,n}),\qquad m,n=0,1,2,\cdots,$$

and A is as in (1.2). When $-1<q<1$, J is the sum of the self-adjoint operator A and the compact operator D; hence the essential spectra of J and A are identical, i.e. the essential spectrum of J is $[-1,1]$ since $2A$ is the forward shift plus the backward shift. When $q>1$ or $q<-1$ the matrix operator A is unbounded (on l^2).

In §2 we treat the case $-1<q<1$. We derive a generating function for the θ_n's and apply Darboux's method to the generating function in order to determine the asymptotic behavior of $\theta_n^{(a)}(x;\ q)$ for fixed x and large n. This asymptotic result is then used to compute the resolvent of the operator J in the case $-1<q<1$. The resolvent and some recent results of P. Nevai [12] are used in §3 to compute the absolutely continuous component of the spectral measure of J explicitly. In the rest of §3 and all of §4 we give some necessary and some sufficient conditions for the absence of a discrete singular component of the spectral measure. In §5 an explicit representation of $\theta_n^{(a)}(x;\ q)$ is derived and the sign of the linearization coefficients is determined. Section 6 contains an analysis of the spectrum for $q>1$ and $q<-1$. We show that the continuous spectrum is empty, hence the spectral measure is purely singular. In §7, we treat the case $q=-1$.

The present work is a sequel to Askey and Ismail's work [4] and uses their approach which is originally due to Pollaczek [14].

The following well-known theorem will be used in the text.

THEOREM 1.1. *Let $\{P_n(x)\}$ be a sequence of monic polynomials generated by*

$$(1.10)\qquad P_0(x)=1,\qquad P_1(x)=x-c_1,$$

$$(1.11)\qquad P_n(x)=(x-c_n)P_{n-1}(x)-\lambda_nP_{n-2}(x),\qquad n>1.$$

The positivity condition $\lambda_n>0$ is satisfied if and only if there exists a nondecreasing function $\sigma(x)$ of bounded variation on $(-\infty,\infty)$ such that

$$(1.12)\qquad \int_{-\infty}^{\infty}P_n(x)P_m(x)\,d\sigma(x)=\lambda_1\lambda_2\lambda_3\cdots\lambda_{n+1}\delta_{m,n},$$

where σ is normalized by

$$\lambda_1:=\int_{-\infty}^{\infty}d\sigma(x)=1.$$

For a proof, see Chihara [8, pp. 16–22 and 56–58]. Akhiezer [1] gives an account of the role played by orthogonal polynomials in the spectral theory of Jacobi matrices.

2. The resolvent when $-1<q<1$. We first establish the generating function

$$\sum_{n=0}^{\infty} \theta_n^{(a)}(x;\ q)t^n=\sum_{n=0}^{\infty} \frac{(-at)^k q^{k(k-1)/2}}{(t/A;\ q)_{k+1}(t/B;\ q)_{k+1}}, \tag{2.1}$$

where A and B are roots of $1-2xt+t^2=0$ and the q-shifted factorial $(\sigma;\ q)_n$ is

$$(\sigma;\ q)_0=1,\quad (\sigma;\ q)_n=\prod_{j=1}^{n}(1-\sigma q^{j-1}),\quad n=1,2,\cdots. \tag{2.2}$$

The case $n=\infty$ is also allowed in (2.2) since the infinite product converges for $q\in(-1,1)$. To fix the notation we set

$$A,B=x\pm\sqrt{x^2-1},\qquad |B|\leq|A|. \tag{2.3}$$

Proof of (2.1). Denote the left side of (2.1) by $H(x,t)$. Multiplying (1.6) by t^{n+1} and adding the resulting equalities for $n=1,2,\cdots$, we see that $H(x,t)$ satisfies the functional equation

$$H(x,t)=(1-2xt+t^2)^{-1}[1-atH(x,qt)]. \tag{2.4}$$

We also used the initial conditions (1.7). Iterating (2.4) formally leads to (2.1). This formal argument can be justified. Observe that the right side of (2.1) is analytic in t in a neighborhood of $t=0$ and it satisfies (2.4). Let $\Sigma_0^\infty w_n t^n$ be its Taylor series about $t=0$. It is easily verified that w_j and $\theta_j^{(a)}(x;\ q)$ agree when $j=0,1$ and that w_n satisfies (1.6). This identifies w_n as $\theta_n^{(a)}(x;\ q)$ since both satisfy the same three term recurrence relation and the same initial conditions. This completes the proof of (2.1).

The polynomials of the second kind $\phi_n^{(a)}(x;\ q)$ satisfy the recursion (1.6) and the initial conditions

$$\phi_0^{(a)}(x;\ q)=0,\qquad \phi_1^{(a)}(x;\ q)=2. \tag{2.5}$$

It is not difficult to identify the ϕ_n's as

$$\phi_n^{(a)}(x;\ q)=2\theta_{n-1}^{(aq)}(x;\ q),\quad n>0,\quad \phi_0^{(a)}(x;\ q)=0. \tag{2.6}$$

We now investigate the asymptotic behavior of $\theta_n^{(a)}(x;\ q)$ and $\phi_n^{(a)}(x;\ q)$ for large n and fixed x. The reason is that if $d\psi$ is the *spectral measure* of $A+D$ then the θ_n's are orthogonal with respect to $d\psi^{(a)}(t;\ q)$ and $d\psi^{(a)}(t;\ q)$, normalized by

$$\int_{-\infty}^{\infty} d\psi^{(a)}(t;\ q)=1,$$

satisfies

$$\lim_{n\to\infty}\frac{\phi_n^{(a)}(x;\ q)}{\theta_n^{(a)}(x;\ q)}=\int_{-\infty}^{\infty}\frac{d\psi^{(a)}(t;\ q)}{x-t},\qquad \operatorname{Im} x\neq 0, \tag{2.7}$$

see [4, Thm. 2.4]. The relationships (2.6) and (2.7) identify the resolvent

$$\int_{-\infty}^{\infty}(d\psi^{(a)}(t;\ q)/(x-t))$$

in terms of the asymptotic behavior of $\theta_n^{(a)}(x;\ q)$.

THEOREM 2.1. *The asymptotic behavior of* $\theta_n^{(a)}(x;\, q)$ *is given by*

$$\theta_n^{(a)}(x;\, q) \sim A^n \sum_0^\infty \frac{(-aB)^k q^{k(k-1)/2}}{(B/A;\, q)_{k+1}(q;\, q)_k}, \qquad x \notin [-1,1], \tag{2.8}$$

$$\theta_n^{(a)}(x;\, q) \sim A^n \sum_0^\infty \frac{(-aB)^k q^{k(k-1)/2}}{(B/A;\, q)_{k+1}(q;\, q)_k} + \textit{conjugate}, \qquad x \in (-1,1), \tag{2.9}$$

$$\theta_n^{(a)}(1;\, q) \sim (n+1) \sum_0^\infty \frac{(-a)^k q^{k(k-1)/2}}{(q;\, q)_k^2}, \tag{2.10}$$

$$\theta_n^{(a)}(-1, q) \sim (-1)^n (n+1) \sum_0^\infty \frac{a^k q^{k(k-1)/2}}{(q;\, q)_k^2}. \tag{2.11}$$

Note that A and B are complex conjugates when x belongs to $[-1,1]$. An essential tool in proving (2.8)–(2.11) is the following asymptotic method of Darboux.

THEOREM 2.2 (Darboux's method). *Assume that* $f(z)=\sum_0^\infty f_n z^n$ *is analytic in* $|z|<r$ *and has a finite number of algebraic singularities on* $|z|=r$. *Let* $g(z)=\sum_0^\infty g_n z^n$ *be a comparison function, that is g is also analytic in* $|z|<r$, *has a finite number of algebraic singularities on* $|z|=r$ *and* $f-g$ *is continuous on* $|z|=r$. *Then* $f_n=g_n+o(r^{-n})$.

Olver [13] used the Riemann–Lebesgue lemma to prove Theorem 2.2, see also Szegö [18].

Proof of Theorem 2.1. Apply Darboux's method with the comparison functions

$$(1-t/B)^{-1} \sum_{k=0}^\infty \frac{(-aB)^k q^{k(k-1)/2}}{(B/A;\, q)_{k+1}(q;\, q)_k} \quad \text{if } x \notin [-1,1],$$

$$(1-t/B)^{-1} \sum_{k=0}^\infty \frac{(-aB)^k q^{k(k-1)/2}}{(B/A;\, q)_{k+1}(q;\, q)_k} + (1-t/A)^{-1} \sum_{k=0}^\infty \frac{(-aA)^k q^{k(k-1)/2}}{(A/B;\, q)_{k+1}(q;\, q)_k} \quad \text{if } x \in (-1,1),$$

$$(1 \pm t)^{-2} \sum_0^\infty \frac{(\pm a)^k q^{k(k-1)/2}}{(q;\, q)_k^2} \quad \text{if } x = \mp 1.$$

The details are straightforward and will be omitted.

THEOREM 2.3. *The polynomials* $\{\theta_n^{(a)}(x;\, q)\}$ *satisfy the orthogonality relation*

$$\int_{-\infty}^\infty \theta_n^{(a)}(x;\, q)\theta_m^{(a)}(x;\, q)\, d\psi^{(a)}(x;\, q) = \delta_{m,n}, \tag{2.12}$$

where $d\psi$ *is a positive measure with bounded support. Furthermore the Stieltjes transform of* $d\psi$ *is given by*

$$\int_{-\infty}^\infty \frac{d\psi^{(a)}(t;\, q)}{x-t} = 2BF_{aq}(x)/F_a(x), \tag{2.13}$$

when $\operatorname{Im} x \neq 0$, *where*

$$F_a(x) := \sum_{k=0}^\infty \frac{(-aB)^k q^{k(k-1)/2}}{(B^2;\, q)_{k+1}(q;\, q)_k}. \tag{2.14}$$

Proof. The monic polynomials associated with the θ_n's are

$$p_n(x) := 2^{-n}\theta_n^{(a)}(x;\, q). \tag{2.15}$$

The p's satisfy

$$p_{n+1}(x)=\left(x-\frac{1}{2}aq^n\right)p_n(x)-\frac{1}{4}p_{n-1}(x), \tag{2.16}$$

so the coefficients in Chihara's notation [8, see Thm. 1.1] are given by

$$c_n=\frac{1}{2}aq^{n-1}, \quad n\geqq 1, \qquad \lambda_n=\frac{1}{4} \quad \text{for } n>1. \tag{2.17}$$

Combining Theorems 4.2 (p. 19) and 4.4 (p. 21) in Chihara [8], we establish (2.12). The boundedness of the support of $d\psi^{(a)}$ follows from Theorem 2.2 in [8, p. 109]. Now Markoff's theorem [8, pp. 89–90] implies (2.7) which when combined with (2.6) and the asymptotic formula (2.8) establishes (2.13). This completes the proof.

Note that (2.13) actually holds for x off the support of $d\psi^{(a)}$.

3. Absolutely continuous component of the spectral measure. We first compute the absolutely continuous component of $d\psi$. If $\{p_n(x)\}$ is a sequence of orthonormal polynomials,

$$\gamma_n := \text{coefficient of } x^n \text{ in } p_n(x),$$

then $\{p_n(x)\}$ satisfies a three term recursion of the type

$$xp_n(x)=\frac{\gamma_n}{\gamma_{n+1}}p_{n+1}(x)+\alpha_n p_n(x)+\frac{\gamma_{n-1}}{\gamma_n}p_{n-1}(x). \tag{3.1}$$

Combining Corollaries 36 (p. 141) and 40 (p. 143) in Nevai [12], we get

THEOREM 3.1 (Nevai). *If*

$$\sum_{k=1}^{\infty}\left\{|\alpha_k|+\left|\frac{\gamma_{k-1}}{\gamma_k}-\frac{1}{2}\right|\right\}<\infty \tag{3.2}$$

then

$$d\psi(t)=\psi'(t)\,dt+d\psi_j(t),$$

where ψ' is continuous and positive in $(-1,1)$, $\operatorname{supp}\psi'=[-1,1]$ *and ψ_j is a step function constant on $(-1,1)$. Furthermore*

$$\limsup_n \psi'(x)\sqrt{1-x^2}\,p_n^2(x)=\frac{2}{\pi}, \tag{3.3}$$

holds for almost every $x\in$ $\operatorname{supp} d\psi$.

In the case of the θ_n's, $\gamma_n=2^n$, $\alpha_n=\frac{1}{2}aq^n$, so (3.2) is satisfied. The θ's are orthonormal. The asymptotic formula (2.9) can be written in the form

$$\theta_n(\cos\theta;\, q)\sim 2\left|\sum_0^{\infty}\frac{\left(-ae^{i\theta}\right)^k q^{k(k-1)/2}}{\left(e^{2i\theta};\, q\right)_{k+1}(q;\, q)_k}\right|\cos(n\theta+\varphi), \tag{3.4}$$

holding for $0<\theta<\pi$, where

$$\varphi=\arg\left(\sum_{k=0}^{\infty}\frac{\left(-ae^{-i\theta}\right)^k q^{k(k-1)/2}}{\left(e^{-2i\theta};\, q\right)_{k+1}(q;\, q)_k}\right). \tag{3.5}$$

This analysis and Nevai's theorem establish the following.

THEOREM 3.2. *We have*

$$(3.6)\qquad \psi'(x)=\frac{2}{\pi}\sqrt{1-x^2}\left|\sum_{0}^{\infty}\frac{\left(-ae^{i\theta}\right)^k q^{k(k-1)/2}}{\left(qe^{2i\theta};\,q\right)_k(q;\,q)_k}\right|^{-2},\qquad x=\cos\theta,$$

holding for $q\in(-1,1)$.

We now analyse the discrete measure $d\psi_j$. Recall that

$$(3.7)\qquad F_a(x):=\sum_{j=0}^{\infty}\frac{(-aB)^j q^{j(j-1)/2}}{\left(B^2;\,q\right)_{j+1}(q;\,q)_j},$$

where A, B are as in (2.13). These functions, as functions of a, are essentially basic Bessel functions, see Al-Salam and Ismail [2] and Ismail [10]. Furthermore $F_a(x)$ and $F_{aq}(x)$ have no common zeros. A jump J in $\psi(t)$ at $t=\xi$ contributes $J(x-\xi)^{-1}$ to the left side of (2.13). Hence the location of the discrete masses coincides with the poles of the right side of (2.13), i.e. $F_{aq}(x)/F_a(x)$. These poles are the zeros of $F_a(x)$. Since computing the zeros of the transcendental function $F_a(x)$ seems impossible, we will make additional assumptions on a and q to guarantee that $F_a(x)$ does not vanish. Theorem 3.1 ensures the nonvanishing of $F_a(x)$ for $-1<x<1$. The zeros, if any, of $F_a(x)$ are all real since for x, y real, $y\neq 0$,

$$\left|\int_{-\infty}^{\infty}\frac{d\psi^{(a)}(t;\,q)}{x+iy-t}\right|\leqq\frac{1}{|y|}\int_{-\infty}^{\infty}d\psi^{(a)}(t;\,q)<\infty.$$

When x is real and lies outside $[-1,1]$ the quantities A, B will take real values and $0<B/A<1$ since $AB=1$. Observe that (3.7) implies the positivity of $F_a(x)$ if $a<0$ and $x>1$, $q>0$ because in this case

$$B=x-\sqrt{x^2-1}>0.$$

On the other hand when $a>0$ and $x<-1$ with q still positive B is $x+\sqrt{x^2-1}$, which is negative, and $F_a(x)$ is strictly positive. This means that when $q>0$ the zeros of $F_a(x)$, if any, are to the right of $[-1,1]$ if $a>0$ and to the left of $[-1,1]$ if $a<0$. In general for fixed real $x\notin[-1,1]$ and $q>0$ the function $F_a(x)$ as a function of a, has infinitely many zeros and all the zeros are real and simple, [2], [10]. So, without restricting a discrete masses will appear. In fact for every prescribed value of $x_0>1$ (<-1) of x and $q>0$ there are infinitely many values of $a>0$ (<0) that will make $F_a(x)$ vanish at $x=x_0$ and, of course, $F_{aq}(x_0)\neq 0$.

THEOREM 3.3 (Chihara [9]). *In order that* $d\psi_j\equiv 0$, *outside* $[-1,1]$, *for* $-1<q<1$ *it is necessary that*

$$(3.8)\qquad -2<a<2.$$

Proof. In Chihara's notation [8, p. 18] the recursion (1.6) is

$$(3.9)\qquad P_{n+1}(x)=\left(x-\frac{a}{2}q^n\right)P_n(x)-\frac{1}{4}P_{n-1}(x)$$

where

$$(3.10)\qquad P_n(x):=2^{-n}\theta_n^{(a)}(x;\,q),$$

so

$$(3.11)\qquad C_n=\frac{a}{2}q^{n-1},\qquad \lambda_{n+1}=\frac{1}{4}\qquad (n\geqq 1).$$

By Theorem 4.1 in [8, p. 122], we see that $\sigma=-1$, $\tau=1$, that is the zeros of $P_n(x)$ are dense in $(-1,1)$ confirming our earlier result that $\operatorname{supp}\psi'(t)=[-1,1]$. Thus, the true interval of orthogonality is $[\xi_1,\eta_1]\supset[-1,1]$. Theorem 2.1 and Corollary 1 in [8, p. 108] imply that if $[\xi_1,\eta_1]=[-1,1]$, i.e. ψ_j is constant outside $[-1,1]$, then $-1<c_n<1$ $(n\geqq 1)$. Clearly $-1<c_n<1$ is equivalent to (3.8). This completes the proof.

In particular there will be mass points <1 (>-1) if $a\leqq -2$ $(\geqq 2)$.

The results presented so far are valid for $-1<q<1$. In the rest of this section we shall restrict ourselves to the case

$$1>q>0. \tag{3.12}$$

The cases $q=0,1$ are essentially the Chebyshev polynomials.

We now prove the following.

THEOREM 3.4. *The points $t=\pm 1$ are not mass points when $1>q>0$ and $|a|\leqq(1-q)^2$.*

Proof. From the theory of moment problems [15, pp. 45–46] we know that ξ is a mass point if and only if

$$\sum_0^\infty\left[\theta_n^{(a)}(\xi;\ q)\right]^2<\infty, \tag{3.13}$$

since the θ's are orthonormal. The asymptotic formulas (2.10) and (2.11) imply that the above series diverged at $\xi=\pm 1$ if

$$\sum_0^\infty\frac{(-|a|)^k q^{k(k-1)/2}}{(q;\ q)_k^2}\neq 0. \tag{3.14}$$

We rewrite (3.14) in the form

$$\sum_0^\infty\frac{a^{2k}q^{k(2k-1)}}{(q;\ q)_{2k+1}^2}\left\{\left(1-q^{2k+1}\right)^2-|a|q^{2k}\right\}$$

which will be positive if $(1-q^{2k+1})^2-|a|q^{2k}$ is positive. Let w be q^{2k}, so $w\in[0,1]$. The function $(1-wq)^2-|a|w$ is positive when $w=0,1$ and its global minimum (on $(-\infty,\infty)$) is at $w=\frac{1}{2}(2q+|a|)q^{-2}$ which lies outside $[-1,1]$. This establishes the positivity of the series (3.14) when $|a|\leqq(1-q)^2$ and completes the proof.

For completeness we include the definition of a chain sequence.

DEFINITION. A sequence $\{a_n\}_1^\infty$, is a chain sequence if there is a parameter sequence $\{g_n\}$ such that $0\leqq g_0<1$, $0<g_n<1$, $(n\geqq 1)$ and $a_n=g_n(1-g_{n-1})$ $(n\geqq 1)$.

Chihara kindly communicated the following theorem and its proof to us.

THEOREM 3.5 (Chihara [9]). *A necessary condition for $\psi^{(a)}(t;\ q)$ to be continuous outside $[-1,1]$ is*

$$|a|<\left(1+q-\sqrt{1+q^2}\right)q^{-1}, \tag{3.15}$$

when $q\in(0,1)$.

Proof. Recall the notation (3.9), (3.10), (3.11). Theorem 2.1 and Corollary 1 in [8, p. 108] characterize the true interval of orthogonality to be $[-1,1]$ if and only if $-1<c_n<1$ and $\{\alpha_n(-1)\}$ and $\{\alpha_n(1)\}$ are both chain sequences, where

$$\alpha_n(x)=\lambda_{n+1}\left[(c_n-x)(c_{n+1}-x)\right]^{-1},$$

c_n, λ_n are as in (3.11). As we saw in Theorem 3.3, $-1<c_n<1$ is equivalent to (3.8). It is clear from (1.6) that

$$\theta_n^{(-a)}(-x;\ q)=(-1)^n\theta_n^{(a)}(x;\ q), \tag{3.16}$$

so in the rest of the proof we shall assume $0<a<2$, from (3.8). Clearly,

$$\alpha_n(-1)=\left[(2+aq^{n-1})(2+aq^n)\right]^{-1}\in\left(0,\frac{1}{4}\right),$$

so $\{\alpha_n(-1)\}$ is a chain sequence because it is dominated by the chain sequence $\{\frac{1}{4},\frac{1}{4},\cdots\}$, see [8, pp. 91, 97]. This implies that there are no mass points to the left of -1. Clearly $q>0$ implies

$$\alpha_n(1)=\frac{1}{(2-aq^{n-1})(2-aq^n)}>\frac{1}{4}.$$

Now $\{\frac{1}{2},\frac{1}{4},\frac{1}{4},\cdots\}$ is a chain sequence that determines its parameters uniquely, see [8, Exercise 5.1, p. 99], so no chain sequence can dominate it [8, p. 97]. Thus $\alpha_1(1)<\frac{1}{2}$ is necessary for $\psi^{(a)}(t;\, q)$ to be continuous to the right of 1. The condition $\alpha_1(1)<\frac{1}{2}$ is

$$qa^2-2(1+q)a+2>0,$$

which is equivalent to (3.15) and the proof is complete.

4. The singular component of the spectral measure. The purpose of the present section is to define conditions sufficient for the absence of the point spectrum, that is $\psi^{(a)}(x;\, q)$ is constant outside $[-1,1]$. Since the point spectrum coincides with the zeros of $F_a(x)$ we now study the function $F_a(x)$.

THEOREM 4.1. *When* $0<q<1$ *and*

$$|a|\leqq(1-q)^2, \tag{4.1}$$

the function $F_a(x)$ will have no zeros and the discrete spectrum will be empty.

Proof. We know that the zeros of $F_a(x)$, if any, are real and lie outside $(-1,1)$. The definition (3.7) of $F_a(x)$ clearly implies the positivity of F_a when $aB<0$, so we shall concentrate on the case $aB>0$. We express $F_a(x)$ in the form

$$F_a(x)=\sum_{j=0}^{\infty}\frac{(aB)^{2j}q^{j(2j-1)}}{(B^2;\, q)_{2j+2}(q;\, q)_{2j+1}} \cdot\left[(1-B^2q^{2j+1})(1-q^{2j+1})-aBq^{2j}\right]. \tag{4.2}$$

Observe that the positivity of $F_a(x)$ will follow from (4.2) if

$$(1-B^2q^{2j+1})(1-q^{2j+1})-aBq^{2j}\geqq 0,\qquad j=0,1,\cdots, \tag{4.3}$$

holds with strict inequality for at least one j. The cases $x\geqq 1$ and $x\leqq -1$ will be considered separately.

Case (i). $x\geqq 1$. In this case $B=x-\sqrt{x^2-1}\in(0,1]$ for finite $x\geqq 1$, and we need only to consider the case $a>0$. Clearly we have

$$1-B^2q^{2j+1}\geqq 1-q^{2j+1}; \tag{4.4}$$

hence (4.3) will hold if

$$(1-qy)^2-ay\geqq 0,\qquad 0\leqq y\leqq 1, \tag{4.5}$$

where $y=q^{2j}$. The polynomial on the left side of (4.5) has a local minimum at

$$y=\frac{1}{q}+\frac{a}{2q^2}$$

which is outside [0, 1]. Thus $(1-qy)^2-ay$ is monotone on [0, 1] and takes the value 1 at $y=0$. Therefore (4.5) holds for $y\in[0,1]$ if and only if it holds at $y=1$, that is

$$(1-q)^2-a\geqq 0,$$

which is (4.1).

Case (ii). $x\leqq -1$. Now B is $x+\sqrt{x^2-1}$, so $0>B\geqq -1$ and we need only to consider the case $a<0$. The analysis in case i goes through until (4.3). In (4.3) we replace aB by $|aB|$ then use $|B|\leqq 1$ and the rest will follow in the same way but a is now replaced by $|a|$.

THEOREM 4.2. *The function $F_a(x)$ does not vanish for x real when $0<q<1$ and*

$$|B|\leqq(1-q)^2/|a|. \tag{4.6}$$

Proof. The proof is similar to our proof of Theorem 4.1. In Case (i) we use (4.4) to replace (4.3) by

$$(1-qy)^2-aBy\geqq 0, \qquad y\in[0,1]. \tag{4.7}$$

In the proof of Theorem 4.1 we used $0<B\leqq 1$ to essentially replace (4.7) by the stronger condition (4.5). Repeating the same argument we establish (4.3) from (4.6). Case (ii) can be handled in a similar fashion.

COROLLARY 4.3. *If $|a|>(1-q)^2$, $0<q<1$ and $F_a(\xi)$ vanishes then*

$$|\xi|<\frac{d^2+1}{2d}, \tag{4.8}$$

where

$$d:=(1-q)^2/|a|. \tag{4.9}$$

Proof. If $\xi>0$ then $\xi>1$ and $B=\xi-\sqrt{\xi^2-1}>0$. The inequality (4.6) should be violated at ξ, so

$$\xi-\sqrt{\xi^2-1}>d.$$

Thus $\xi-d>\sqrt{\xi^2-1}$ which is equivalent to (4.8) because $\xi-d>0$ due to the fact $d<1<\xi$. The case $\xi<0$ can be treated similarly.

Observe that (4.8) provides a bound for the spectrum in general. We now treat the case $-1<q<0$. Define Q by

$$Q:=\frac{q^{-1}}{2}\left[1-q^2-\sqrt{(1-q^2)^2-4q}\right], \qquad 0>q>-1. \tag{4.10}$$

It is easy to see that $Q>0$.

THEOREM 4.4. *Let $0>q>-1$. The zeros, if any, of $F_a(x)$ lie outside the region*

$$|B|\leqq(1-q^2)Q/|a|. \tag{4.11}$$

In particular if

$$|a|\leqq(1-q^2)Q, \tag{4.12}$$

then $F_a(x)$ has no zeros and the discrete spectrum is empty.

Proof. There is no loss of generality in assuming $a > 0$ because

$$\theta_n^{(-a)}(-x;\, q) = (-1)^n \theta_n^{(a)}(x;\, q),$$

see (3.16), and the θ_n's reduce to Chebyshev polynomials when $a = 0$.

We express the series (3.7) defining $F_a(x)$ as the sum of four series according to whether the summation index k is $\equiv 0, 1, 2$ or $3 \pmod 4$. We then regroup the series in the form

$$F_a(x) = \sum_{j=0}^{\infty} \frac{(aB)^{4j} q^{2j(4j-1)} H_j}{(B^2;\, q)_{4j+4} (q;\, q)_{4j+3}}, \tag{4.13}$$

where

$$\begin{aligned} H_j := {} & \prod_{l=1}^{3} (1 - q^{4j+l})(1 - B^2 q^{4j+l}) \\ & - aBq^{4j} \prod_{l=2}^{3} (1 - q^{4j+l})(1 - B^2 q^{4j+l}) \\ & + a^2 B^2 q^{8j+1} (1 - q^{4j+3})(1 - B^2 q^{4j+3}) - a^3 B^3 q^{12j+3}. \end{aligned} \tag{4.14}$$

The cases $x \geqq 1$ and $x \leqq -1$ require separate treatment.

Case (i). $x \geqq 1$. Now $B = x - \sqrt{x^2 - 1} \in (0, 1]$, so $aB > 0$ and the last term in H_j is positive. From (4.14) we get

$$H_j > (1 - q^3)(1 - B^2 q^3) \varphi(q^{4j}), \tag{4.15}$$

if $\varphi(q^{4j}) > 0$, where

$$\varphi(y) = (1 - q^2)^2 - aB(1 - q^2)^2 y + a^2 B^2 q y^2. \tag{4.16}$$

The quadratic polynomial $\varphi(y)$ has a local maximum at the point

$$y = (2aBq)^{-1} (1 - q^2)^2 < 0,$$

and $\varphi(0) = (1 - q^2)^2 > 0$, hence $\varphi(y)$ is monotone on $(0, 1)$ and (4.13) and (4.15) will imply the positivity of $F_a(x)$ if we show that $\varphi(1) \geqq 0$. It is not difficult to see that $\varphi(1) \geqq 0$ is equivalent to (4.11).

Case (ii). $x \leqq -1$. In this case $B = x + \sqrt{x^2 - 1} \in [-1, 0)$ and $aB < 0$, so the first and second terms in H_j are positive while the third and fourth terms are negative. The sum of the first and third terms is bounded below by

$$(1 - q^{4j+3})(1 - B^2 q^{4j+3}) \left[(1 - q^2)^2 + a^2 B^2 q \right]$$

which is nonnegative if

$$-aB \leqq (1 - q^2)/\sqrt{-q}\,. \tag{4.17}$$

Similarly we prove that (4.17) implies the nonnegativity of the sum of the second and fourth terms in H_j. Finally observe that the condition (4.17) is weaker than (4.11), so (4.11) implies the positivity of $F_a(x)$ in all cases.

COROLLARY 4.5. *If* $|a|>(1-q^2)Q$, $0>q>-1$ *and* $F(\xi)=0$ *then*

$$|\xi|<\frac{c^2+1}{2c}, \tag{4.18}$$

and

$$c=(1-q^2)Q/|a|. \tag{4.19}$$

The proof of Corollary 4.5 is very similar to the proof of Corollary 4.3 and will be omitted.

5. Explicit representation and linearization of products. The present section contains two additional results concerning the θ_n's. The first is the following explicit representation for $\theta_n^{(a)}(\cos\theta;\,q)$ as a trigonometric polynomial of degree n.

THEOREM 5.1. *We have*

(5.1)
$$\theta_n^{(a)}(\cos\theta;\,q)=\sum_{\substack{j,k\geqq 0\\ j+k\leqq n}}\frac{q^{k(k-1)/2}(q;\,q)_{n-j}(q;\,q)_{j+k}}{(q;\,q)_k^2(q;\,q)_j(q;\,q)_{n-k-j}}(-a)^k\cos[(n-k-2j)\theta].$$

Proof. We use the q-binomial theorem

$$\frac{(az;\,q)_\infty}{(z;\,q)_\infty}=\sum_0^\infty\frac{(a;\,q)_n}{(q;\,q)_n}z^n$$

(Slater [16, p. 92]) to expand $(t/A;\,q)_{k+1}^{-1}$ and $(t/B;\,q)_{k+1}^{-1}$ in the generating function (2.1) in the form

$$(t/c;\,q)_{k+1}^{-1}=(q^{k+1}t/c;\,q)_\infty/(t/c;\,q)_\infty=\sum_{j=0}^\infty\frac{(q^{k+1};\,q)_j}{(q;\,q)_j}t^jc^{-j}$$

$$=\sum_{j=0}^\infty\frac{(q;\,q)_{k+j}}{(q;\,q)_k(q;\,q)_j}t^jc^{-j}.$$

This identity, the substitutions $A=e^{i\theta}$, $B=e^{-i\theta}$ and some simple manipulations establish (5.1) and the proof is complete.

In the process of proving Theorem 5.1, we essentially proved the following.

COROLLARY 5.2. *The* θ_n*'s have the explicit representation*

$$\theta_n^{(a)}(x;\,q)=\sum_{\substack{k,l,m\geqq 0\\ k+l+m=n}}\frac{(-a)^kq^{k/(k-1)/2}(q;\,q)_{k+l}(q;\,q)_{k+m}}{(q;\,q)_k^2(q;\,q)_l(q;\,q)_m}A^{-l}B^{-m}, \tag{5.2}$$

where A and B are as in (2.3).

The second result in this section concerns coefficients in the linearization of a product of two θ_n's as a sum of θ_n's. We shall show that these linearization coefficients are nonnegative. Our proof of the nonnegativity of the linearization coefficients depends on the following key lemma of Askey [3].

LEMMA 5.3. *Let* $\{P_n(x)\}$ *be a sequence of monic polynomials, that is*

$$P_n(x)=x^n+\text{a polynomial of degree at most } n-1,$$

that satisfies

$$P_1(x)P_n(x)=P_{n+1}(x)+a_nP_n(x)+b_nP_{n-1}(x). \tag{5.3}$$

If $a_n \geqq 0$, $b_n > 0$ *and* $a_{n+1} \geqq a_n$, $b_{n+1} \geqq b_n$, *then*

$$P_n(x)P_m(x) = \sum_{k=|m-n|}^{m+n} \alpha_k P_k(x) \tag{5.4}$$

with $\alpha_k \geqq 0$.

THEOREM 5.4. *The coefficients* $\alpha(k,m,n)$ *in*

$$\theta_m^{(a)}(x;\ q)\theta_n^{(a)}(x;\ q) = \sum_{k=|m-n|}^{m+n} \alpha(k,m,n)\theta_k^{(a)}(x;\ q)$$

are nonnegative for $a \leqq 0$.

Proof. The associated monic set is $P_n(x) := 2^{-n}\theta_n^{(a)}(x;\ q)$. Hence $P_1(x) = x - \frac{1}{2}a$. Therefore

$$P_1(x)P_n(x) = P_{n+1}(x) + \frac{1}{4}P_{n-1}(x) - \frac{a}{2}(1-q^n)P_n(x),$$

since the P's satisfies (3.9), i.e. $b_n = \frac{1}{4}$, $a_n = -\frac{a}{2}(1-q^n)$. If $a \leqq 0$ then $a_n \geqq 0$ and $a_{n+1} \geqq a_n$, so by Lemma 5.3 we establish the nonnegativity of the linearization coefficients in this case.

COROLLARY 5.5. *If* $a > 0$ *then* $(-1)^{k+m+n}\alpha(k,m,n) \geqq 0$.

Proof. This follows from

$$(-1)^n\theta_n^{(a)}(-x;\ q) = \theta_n^{(a)}(x;\ q),$$

see (3.16).

6. The case $|q| > 1$. In this case the generating function (2.1) no longer holds but the explicit formula (5.2) remains valid because both sides of (5.2) are well defined for $q > 1$ or $q < -1$ and computing any θ_n involves only a finite number of steps. We now determine the asymptotic behavior of θ_n for large n and fixed x. We set

$$p := 1/q. \tag{6.1}$$

THEOREM 6.1. *The following asymptotic formula*

$$\theta_n^{(a)}(x;\ q) \sim (-a)^n q^{n(n-1)/2}\left(\frac{A}{a};\ p\right)_\infty \sum_{k=0}^{\infty} \frac{p^{k(k-1)/2}(-B/a)^k}{(p;\ p)_k(A/a;\ p)_k}, \tag{6.2}$$

holds for x *off the support of* $d\psi^{(a)}$.

Proof. We replace q in (5.2) by $1/p$, use the observation

$$(q;\ q)_l = (-1)^l q^{l(l+1)/2}(p;\ p)_l,$$

and replace the summation index k in (5.2) by $n-k$ to obtain

(6.3)

$$(-a)^{-n}q^{-n(n-1)/2}\theta_n^{(a)} = \sum_{0\leqq l\leqq k\leqq n} \frac{p^{k(k-1)/2}(-a)^{-k}(p;\ p)_{n+l-k}(p;\ p)_{n-l}A^{-l}B^{l-k}}{(p;\ p)_l(p;\ p)_{k-l}(p;\ p)_{n-k}^2}.$$

We now apply Tannery's theorem, a discrete and earlier version of the Lebesgue dominated convergence theorem, see Bromwich [6], to the right side of (6.3). The pointwise limit exists and since the infinite product $(p;\ p)_\infty$ exists the right side of (6.3) is bounded by a constant multiple of

$$\sum_{k,l=0}^{\infty} \frac{p^{(k+l)(k+l-1)/2}|A|^{-l}|B|^{-k}|a|^{-k-l}}{(p;\ p)_k(p;\ p)_l}$$

which is a convergent double series. Thus

$$\lim_{n\to\infty}(-a)^{-n}p^{n(n-1)/2}\theta_n^{(a)}(x;\ a)=\sum_{l=0}^{\infty}\frac{(-aA)^{-l}p^{l(l-1)/2}}{(p;\ p)_l}\sum_{k=0}^{\infty}\frac{p^{k(k-1)/2}(-aBp^{-l})^{-k}}{(p;\ p)_k}.$$

The inner sum can be evaluated by Euler's formula

$$\sum_{0}^{\infty}\frac{(-x)^n q^{n(n-1)/2}}{(q;\ q)_n}=(x;\ q)_\infty,$$

Slater [16, p. 93], and we get

$$(-a)^{-n}p^{n(n-1)/2}\theta_n^{(a)}(x;\ q)\sim\sum_{l=0}^{\infty}\frac{(-aA)^{-l}p^{l(l-1)/2}}{(p;\ p)_l}\left(\frac{p^l}{aB};\ p\right)_\infty,$$

which can be easily reduced to (6.2) since $AB=1$ and $(\sigma p^l;\ p)_\infty$ is nothing but $(\sigma;\ p)_\infty/(\sigma;\ p)_l$. This completes the proof.

We define

$$G(r,s):=(s;\ p)_\infty\sum_{k=0}^{\infty}\frac{p^{k(k-1)/2}r^k}{(p;\ p)_k(s;\ p)_k}. \tag{6.4}$$

THEOREM 6.2. *There exists a unique measure $d\psi^{(a)}(x;\ q)$ such that*

$$\int_{-\infty}^{\infty}\theta_n^{(a)}(x;\ q)\theta_m^{(a)}(x;\ q)\,d\psi^{(a)}(x;\ q)=\delta_{mn}. \tag{6.5}$$

The support of $d\psi^{(a)}$ is unbounded and the Stieltjes transform of $d\psi^{(a)}$ is given by

$$\int_{-\infty}^{\infty}\frac{d\psi^{(a)}(t;\ q)}{x-t}=-2a^{-1}G\left(-\frac{B}{a}q,\ \frac{A}{a}\right)\Big/G\left(-\frac{B}{a},\ \frac{A}{a}\right), \tag{6.6}$$

valid for $x\notin\operatorname{supp}\{d\psi^{(a)}\}$. Furthermore $d\psi^{(a)}$ is singular with respect to the Lebesgue measure, that is its absolutely continuous component vanishes almost everywhere.

Proof. The existence of $d\psi^{(a)}$ follows from Theorem 1.1 because the positivity condition $\lambda_n>0$, (2.15) and (2.17) remain valid when $|q|>1$. In the terminology of the moment problem, [15], the uniqueness of $d\psi^{(a)}$ is called the determinacy of the moment problem. Theorem 2.9 in Shohat and Tamarkin [15, p. 50] insures the uniqueness of $d\psi^{(a)}$ if the series

$$\sum_{0}^{\infty}\left|\theta_n^{(a)}(x;\ q)\right|^2 \tag{6.7}$$

diverges at one point x, real or complex. The existence of an x that makes the series (6.7) diverge when $a\neq 0$ is clear from (6.2) because $G(-B/a,\ A/a)$ does not vanish identically and $|q|>1$. This proves (6.5). The polynomials of the second kind are $\{\theta_{n-1}^{(aq)}(x;\ q)\}_1^\infty$, so the continued fraction

$$\frac{2|}{|2x-a}-\frac{2|}{|2x-aq}-\cdots-\frac{2|}{|2x-aq^n}-\cdots \tag{6.8}$$

converges to the left side of (6.6), see the discussion in [15, p. 46]. On the other hand the continued fraction (6.8) is

$$\lim_{n\to\infty} \theta_{n-1}^{(aq)}(x,q)/\theta_n^{(a)}(x;\,q)$$

which, in view of (6.2), is the right side of (6.6). It only remains to show that $d\psi^{(a)}$ is singular. Recall that (Stone [17])

$$F(z)=\int_{-\infty}^{\infty}\frac{d\mu(t)}{z-t},\qquad z\notin\{\operatorname{supp} d\mu\} \tag{6.9}$$

if and only if

$$\mu(t_2)-\mu(t_1)=\lim_{\varepsilon\to0^+}\frac{1}{2\pi i}\int_{t_1}^{t_2}\{F(t-i\varepsilon)-F(t+i\varepsilon)\}\,dt, \tag{6.10}$$

where μ is normalized by $\mu(-\infty)=0$, $\mu(t)=\frac{1}{2}\{\mu(t+)+\mu(t-)\}$. Therefore the absolutely continuous component of $d\mu$ is

$$\mu'(t)=\frac{1}{2\pi i}\{F(t-i0)-F(t+i0)\}\quad \text{a.e.} \tag{6.11}$$

so if $F(z)$ is single-valued across the x-axis μ' will vanish almost everywhere. The functions $G(-B/a,\ A/a)$ appearing in the right side of (6.6) is

$$\sum_{k,l=0}^{\infty}\frac{p^{(k+l)(k+l-1)/2}}{(p;\,p)_k(p;\,p)_l}\left(-\frac{B}{a}\right)^l\left(-\frac{A}{a}\right)^k,$$

hence both $G(-B/a,\ A/a)$ and $G(-Bp/a,\ A/a)$ are uniform limits of polynomials symmetric in A, B. We now show that the polynomials symmetric in A and B are single-valued functions of x. Any such polynomial is a sum of terms of the type

$$a_{mn}\{A^mB^n+B^mA^n\},\qquad m\geqq n.$$

But

$$A^mB^n+B^mA^n=(AB)^n(A^{m-n}+B^{m-n})=A^{m-n}+B^{m-n},$$

since $AB=1$. Therefore

$$\begin{aligned}A^mB^n+B^mA^n&=(A+B)^{m-n}-\sum_{j=1}^{m-n-1}\binom{m-n}{j}A^jB^{m-n-j}\\&=(A+B)^{m-n}-\sum_{j=1}^{m-n-1}\binom{m-n}{j}A^{j-1}B^{m-n-j-1},\end{aligned}$$

where we used $AB=1$. Recall that $A+B=2x$. The above relationship expresses $B^mA^n+B^nA^m$ as $(2x)^{m-n}$ plus some symmetric lower degree polynomial. Repeating this process shows that $A^mB^n+B^mA^n$ is a polynomial in x of degree $m-n$; hence is single-valued. Therefore both $G(-B/a,\ A/a)$ and $G(-Bq/a,\ A/a)$ are entire functions of x, hence single-valued across the x-axis. This completes the proof.

The idea of using Tannery's theorem to derive asymptotic expansions for certain polynomials has been used successfully in other cases, see [2], [10] and [11].

7. The case $q=-1$. When $q=-1$ the generating function

$$G(x,t):=\sum_{0}^{\infty}t^n\theta_n^{(a)}(x;\ -1), \tag{7.1}$$

satisfies the functional equation, see (2.4),

$$(1-2xt+t^2)G(x,t)=1-at\,G(x,-t). \tag{7.2}$$

We now iterate (7.2), that is replace t by $-t$, then use the result and (7.2) to eliminate $G(x,-t)$. The result is

$$\sum_{0}^{\infty}t^n\theta_n^{(a)}(x;\ -1)=\{1+(2x-a)t+t^2\}\{(1+t^2)^2+(a^2-4x^2)t^2\}^{-1}. \tag{7.3}$$

THEOREM 7.1. *For fixed x, let α, β be the roots of*

$$(1+u)^2+(a^2-4x^2)u=0 \quad \textit{with } |\beta|\leqq|\alpha|. \tag{7.4}$$

The polynomials $\{\theta_n^{(a)}(x;\ -1)\}$ can be expressed explicitly as

$$\begin{aligned}\theta_{2n}^{(a)}(x;\ -1)&=(\alpha-\alpha)^{-1}[(\alpha+\beta)\alpha^n-(\beta+1)\beta^n],\\ \theta_{2n+1}^{(a)}(x;\ -1)&=(2x-a)(\alpha-\beta)^{-1}(\alpha^{n+1}-\beta^{n+1}),\end{aligned}\qquad n=0,1,2,\cdots. \tag{7.5}$$

Proof. Rewrite the right side of (7.3) in the form

$$\frac{1}{\alpha-\beta}\left[\frac{\alpha+1}{1-t^2\alpha}-\frac{\beta+1}{1-t^2\beta}+(2x-a)t\left(\frac{\alpha}{1-t^2\alpha}-\frac{\beta}{1-t^2\beta}\right)\right],$$

then expand $(1-t^2\alpha^{-1})$ and $(1-t^2\beta^{-1})$ in powers of t^2 and equate coefficients of like powers of t. A simple calculation yields (7.5) and the proof is complete.

THEOREM 7.2. *The polynomial $\{\theta_n^{(a)}(x;\ -1)\}$ satisfies the orthogonality relation*

$$\int_E\left[\frac{2x+a}{2x-a}\left(1+\frac{a^2}{4}-x^2\right)\right]^{1/2}\theta_n^{(a)}(x;\ -1)\theta_m^{(a)}(x;\ -1)\,dx=\pi\delta_{m,n}, \tag{7.6}$$

where E is given by

$$E=\left[-\frac{1}{2}\sqrt{4+a^2},\ -\frac{|a|}{2}\right]\cup\left[\frac{|a|}{2},\ \frac{1}{2}\sqrt{a^2+4}\right]. \tag{7.7}$$

Proof. The corresponding continued fraction $\chi(x)$ is periodic and satisfies

$$\chi(x)=\frac{1}{2x-a-1/(2x+a-\frac{1}{2}\chi(x))}.$$

Therefore

$$\chi(x)=\int_{-\infty}^{\infty}\frac{d\psi^{(a)}(t,-1)}{x-t}=2(\beta+1)(2x-a)^{-1}. \tag{7.8}$$

We now invert (7.8) to compute $d\psi^{(a)}$. It is clear from (7.8) that $d\psi^{(a)}$ is absolutely continuous because $\beta=-1$ when $x=\frac{1}{2}a$. We now apply (6.9), (6.10) and (6.11) to get

$$\frac{d}{dx}\psi^{(a)}(x,-1)=\lim_{\varepsilon\to0^+}2(\beta_1-\beta_2)\frac{(2\pi i)^{-1}}{2x-a}, \tag{7.9}$$

where β_1 and β_2 are the values of β when $x = x - i0$ and $x = x + i0$, respectively. It is not difficult to see from (7.8) and (7.4) that $\psi^{(a)}$ is constant outside E and

$$(7.10) \quad \frac{d}{dx}\psi^{(a)}(x, -1) = \left\{4 - (2 + a^2 - 4x^2)^2\right\}^{1/2}\left\{\pi(2x - a)\right\}^{-1}, \qquad x \in E,$$

which implies the orthogonality relation (7.6) because $\psi^{(a)}$ is normalized by

$$\int_{-\infty}^{\infty} d\psi^{(a)}(x; -1) = 1.$$

Remark 7.3. The fact that $d/dx\psi^{(a)}(x; -1)$ vanishes outside E is predicted in Chihara [8, Thm. 4.1, p. 122].

Observe that as $a \to 0$ the orthogonality relation (7.6) reduces to the orthogonality relation for $\{U_n(x)\}$. The results obtained here also follow from Chihara's work [7].

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem*, Hafner, New York, 1965.

[2] W. AL-SALAM AND M. E. H. ISMAIL, *Orthogonal polynomials associated with the Rogers-Ramanujan continued fraction*, Pacific J. Math., 105 (1983), pp. 269–283.

[3] R. ASKEY, *Linearization of the products of orthogonal polynomials*, in Problems in Analysis, R. C. Grunning, ed., Princeton Univ. Press, Princeton, NJ, 1970, pp. 223–238.

[4] R. ASKEY AND M. E. H. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Memoirs Amer. Math. Soc. 300, 1984.

[5] J. AVRON AND B. SIMON, *Singular continuous spectrum for a class of almost periodic Jacobi matrices*, Bull. Amer. Math. Soc. N.S., 6 (1982), pp. 81–86.

[6] T. J. I. A. BROMWICH, *An Introduction to the Theory of Infinite Series*, second edition, Macmillan, New York, 1955.

[7] T. S. CHIHARA, *On kernel polynomials and related systems*, Bull. U. M.I., 19 (1964), pp. 451–459.

[8] ______, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[9] ______, private communication.

[10] M. E. H. ISMAIL, *The zeros of basic Bessel functions, the functions $J_{\nu+ax}(x)$ and associated orthogonal polynomials*, J. Math. Anal. Appl., 86 (1982), pp. 1–19.

[11] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and ${}_4\phi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54.

[12] P. G. NEVAI, *Orthogonal polynomials*, Memoirs Amer. Math. Soc. 213, 1979.

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

[14] F. POLLACZEK, *Sur une géneralisation des polynômes de Jacobi*, Memorial des Sciences Mathématiques, Volume 131, 1956.

[15] J. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, revised edition, Mathematical Surveys, No. 1, American Mathematical Society, Providence, RI, 1963.

[16] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[17] M. H. STONE, *Linear Transformations in Hilbert Spaces*, Colloquium Publications, Volume 15, American Mathematical Society, New York, 1932.

[18] G. SZEGÖ, *Orthogonal Polynomials*, fourth edition, Colloquium Publications, Volume 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

020

GENERALIZED FEJÉR AND LANCZOS KERNELS*

M. CLUTTON-BROCK[†]

Abstract. On a circle, the Fejér kernel is nonnegative and is useful for approximating positive densities, and the Lanczos kernel is an approximate identity and gives rapid convergence when applied to discontinuous functions. These kernels are easy to construct from trigonometric functions because the Dirichlet kernel is invariant under translations. This paper uses the invariance under rotations of the Dirichlet kernel constructed from spherical harmonics to construct Fejér and Lanczos kernels on a sphere. On a $(q+2)$-dimensional sphere, the kth order Lanczos kernel is an approximate identity for $q<2k$. The corresponding Fejér and Lanczos kernels in Euclidean space can be constructed by a simple mapping.

Key words. kernels, approximate identity, convergence

AMS(MOS) subject classifications. Primary 41, 33

1. Introduction. One sometimes wants to approximate a function which is everywhere nonnegative. For example, Smoothed Particle Hydrodynamics [5] fits a mass density to discrete masses at points $x_1..x_J$, and in statistical estimation the kernel method [12] fits a frequency density to a sample $x_1..x_J$. These methods generally use something like a Gaussian for a nonnegative kernel, but a more economic representation is obtained if the kernel is constructed from a set $\{\phi_n\}$ of orthonormal functions:

$$K_N(x,y)=\sum_{n=0}^{N}\sigma_n\phi_n(x)\phi_n(y). \tag{1.1}$$

The problem is, how do we choose the sigma factors σ_n so that the kernel (1.1) is everywhere nonnegative?

The best known nonnegative kernel is the Fejér kernel $F_N(x,y)$ constructed from trigonometric functions with

$$\sigma_n=1-n(N+1). \tag{1.2}$$

It can be derived from the Dirichlet kernel either as the $(C,1)$ mean or by squaring the Dirichlet kernel of order $N/2$ and normalizing. These methods work essentially because the trigonometric functions are invariant under translation so that the Dirichlet kernel $D_N(x,y)$ is a function of $x-y$ only. Now spherical harmonics are invariant under rotations, so the Dirichlet kernel $D_N(A,B)$ on a sphere is a function only of the angle $\widehat{AB}$. Because of this it is possible to construct a Fejér kernel on the sphere, as we see in §3. How to do this using Cesàro means is already known: the $(C,q+1)$ mean of the Dirichlet kernel on a $(q+2)$-dimensional sphere is everywhere nonnegative (Kogbetlianz [9], Askey and Pollard [3]). For even N, however, the $(C,q+1)$ mean is everywhere greater than zero and is not the sharpest possible kernel. It may therefore be better for even N to use the kernel obtained by squaring the Dirichlet kernel of order $N/2$ and normalizing, and in §3 we derive the corresponding sigma factors.

In practical problems we more often want kernels in Euclidean space. It is easy, however, to map Euclidean space E_{q+1} onto the surface S_{q+1} of a $(q+2)$-dimensional sphere; we discuss this briefly in §6. The case $q=2$ is especially important because it

*Received by the editors June 4, 1982, and in final revised form May 15, 1985. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

[†] Department of Mathematics and Astronomy, University of Manitoba, Winnipeg R3T 2N2, Canada.

corresponds to 3-dimensional Euclidean space, and fortunately the ultraspherical polynomials from which the Dirichlet kernel on a 4-dimensional sphere is constructed are especially simple.

The Fejér kernel has many desirable properties apart from being nonnegative: it captures all continuous functions, and it avoids the notorious Gibbs oscillations. The Gibbs oscillations induced when the Dirichlet kernel is applied to a discontinuous function get worse as the number of dimensions increases, and in 3 dimensions they can be very serious indeed. Of course, if the function we are approximating is smooth enough, there is no trouble, and the Dirichlet kernel will converge rapidly. But in practical problems functions are not always smooth; compressible flows develop discontinuities, and in particle methods of simulating fluids and plasmas the data is noisy. In these circumstances some form of smoothing is essential.

The Fejér kernel, however, may provide too much smoothing. The Nth order Fejér kernel is about as sharp as the Dirichlet kernel of order $N/2$, and in 3 dimensions the Fejér kernel needs about 8 times as many terms to achieve the same sharpness as the Dirichlet kernel. We therefore want a kernel which is better behaved than the Dirichlet kernel but is sharper than the Fejér kernel.

Lanczos [10], [11] showed how to construct a kernel for trigonometric functions that gives rapid convergence when applied to a discontinuous function. The Lanczos kernel $L_N(x,y)$ is obtained from the Dirichlet kernel $D_N(x,z)$ by integrating over a window in z:

$$L_N(x,y)=\frac{1}{\Delta}\int_{y-\Delta/2}^{y+\Delta/2} D_N(x,z)\,dz. \tag{1.3}$$

In other words, the Lanczos kernel is a moving average of the Dirichlet kernel. This provides the same amount of smoothing for all x because the Dirichlet kernel is a function of $x-z$ only. Because the Dirichlet kernel $D_N(A,B)$ on a sphere is a function of the angle $\widehat{AB}$ only, we can also construct a Lanczos kernel on the sphere, and we derive the Lanczos sigma factors in §4.

The Lanczos kernel constructed from trigonometric functions is well behaved because it is an approximate identity, that is, it fulfills the following conditions:

$$\int_{\text{all } y} L_N(x,y)\,dy=1, \tag{1.4}$$

$$\sup_N \int_{\text{all } y} |L_N(x,y)|\,dy<\infty, \tag{1.5}$$

$$\lim_{N\to\infty} \int_{|x-y|>\varepsilon} |L_N(x,y)|\,dy=0. \tag{1.6}$$

An approximate identity captures all continuous functions and avoids Gibbs oscillations. Many of the troubles of the Dirichlet kernel stem from the fact that it is not an approximate kernel. On a $(q+2)$-dimensional sphere, for example, we have

$$\int_{\text{all } B} |D_N(A,B)|\,d\Omega(B)=\begin{cases} O(\ln N) & \text{for } q=0 \text{ (a circle)},\\ O(N^{q/2}) & \text{for } q>0, \end{cases} \tag{1.7}$$

which shows how a Dirichlet kernel gets worse behaved and farther from being an approximate identity as the number of dimensions increases.

In §5 we show that the Lanczos kernel is an approximate identity for $q=1$ (the 2-dimensional surface of a 3-dimensional sphere), but not for $q=2$. To obtain an approximate identity for $q \geq 2$, we must use higher order Lanczos kernels, which are obtained by repeated Lanczos averaging, that is, by repeated integration over the window. The second order Lanczos kernel is an approximate identity for $q=2$ or 3, and the kth order Lanczos kernel is an approximate identity for $q<2k$.

For ease of reference and to define the notation §2 gives the most important formulae involving spherical harmonics in $(q+2)$-dimensions. An excellent account is given by Hochstadt [6].

2. Notation and useful formulae for spherical harmonics. In terms of the angular coordinates $(\theta_0, \theta_1 \cdots \theta_q)$ the element of area of a $(q+2)$-dimensional sphere can be written

$$d\Omega_q = (\sin\theta_0)^q d_0 \times (\sin\theta_1)^{q-1} d\theta_1 \times \cdots \times d\theta_q. \tag{2.1}$$

This suggests representing a $(q+2)$-dimensional spherical harmonic in terms of ultraspherical polynomials $C(\begin{smallmatrix}\lambda\\ n\end{smallmatrix}|\cos\theta)$ orthogonal under the weight function $(\sin\theta)^{2\lambda}$

$$\begin{aligned} &Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|\theta_0\theta_1..\theta_q\right) \\ &= \left\{H\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\right)\right\}^{-1} C\left(\begin{matrix}k_1+q/2\\ n-k_1\end{matrix}\middle|\cos\theta_0\right)(\sin\theta_0)^{k_1} \times C\left(\begin{matrix}k_2+(q-1)/2\\ k_1-k_2\end{matrix}\middle|\cos\theta_1\right)(\sin\theta_1)^{k_2} \\ &\qquad \times \cdots \times C\left(\begin{matrix}|k_q|+1/2\\ k_{q-1}-|k_q|\end{matrix}\middle|\cos\theta_{q-1}\right)(\sin\theta_{q-1})^{|k_q|} \times \exp(ik_q\theta_q), \end{aligned} \tag{2.2}$$

where H is the normalizing integral

$$H\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\right) = h\left(\begin{matrix}k_1+q/2\\ n-k_1\end{matrix}\right) \times h\left(\begin{matrix}k_2+(q-1)/2\\ k_1-k_2\end{matrix}\right) \times \cdots \times h\left(\begin{matrix}|k_q|+1/2\\ k_{q-1}-|k_q|\end{matrix}\right) \times 2\pi, \tag{2.3}$$

and h is the normalizing integral for ultraspherical polynomials

$$h\left(\begin{matrix}\lambda\\ n\end{matrix}\right) = \int_0^\pi \left\{C\left(\begin{matrix}\lambda\\ n\end{matrix}\middle|\cos\theta\right)\right\}^2 (\sin\theta)^{2\lambda} d\theta = \frac{C\left(\begin{matrix}\lambda\\ n\end{matrix}\middle|1\right)}{(1+n/\lambda)} h\left(\begin{matrix}\lambda\\ 0\end{matrix}\right) \tag{2.4}$$

with

$$h\left(\begin{matrix}\lambda\\ 0\end{matrix}\right) = \frac{\pi 2^{1-2\lambda}\Gamma(2\lambda)}{\lambda\{\Gamma(\lambda)\}^2}, \tag{2.5}$$

and the ultraspherical polynomials are standardised so that

$$C\left(\begin{matrix}\lambda\\ n\end{matrix}\middle|1\right) = \left(\begin{matrix}n+2\lambda-1\\ n\end{matrix}\right). \tag{2.6}$$

The Dirichlet kernel on a sphere is

$$D_N(A,B) = \sum_{n=0}^{N} \sum_{k_1 \cdots k_q} Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|A\right) Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|B\right). \tag{2.7}$$

The sum over $k_1..k_q$ is invariant under rotations [6], and on rotating B to the pole $(0,0..0)$ and A to the point $(\widehat{AB},0..0)$, all terms vanish except those with $k_1..k_q=0$. We therefore obtain the addition formula

(2.8)
$$\sum_{k_1\cdots k_q} Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|A\right)Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|B\right)=Y\left(\begin{matrix}0..0\\ n\end{matrix}\middle|\widehat{AB},0..0\right)Y\left(\begin{matrix}0..0\\ n\end{matrix}\middle|0,0..0\right)$$
$$=\left\{H\begin{pmatrix}0..0\\ n\end{pmatrix}\right\}^{-1}C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|1\right)C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\widehat{AB}\right).$$

Recalling (2.3) and (2.4), we have

$$H\begin{pmatrix}0..0\\ n\end{pmatrix}=\Omega_q C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|1\right)\Big/(1+2n/q), \tag{2.9}$$

where

$$\Omega_q=h\begin{pmatrix}q/2\\ 0\end{pmatrix}\times h\begin{pmatrix}(q-1)/2\\ 0\end{pmatrix}\times\cdots\times h\begin{pmatrix}1/2\\ 0\end{pmatrix}\times 2\pi \tag{2.10}$$

is the surface area of a $(q+2)$-dimensional sphere. The Dirichlet kernel is therefore

$$D_N(A,B)=\Omega_q^{-1}\sum_{n=0}^{N}(1+2n/q)C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\widehat{AB}\right). \tag{2.11}$$

Our task is now to find the sigma factors in the Fejér and Lanczos kernels which take the form

$$K_N(A,B)=\sum_{n=0}^{N}\sigma_n\sum_{k_1..k_q}Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|A\right)Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|B\right) \tag{2.12}$$
$$=\Omega_q^{-1}\sum_{n=0}^{N}\sigma_n(1+2n/q)C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\widehat{AB}\right).$$

3. The Fejér sigma factors. The simplest way to obtain a nonnegative kernel is to use Cesàro means. Recall that the (C,β) mean of Σt_n is

$$S_N(C,\beta;\,t)=\sum_{n=0}^{N}\frac{[N]_n}{[N+\beta]_n}t_n \tag{3.1}$$

where $[\alpha]_n$ is the nth descending factorial of α. Kogbetlianz [9] first stated that the $(C,2\lambda+1)$ means of

$$\sum_n(1+n/\lambda)C\left(\begin{matrix}\lambda\\ n\end{matrix}\middle|x\right) \tag{3.2}$$

are nonnegative. Askey and Pollard [3] gave a simple proof, which is also outlined in Askey's book [1]. Now (3.2) is simply proportional to the Dirichlet kernel (2.11) with $q=2\lambda$, so it is clear that the $(C,q+1)$ means of the Dirichlet kernel should be nonnegative. The corresponding Fejér sigma factors are

$$\sigma_n=\frac{[N]_n}{[N+q+1]_n}. \tag{3.3}$$

It is desirable that the Fejér kernel should be as sharp as possible consistent with being nonnegative. The smallest value of β which will make the (C,β) mean of the Dirichlet kernel nonnegative is $q+1$ for odd N, but less than $q+1$ for even N. For example, with $q=2$, $N=6$, the (C,β) mean of the Dirichlet kernel is positive for

$$\beta=2.28<q+1=3. \tag{3.4}$$

Thus, for even N the optimum value of β which will give the sharpest nonnegative Fejér kernel is less than $q+1$. I would like to be able to find the optimum value of β by using the elegant methods in Chapters 8 and 9 of Askey's book [1], or perhaps to extend Askey's results [2] for positive radial functions on a sphere, but I am not clever enough.

For even N we can obtain another nonnegative kernel by squaring the Dirichlet kernel of order $N/2$ and normalizing. The square of the Dirichlet kernel (2.11) is

$$\begin{aligned}\{D_{N/2}(A,B)\}^2=\Omega_q^{-2}\sum_{n,m=0}^{N/2}&(1+2n/q)(1+2m/q)\\ &\times C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\widehat{AB}\right)C\left(\begin{matrix}q/2\\ m\end{matrix}\middle|\cos\widehat{AB}\right).\end{aligned}\tag{3.5}$$

The product of two ultraspherical polynomials can be linearized:

$$C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|x\right)C\left(\begin{matrix}q/2\\ m\end{matrix}\middle|x\right)=\sum_l \Pi(n,m;\ l)C\left(\begin{matrix}q/2\\ l\end{matrix}\middle|x\right).\tag{3.6}$$

The product linearization coefficients are given by the explicit formula

$$\begin{aligned}\Pi(n,m;\ n+m-2k)=&\frac{(m+n+q/2-2k)}{(m+n+q/2-k)}\\ &\times\frac{(q/2)_k(q/2)_{m-k}(q/2)_{n-k}(q)_{m+n-k}(m+n-2k)!}{k!(m-k)!(n-k)!(q/2)_{m+n-k}(q)_{m+n-2k}},\end{aligned}\tag{3.7}$$

where $(\alpha)_n$ is the nth ascending factorial of α. This result was first stated by Dougall [4] and proved by Hsü [7]; see also Askey's book. We can see from (3.7) that all the coefficients are nonnegative, and the only nonzero coefficients in (3.6) are those for which $n+m+l$ is even and $|n-m|\leqq l\leqq n+n$.

When many product linearizations are needed, it is easier to generate them using a recurrence relation such as the one found by Hyllaraas [8] or, more simply still, using the one found directly from the recurrence relation for the ultraspherical polynomials:

$$a_n\Pi(n+1,m;\ l)+b_n\Pi(n-1,m;\ l)=a_{l-1}\Pi(n,m;\ l-1)+b_{l+1}\Pi(n,m;\ l+1)\tag{3.8}$$

where

$$a_n=(n+1)/(2n+q)\quad\text{and}\quad b_n=(n+q-1)/(2n+q)\tag{3.9}$$

are the coefficients in the recurrence relation

$$xC\left(\begin{matrix}q/2\\ n\end{matrix}\middle|x\right)=a_nC\left(\begin{matrix}q/2\\ n+1\end{matrix}\middle|x\right)+b_nC\left(\begin{matrix}q/2\\ n-1\end{matrix}\middle|x\right).\tag{3.10}$$

Linearizing the square (3.5) of the Dirichlet kernel gives

$$\text{(3.11)}\quad \{D_{N/2}(A,B)\}^2=\Omega_q^{-2}\sum_{m,n=0}^{N/2}(1+2n/q)(1+2m/q)$$
$$\times\sum_{\left\{\substack{|n-m|\leqq l\leqq n+m\\ n+m+l\text{ even}}\right\}}\Pi(n,m;\,l)C\left(\begin{matrix}q/2\\ l\end{matrix}\middle|\cos\widehat{AB}\right).$$

The Fejér kernel has the form

$$\text{(3.12)}\quad F_N(A,B)=\Omega_q^{-1}\sum_{l=0}^{N}\sigma_l(1+2l/q)C\left(\begin{matrix}q/2\\ l\end{matrix}\middle|\cos\widehat{AB}\right)$$

with $\sigma_0=1$ for proper normalization. Comparing (3.11) and (3.12), we see that the Fejér sigma factors are

$$\text{(3.13)}\quad \sigma_l=\frac{S_l/S_0}{(1+2l/q)}$$

where

$$\text{(3.14)}\quad S_l=\sum_{\left\{\substack{0\leqq n,m\leqq N/2\\ |n-m|\leqq l\leqq n+m\\ n+m+l\text{ even}}\right\}}(1+2n/q)(1+2m/q)\Pi(n,m;\,l).$$

For even N, the Fejér kernel with sigma factors (3.13) should be sharper than the Fejér kernel with sigma factors (3.3). The sigma factors (3.3) are easier to obtain, and of course there may be some purposes for which the resulting kernel is actually better. For the important case $q=2$ the sigma factors (3.13) are, however, easy to obtain.

For $q=2$ the ultraspherical polynomials are Chebyshev polynomials of the second kind, and all the nonzero product linearization coefficients are unity, as may be seen directly from the explicit formula (3.7). Then (3.14) becomes

$$\text{(3.15)}\quad S_l=\sum_{\left\{\substack{0\leqq n,m\leqq N/2\\ |n-m|\leqq l\leqq n+m\\ n+m+l\text{ even}}\right\}}(1+n)(1+m).$$

On evaluating these sums we obtain

$$\text{(3.16)}\quad S_{2l}=\left(\frac{2l+1}{6}N-\frac{2l^2-1}{2}\right)(N/2+1)(N/2+2)$$
$$+\left(\frac{l(l+2)}{3}-\frac{2l+1}{6}\right)l(l+1),$$

$$\text{(3.17)}\quad S_{2l+1}=\frac{l+1}{3}\left[(N-3l)(N/2+1)(N/2+2)+l(l+1)(l+2)\right].$$

In particular

$$\text{(3.18)}\quad S_0=(N/6+1/2)(N/2+1)(N/2+2),\qquad S_{N+1}=0.$$

4. The Lanczos sigma factors. The Lanczos kernel on a sphere is obtained by integrating the Dirichlet kernel $D_N(A,C)$ over a window in C centered at B with radius Δ:

$$L_N(A,B)=\int_{\overset{\frown}{CB}\leqq\Delta} D_N(A,C)\,d\Omega(C)\Big/\int_{\overset{\frown}{CB}\leqq\Delta} d\Omega(C). \tag{4.1}$$

The second order Lanczos kernel is formed by integrating again over the window:

$$L_N^{(2)}(A,B)=\int_{\overset{\frown}{CB}\leqq\Delta} L_N(A,C)\,d\Omega(C)\Big/\int_{\overset{\frown}{CB}\leqq\Delta} d\Omega(C). \tag{4.2}$$

Higher order Lanczos kernels are obtained similarly by integrating repeatedly over the window.

To obtain the Lanczos sigma factors, we expand the Dirichlet kernel in spherical harmonics as in (2.7), so that

(4.3)

$$\int_{\overset{\frown}{CB}\leqq\Delta} D_N(A,C)\,d\Omega(C)=\sum_{n=0}^{N}\sum_{k_1\cdots k_q} Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|A\right)\int_{\overset{\frown}{CB}\leqq\Delta} Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|C\right)d\Omega(C).$$

We put B at the pole $(0,0..0)$, A at $(\overset{\frown}{AB},0..0)$, and give to C coordinates $(\gamma_0,\gamma_1..\gamma_q)$ where $\gamma_0=\overset{\frown}{CB}$. In the representation (2.7) all spherical harmonics vanish at A unless $k_2..k_q=0$. The integral over the window $\gamma_0=\overset{\frown}{CB}\leqq\Delta$ then vanishes unless $k_1=0$, because the part of the spherical harmonic which depends on γ_1 is

$$\left(\begin{matrix}k_1,0..0\\ n\end{matrix}\middle|C\right)\propto C\left(\begin{matrix}(q-1)/2\\ k_1\end{matrix}\middle|\cos\gamma_1\right), \tag{4.4}$$

and the integral over the window is proportional to

$$\int_0^{\pi} C\left(\begin{matrix}(q-1)/2\\ k_1\end{matrix}\middle|\cos\gamma_1\right)(\sin\gamma_1)^{q-1}d\gamma_1=0 \quad \text{if } k_1\neq 0. \tag{4.5}$$

The integral (4.3) of the Dirichlet kernel then becomes

$$\begin{aligned}(4.6)\quad &\int_{\overset{\frown}{CB}\leqq\Delta} D_N(A,C)\,d\Omega(C)\\ &=\sum_{n=0}^{N} Y\left(\begin{matrix}0..0\\ n\end{matrix}\middle|A\right)\int Y\left(\begin{matrix}0..0\\ n\end{matrix}\middle|C\right)d\Omega(C)\\ &=\sum_{n=0}^{N}\left\{H\left(\begin{matrix}0..0\\ n\end{matrix}\right)\right\}^{-1} C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\overset{\frown}{AB}\right)\times\Omega_{q-1}\int_0^{\Delta} C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\gamma_0\right)(\sin\gamma_0)^q d\gamma_0.\end{aligned}$$

Since

$$\int_{\overset{\frown}{CB}\leqq\Delta} d\Omega(C)=\Omega_{q-1}\int_0^{\Delta}(\sin\gamma_0)^q d\gamma_0 \tag{4.7}$$

we find, using (2.9) for H, that the Lanczos kernel becomes

$$L_N(A,B)=\Omega_q^{-1}\sum_{n=0}^{N}\sigma_n(1+2n/q)C\left(\begin{matrix}q/2\\ n\end{matrix}\middle|\cos\overset{\frown}{AB}\right) \tag{4.8}$$

with

$$\sigma_n = \frac{\int_0^\Delta C\left(\begin{smallmatrix} q/2 \\ n \end{smallmatrix} \middle| \cos\gamma_0\right)(\sin\gamma_0)^q\,d\gamma_0}{C\left(\begin{smallmatrix} q/2 \\ n \end{smallmatrix} \middle| 1\right)\int_0^\Delta (\sin\gamma_0)^q\,d\gamma_0}. \tag{4.9}$$

From the Rodrigues formula

$$C\left(\begin{matrix} q/2 \\ n \end{matrix} \,\middle|\, \cos\theta\right) = \frac{A_n^q}{(\sin\theta)^{q-1}}\left(\frac{1}{\sin\theta}\frac{d}{d\theta}\right)^n (\sin\theta)^{2n+q-1}, \tag{4.10}$$

with

$$A_n^q = \frac{\Gamma((q+1)/2)\Gamma(n+q)}{2^n n!\Gamma(n+(q+1)/2)\Gamma(q)}, \tag{4.11}$$

we find for $n>0$ that

$$\int_0^\Delta C\left(\begin{matrix} q/2 \\ n \end{matrix} \,\middle|\, \cos\gamma_0\right)(\sin\gamma_0)^q\,d\gamma_0 = \frac{q(\sin\Delta)^{q+1}}{n(q+n)}\,C\left(\begin{matrix} q/2+1 \\ n-1 \end{matrix} \,\middle|\, \cos\Delta\right). \tag{4.12}$$

Since

$$C\left(\begin{matrix} q/2 \\ n \end{matrix} \,\middle|\, 1\right) = \binom{n+q-1}{n} = \frac{q(q+1)}{n(q+n)}\,C\left(\begin{matrix} q/2+1 \\ n-1 \end{matrix} \,\middle|\, 1\right), \tag{4.13}$$

we have for $n>0$

$$\sigma_n = \frac{(\sin\Delta)^{q+1}/(q+1)}{\int_0^\Delta (\sin\gamma_0)^q\,d\gamma_0} \times \frac{C\left(\begin{matrix} q/2+1 \\ n-1 \end{matrix} \,\middle|\, \cos\Delta\right)}{C\left(\begin{matrix} q/2+1 \\ n-1 \end{matrix} \,\middle|\, 1\right)}. \tag{4.14}$$

We shall fix Δ by the requirement that

$$\sigma_{N+1} = 0, \tag{4.15}$$

or

$$C\left(\begin{matrix} q/2+1 \\ N \end{matrix} \,\middle|\, \cos\Delta\right) = 0. \tag{4.16}$$

For the important case $q=2$, the ultraspherical polynomials are Chebyshev polynomials of the second kind, $U_n(\cos\theta) = \sin((n+1)\theta)/\sin\theta$. It is then easiest to obtain the sigma factors by direct integration of (4.9):

$$\begin{aligned} \sigma_n &= \frac{\int_0^\Delta \sin((n+1)\gamma_0)(\sin\gamma_0)\,d\gamma_0}{(n+1)\int_0^\Delta (\sin\gamma_0)^2\,d\gamma_0} \\ &= \frac{1}{(n+1)}\left[\frac{\sin(n\Delta)}{n} - \frac{\sin((n+2)\Delta)}{(n+2)}\right]\bigg/(\Delta - \sin(2\Delta)/2). \end{aligned} \tag{4.17}$$

The condition $\sigma_{N+1} = 0$ gives

$$\tan((N+2)\Delta) - (N+2)\tan\Delta = 0 \tag{4.18}$$

which is easily solved by the iteration

$$\Delta = \frac{\pi + \arctan((N+2)\tan\Delta)}{(N+2)} \tag{4.19}$$

with the starting value

$$\Delta = \frac{Z}{N+2} \tag{4.20}$$

where $Z = 4.4934$ is the solution of

$$Z = \pi + \arctan(Z) \quad \text{or} \quad J_{3/2}(Z) = 0 \tag{4.21}$$

with $\pi < Z < 3\pi/2$.

On integrating again over the window, each spherical harmonic in the expansion (2.7) is simply multiplied by another factor of σ_n, so that the mth order Lanczos kernel is

$$L_N^{(m)}(A,B) = \sum_{n=0}^{N} \sigma_n^m \sum_{k_1..k_q} Y\left(\begin{matrix} k_1..k_q \\ n \end{matrix} \middle| A \right) Y\left(\begin{matrix} k_1..k_q \\ n \end{matrix} \middle| B \right). \tag{4.22}$$

5. The Lanczos kernel as an approximate identity. To show that the Lanczos kernel $L_N(A,B) = L_N(\cos\widehat{AB})$ is an approximate identity it is sufficient to show that

$$\int_{\text{all } B} \left| L_N(\cos\widehat{AB}) \right| d\Omega(B) = O(1), \tag{5.1}$$

$$\int_{\widehat{AB} \geqq \varepsilon} \left| L_N(\cos\widehat{AB}) \right| d\Omega(B) = O(N^{-\alpha}) \quad \text{with } \alpha > 0. \tag{5.2}$$

Over most of the interval $0 \leqq \widehat{AB} \leqq \pi$ the Lanczos kernel for large N oscillates rapidly, but to get the asymptotic behavior of the integrals (5.1) and (5.2) we need only concern ourselves with the amplitude of the oscillations. We get this amplitude from the integral

$$L_N(\cos\widehat{AB}) = \int_{\widehat{CB} \leqq \Delta} D_N(\cos\widehat{AC})\, d\Omega(C) \Big/ \int_{\widehat{CB} \leqq \Delta} d\Omega(C) \tag{5.3}$$

by substituting the asymptotic form of the Dirichlet kernel.

The angle $\widehat{AC}$ is

$$\cos\widehat{AC} = \cos\widehat{AB}\cos\widehat{CB} + \sin\widehat{AB}\sin\widehat{CB}\cos\widehat{ABC}. \tag{5.4}$$

As before we put B at $(0,0..0)$, A at $(\theta,0..0)$ where $\theta = \widehat{AB}$, and C at $(\gamma_0, \gamma_1 .. \gamma_q)$ where $\gamma_0 = \widehat{CB}$ and $\gamma_1 = \widehat{ABC}$. Then

$$\begin{aligned} \cos\widehat{AC} &= \cos(\theta - \delta) = \cos\theta\cos\delta + \sin\theta\sin\delta \\ &= \cos\theta\cos\gamma_0 + \sin\theta\sin\gamma_0\cos\gamma_1. \end{aligned} \tag{5.5}$$

Now $\gamma_0 = \widehat{CB} \leqq \Delta$ in the integral (5.3), and we saw in the last section that Δ is $O(N^{-1})$, so γ_0 is $O(N^{-1})$ also. Then

$$\begin{aligned} \delta &= \tan\theta - \left[(\tan\theta)^2 + \gamma_0^2 - 2\gamma_0\tan\theta\cos\gamma_1 \right]^{1/2} + O(N^{-3}) \\ &= \gamma_0\cos\gamma_1 - \tfrac{1}{2}\gamma_0^2\cot\theta(\sin\gamma_1)^2 + O\left(N^{-3}(\cot\theta)^2\right). \end{aligned} \tag{5.6}$$

We have to include factors of $\cot\theta$ when θ is near 0 or π. To accuracy $O(N^{-2})$, we can replace $\sin\gamma_0$ by γ_0, and (5.3) becomes

$$L_N(\cos\theta)=\frac{O(1)}{\Delta^{q+1}}\int_0^\Delta\gamma_0^q d\gamma_0\int_0^\pi(\sin\gamma_1)^{q-1}d\gamma_1 D_N(\cos(\theta-\delta)). \tag{5.7}$$

We can obtain the asymptotic forms for the Dirichlet kernel by starting with a simple closed expression in terms of Jacobi polynomials (Szegö [14, 4.5.3]):

$$D_N(\cos\theta)=O(N^{(q+1)/2})P\left(\begin{matrix}(q+1)/2,(q-1)/2\\N\end{matrix}\middle|\cos\theta\right). \tag{5.8}$$

For the interval $c/N\leqq\theta\leqq\pi-c/N$ with fixed c, the Jacobi polynomials have the following asymptotic form (Szegö [14, 8.21.12 and 8.21.18]; see also [15]):

$$\begin{aligned}P\left(\begin{matrix}\alpha,\beta\\N\end{matrix}\middle|\cos\theta\right)&=O(N^{-1/2})(\sin\theta/2)^{-\alpha-1/2}(\cos\theta/2)^{-\beta-1/2}\\&\quad\times\left\{\cos(N_0\theta-\phi_0)+\frac{O(N^{-1})}{\sin\theta}\cos(N_1\theta-\phi_1)+\cdots\right\},\end{aligned} \tag{5.9}$$

with $N_0,N_1=N+O(1)$ and $\phi_0,\phi_1=O(1)$. For the interval $0\leqq\theta\leqq c/N$ we have (Szegö [14, 8.21.17]):

$$P\left(\begin{matrix}\alpha,\beta\\N\end{matrix}\middle|\cos\theta\right)=O(1)\theta^{-\alpha}J_\alpha(N_0\theta)+\theta^2O(N^\alpha). \tag{5.10}$$

We will consider first the interval $c/N\leqq\theta\leqq\pi-c/N$. Combining (5.8) with (5.9), we have

$$\begin{aligned}D_N(\cos\theta)&=O(N^{q/2})(\sin\theta/2)^{-q/2-1}(\cos\theta/2)^{-q/2}\\&\quad\times\left\{\cos(N_0\theta-\phi_0)+\frac{O(N^{-1})}{\sin\theta}\cos(N_1\theta-\phi_1)+\cdots\right\}.\end{aligned} \tag{5.11}$$

When we substitute the asymptotic form (5.11) into the Lanczos kernel (5.7), we find the Lanczos kernel contains the integral

$$\begin{aligned}&\int_0^\pi(\sin\gamma_1)^{q-1}d\gamma_1\cos(N_0(\theta-\delta)-\phi_0)\\&=\cos(N_0\theta-\phi_0)\int_0^\pi(\sin\gamma_1)^{q-1}d\gamma_1\cos\left[N_0\gamma_0\cos\gamma_1+O(N^{-1})\cot\theta(\sin\gamma_1)^2\right]\\&\quad+\sin(N_0\theta-\phi_0)\int_0^\pi(\sin\gamma_1)^{q-1}d\gamma_1\sin\left[N_0\gamma_0\cos\gamma_1+O(N^{-1})\cot\theta(\sin\gamma_1)^2\right].\end{aligned} \tag{5.12}$$

Now we use the integral representation for Bessel functions

$$J_\nu(z)=O(1)z^\nu\int_0^\pi\cos(z\cos\gamma)(\sin\gamma)^{2\nu}d\gamma, \tag{5.13}$$

together with the fact that

$$\int_0^\pi\sin(z\cos\gamma)(\sin\gamma)^{2\nu}d\gamma=0, \tag{5.14}$$

and we find that (5.12) becomes

$$(5.15)\quad \int_0^{\pi}(\sin\gamma_1)^{q-1}d\gamma_1\cos(N_0(\theta-\delta)-\phi_0)$$

$$=\cos(N_0\theta-\phi_0)\times\left\{O(1)(N_0\gamma_0)^{-(q-1)/2}J_{(q-1)/2}(N_0\gamma_0)+O(N^{-2})(\cot\theta)^2\right\}$$

$$+\sin(N_0\theta-\phi_0)\times O(N^{-1})\cot\theta.$$

The integral over γ_0 becomes

$$(5.16)\quad \frac{1}{\Delta^{q+1}}\int_0^{\Delta}\gamma_0^q d\gamma_0\int_0^{\pi}(\sin\gamma_1)^{q-1}d\gamma_1\cos(N_0(\theta-\delta)-\phi_0)$$

$$=\cos(N_0\theta-\phi_0)\times\frac{O(1)}{(N_0\Delta)^{q+1}}\int_0^{N_0\Delta}x^{(q+1)/2}J_{(q-1)/2}(x)\,dx$$

$$+\sin(N_0\theta-\phi_0)\times O(N^{-1})\,\cot\theta.$$

From the ascending series for Bessel functions, it is easy to see that

$$(5.17)\qquad \int_0^{N_0\Delta}x^{(q+1)/2}J_{(q-1)/2}(x)\,dx=(N_0\Delta)^{(q+1)/2}J_{(q+1)/2}(N_0\Delta).$$

We shall now see that this vanishes to order $O(N^{-1})$ if Δ is fixed by

$$(4.16)\qquad C\left(\begin{matrix}q/2+1\\ N\end{matrix}\middle|\cos\Delta\right)=0.$$

The asymptotic form of the ultraspherical polynomial is obtained from (5.10) and $\Delta=O(N^{-1})$ as

$$(5.18)\quad C\left(\begin{matrix}q/2+1\\ N\end{matrix}\middle|\cos\Delta\right)=O(1)P\left(\begin{matrix}(q+1)/2,\,(q+1)/2\\ N\end{matrix}\middle|\cos\Delta\right)$$

$$=O(N^{(q+1)/2})\left\{J_{(q+1)/2}(N\Delta)+O(N^{-2})\right\}.$$

Since $N_0=N+O(1)$, (4.16) and (5.18) together imply that

$$(5.19)\qquad J_{(q+1)/2}(N\Delta)=O(N^{-2})\quad\text{and}\quad J_{(q+1)/2}(N_0\Delta)=O(N^{-1}).$$

Thus the Lanczos kernel becomes, for $c/N\leqq\theta\leqq\pi-c/N$,

$$(5.20)\quad L_N(\cos\theta)=O(N^{q/2})(\sin\theta/2)^{-q/2-1}(\cos\theta/2)^{-q/2}$$

$$\times\left\{O(N^{-1})\,\cos(N_0\theta-\phi_0)+O(N^{-1})\,\cot\theta\sin(N_0\theta-\phi_0)\right.$$

$$\left.+\frac{O(N^{-2})}{\sin\theta}\cos(N_1\theta-\phi_1)+\frac{O(N^{-2})}{\sin\theta}\cot\theta\sin(N_1\theta-\phi_1)+\cdots\right\}.$$

We see that the effect of Lanczos averaging is to reduce the amplitude of the kernel in the interval $c/N\leqq\theta\leqq\pi-c/N$ by a factor $O(N^{-1})$ $\cot\theta$. The amplitude will be reduced by the same factor every time the averaging is carried out, so the amplitude of the kth order Lanczos kernel in this interval will be less than the amplitude of the Dirichlet kernel by a factor of $O(N^{-k})(\cot\theta)^k$.

In the interval $\varepsilon \leqq \theta \leqq \pi - \varepsilon$ for fixed ε, $\cot\theta$ is $O(1)$, so

$$(5.21)\qquad \int_{\varepsilon}^{\pi-\varepsilon} |L_N^{(k)}(\cos\theta)|(\sin\theta)^q d\theta = O(N^{q/2-k}).$$

Over the interval $c/N \leqq \theta \leqq \varepsilon$ we can replace $\cot\theta$ by θ^{-1}, $\sin\theta/2$ by $\theta/2$, and $\cos\theta/2$ is $O(1)$, so

$$(5.22)\qquad \int_{c/N}^{\varepsilon} |L_N^{(k)}(\cos\theta)|\,(\sin\theta)^q d\theta = O(N^{q/2-k})\int_{c/N}^{\varepsilon} \theta^{-q/2-1}\theta^{-k}\theta^q d\theta$$

$$= \begin{cases} O(1) & \text{for } q/2 < k, \\ O(\ln N) & \text{for } q/2 = k, \\ O(N^{q/2-k}) & \text{for } q/2 > k. \end{cases}$$

Over the interval $\pi - \varepsilon \leqq \theta \leqq \pi - c/N$, $\sin\theta/2$ is $O(1)$ and we can replace $\cos\theta/2$ by $(\pi-\theta)/2$, so

$$(5.23)\qquad \int_{\pi-\varepsilon}^{\pi-c/N} |L_N(\cos\theta)|\,(\sin\theta)^q d\theta = O(N^{q/2-k})\int_{c/N}^{\varepsilon} \theta^{-q/2}\theta^{-k}\theta^q d\theta$$

$$= \begin{cases} O(N^{-1}) & \text{for } q/2 \leqq k-1, \\ O(N^{q/2-k}) & \text{for } q/2 > k-1. \end{cases}$$

In the intervals $0 \leqq \theta \leqq c/N$ and $\pi - c/N \leqq \theta \leqq \pi$, Lanczos averaging does not produce any effective cancellation, and the Lanczos kernel is of the same order of magnitude as the Dirichlet kernel. (5.8) together with the asymptotic form (5.10) gives, for $0 \leqq \theta \leqq c/N$

$$(5.24)\qquad D_N(\cos\theta) = O(N^{(q+1)/2})\theta^{-(q+1)/2}J_{(q+1)/2}(N_0\theta) + \theta^2 O(N^{(q+1)/2}),$$

so that

$$(5.25)\qquad \int_0^{c/N} |L_N(\cos\theta)|(\sin\theta)^q d\theta = O(1)\int_0^{c/N} |D_N(\cos\theta)|\,(\sin\theta)^q d\theta$$

$$= O(N^{(q+1)/2})\int_0^{c/N} \theta^{-(q+1)/2}\theta^q d\theta$$

$$= O(1) \quad \text{for all } q.$$

Recalling that

$$(5.26)\qquad P\left(\begin{matrix}\alpha,\beta \\ N\end{matrix}\middle|\cos\theta\right) = (-)^N P\left(\begin{matrix}\beta,\alpha \\ N\end{matrix}\middle|\cos(\pi-\theta)\right),$$

we obtain from (5.8) and (5.10)

$$(5.27)\qquad D_N(\cos(\pi-\theta)) = O(N^{q/2+1})\theta^{-(q-1)/2}J_{(q-1)/2}(N_0\theta) + \theta^2 O(N^{(q-1)/2}),$$

so that

$$(5.28)\qquad \int_{\pi-c/N}^{\pi} |L_n(\cos\theta)|(\sin\theta)^q d\theta = O(1)\int_0^{c/N} |D_N(\cos(\pi-\theta))|(\sin\theta)^q d\theta$$

$$= O(N^{(q+1)/2})\int_0^{c/N} \theta^{-(q-1)/2}\theta^q d\theta$$

$$= O(N^{-1}) \quad \text{for all } q.$$

Combining (5.21) through (5.23), (5.25) and (5.28), we have

$$(5.29)\qquad \int_{\text{all }B} |L_N^{(k)}(A,B)|\,d\Omega(B) = \int_0^\pi |L_N^{(k)}(\cos\theta)|(\sin\theta)^q d\theta$$

$$= \begin{cases} O(1) & \text{for } q/2<k, \\ O(\ln N) & \text{for } q/2=k, \\ O(N^{q/2-k}) & \text{for } q/2>k, \end{cases}$$

and

$$(5.30)\qquad \int_{\overline{AB}\geqq\varepsilon} |L_N^{(k)}(A,B)|\,d\Omega(B) = \int_\varepsilon^\pi |L_N^{(k)}(\cos\theta)|(\sin\theta)^q d\theta$$

$$= \begin{cases} O(N^{-1}) & \text{for } q/2\leqq k-1, \\ O(N^{q/2-k}) & \text{for } q/2>k-1. \end{cases}$$

Thus the kth order Lanczos kernel is an approximate identity if $q<2k$. In particular, the first order Lanczos kernel is an approximate identity for $q=0$ or 1, the second order Lanczos kernel is an approximate identity for $q=2$ or 3, and so on.

6. Fejér and Lanczos kernels in Euclidean space. We can obtain kernels $K_V(x,y)$ in E_{q+1} from the corresponding kernels $K_\Omega(A,B)$ on the surface S_{q+1} of a $(q+2)$-dimensional sphere by a mapping $f: x\to A$ such that

$$(6.1)\qquad f:(r,\theta_1..\theta_q)\to(\theta_0(r),\theta_1..\theta_q).$$

A simple form for $\theta_0(r)$ which maps E_{q+1} onto the whole of S_{q+1} is

$$(6.2)\qquad \theta_0=2\arctan(r/a).$$

This maps widely separated points at large r in E_{q+1} onto points near the pole $\theta_0=\pi$ of S_{q+1}. This will not matter if the function we are approximating either vanishes rapidly or tends to a constant as $r\to\infty$. If however the function varies significantly with $\theta_1..\theta_q$ as $r\to\infty$, it is better to map E_{q+1} onto the half sphere $0\leqq\theta\leqq\pi/2$ by a transformation such as

$$(6.3)\qquad \theta_0=\arctan(r/a),$$

which maps points at large r with different $\theta_1..\theta_q$ onto well separated points on the sphere. Functions such as a radiating wave which oscillate at constant $\theta_1..\theta_q$ even at large r need special treatment.

The kernel $K_\Omega(A,B)$ is normalized over the full sphere; if we wish to use (6.3), we need a kernel normalized over the half sphere. Such a kernel is

$$(6.4)\qquad K_H(A,B)=K_\Omega(RA,B)+K_\Omega(A,B)$$

where R is the reflection operator

$$(6.5)\qquad R:(\theta_0,\theta_1..\theta_1)\to(\pi-\theta_0,\theta_1..\theta_q).$$

Now the part of the spherical harmonic Y which depends on θ_0 is

$$(6.6)\qquad Y\left(\begin{matrix} k_1..k_q \\ n \end{matrix}\middle|\,\theta_0,\theta_1..\theta_q\right)\propto C\left(\begin{matrix} k_1+q/2 \\ n-k_1 \end{matrix}\middle|\,\cos\theta_0\right).$$

Terms in $K_\Omega(A,B)$ which are odd functions of $\theta_0-\pi/2$ will cancel in the sum (6.4) of K_H, so only those terms for which $n-k_1$ is even will survive in the kernel K_H. The expansion of K_H is therefore

$$K_H(A,B)=2\sum_{n=0}^{N}\sigma_n\sum_{\left\{\substack{k_1..k_q\\ n-k_1\ \text{even}}\right\}}Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|A\right)Y\left(\begin{matrix}k_1..k_q\\ n\end{matrix}\middle|B\right). \tag{6.7}$$

The kernels $K_V(x,y)$ obtained from $K_\Omega(A,B)$ by a mapping of the type (5.1) are not normalized over the volume element dV_q of E_{q+1} but over the element $d\Omega_q$ of S_{q+1}. Consequently the normalization is not

$$\int K_V(x,y)\,dV(y)=1 \quad\text{or}\quad \int dV(x)K_V(x,y)=1, \tag{6.8}$$

but is instead

$$\int K_V(x,y)W(y)\,dV(y)=1 \quad\text{or}\quad \int dV(x)W(x)K_V(x,y)=1, \tag{6.9}$$

where

$$W=\frac{d\Omega_q}{dV_q}=\left(\frac{\sin\theta_0}{r}\right)^q\frac{d\theta_0}{dr} \tag{6.10}$$

is the appropriate weighting function.

REFERENCES

[1] R. Askey, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] ______, *Refinement of Abel summability for Jacobi series*, Proc. Symposium Pure Mathematics, 26, Harmonic Analysis on Homogeneous Spaces, C. Moore, ed., American Mathematical Society, Providence, RI, 1973, pp. 335–338.

[3] R. Askey and H. Pollard, *Some absolutely monotonic and completely monotonic functions*, this Journal, 5 (1974), pp. 58–63.

[4] J. Dougall, *A theorem of Sonine in Bessel functions, with two extensions to spherical harmonics*, Proc. Edinburgh Math. Soc., 37 (1919), pp. 33–47.

[5] R. A. Gingold and J. J. Monaghan, *Smoothed particle hydrodynamics: theory and application to non-spherical stars*, Monthly Notices Roy. Astronom. Soc., 181 (1977), pp. 375–389.

[6] H. Hochstadt, *The Functions of Mathematical Physics*, Wiley-Interscience, New York, 1971.

[7] H. Y. Hsü, *Certain integrals and infinite series involving ultraspherical polynomials and Bessel functions*, Duke Math. J., 4 (1938), pp. 374–383.

[8] E. Hylleraas, *Linearizations of products of Jacobi polynomials*, Math. Scand., 10 (1962), pp. 189–200.

[9] E. Kogbetliantz, *Recherches sur la sommabilité des séries ultra-sphériques par la méthode des moyennes arithmetiques*, J. Math. Pures Appl. (9), 3 (1924), pp. 107–187.

[10] C. Lanczos, *Applied Analysis*, Pitman, London, 1956.

[11] ______, *Trigonometric interpolation of empirical and analytical functions*, J. Math. Phys., 17 (1938), pp. 123–199.

[12] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Stat., 33 (1962), pp. 1065–1076.

[13] V. L. Shapiro, *The symmetric derivative on the $(k-1)$ dimensional sphere*, Trans. Amer. Math. Soc., 81 (1965), pp. 514–524.

[14] G. Szegö, *Orthogonal Polynomials*, Colloquium Publications, vol. 23, 4th ed., American Mathematical Society, Providence, RI, 1967.

[15] ______, *Asymptotische Entwicklungen der Jacobische Polynome*, Schriften der Konigsberger Gelehrten Gesellschaft, naturwissenschaftliche Klasse, 10 (1933), pp. 35–112.

SIAM J. MATH. ANAL.
Vol. 18, No. 1, January 1987

021

ON THE ASYMPTOTIC EXPANSION OF MELLIN TRANSFORMS*

C. L. FRENZEN†

Abstract. The asymptotic behavior of the Mellin transform $M[f;\ x]$ is studied as $x \to +\infty$, and it is shown that the Mellin transform of a certain class of asymptotic sequences $\{\phi_n(t)\}$ $(t \to \infty)$ yields another asymptotic sequence $\{M[\phi_n;\ x]\}$ $(x \to \infty)$. An analogue of Watson's lemma is also established; i.e., under certain circumstances, the asymptotic expansion of $f(t)$ $(t \to \infty)$ induces an asymptotic expansion of $M[f;\ x]$ $(x \to \infty)$.

Key words. asymptotic expansion, Mellin transform

AMS(MOS) subject classifications. Primary 41A60; secondary 44A15

1. Introduction. Recently A. Sidi [7] studied the asymptotic behavior of the Mellin transform

$$M[f;\ x] = \int_0^\infty t^{x-1} f(t)\, dt \tag{1.1}$$

as $x \to +\infty$. This transform and its associated convolution have played an important role in recent developments in the asymptotic expansion of integrals; see Wong [9], and Handelsman and Lew [3]. Sidi, however, was interested in establishing a Mellin transform analogue of Watson's lemma—a result which would say that, under certain circumstances, an asymptotic expansion of $f(t)$ induces an asymptotic expansion of the Mellin transform of $f(t)$, $M[f;\ x]$. Until Sidi's work, this problem had received little attention, although Doetsch [1] and Handelsman and Lew [3] had considered the problem of analytic continuation of the Mellin transform, Riekstins [6] had considered the asymptotic expansion of the inverse Mellin transform and Wagner [8] had obtained some Tauberian theorems for Mellin transforms.

By giving several specific examples, Sidi established some evidence to support the following: under certain circumstances, (1) the dominant contribution to the Mellin transform $M[f;\ x]$ as $x \to +\infty$ comes from the large t-behavior of $f(t)$; (2) the Mellin transform of an asymptotic sequence (for $t \to \infty$) yields an asymptotic sequence (for $x \to +\infty$); (3) the Mellin transform of an asymptotic expansion (for $t \to \infty$) yields an asymptotic expansion (for $x \to +\infty$). In this note we show that (1), (2) and (3) hold for a general class of functions, sequences, and asymptotic expansions, respectively. In particular we dispense with Sidi's assumption that the Mellin transform of a specific asymptotic sequence is again an asymptotic sequence by showing that (2) holds for a general class of asymptotic sequences.

Suppose that a function $f(t)$ is locally integrable for $0 < t < \infty$, such that for some real constant σ, $t^{\sigma-1} f(t)$ is absolutely integrable in any finite interval of the form $[0, a]$, and $f(t) = o(t^{-\mu})$ as $t \to \infty$ for any $\mu > 0$. It follows that $M[f;\ z]$ exists for all sufficiently large $\operatorname{Re} z = x$. We shall say that a function $f(t)$ is in the class M (or in M) if $M[f;\ z]$ exists for all $\operatorname{Re} z = x > x_0$, where x_0 may depend on f. It is well known that

*Received by the editors February 20, 1985, and in revised form July 25, 1985.

† Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

if $f(t)$ is in M, then $M[f;\, z]$ converges absolutely and is holomorphic for all $\operatorname{Re} z = x > x_0$. Here we show that (1) holds for all real nonnegative functions in M which are strictly positive beyond some (perhaps large) value of their argument. For (2) to hold, each individual function of the asymptotic sequence must satisfy the above requirements for (1). Finally in (3) we make use of generalized asymptotic expansions. We require that the function itself and each of the functions in its generalized asymptotic expansion be in M, and that the asymptotic sequence or scale be equivalent to another asymptotic sequence which satisfies the requirements for (2).

In many ways this paper is the Mellin transform analogue of A. Erdélyi's [2] fundamental paper on generalized asymptotic expansions of Laplace integrals. However there are some differences. Erdélyi considered a one-sided Laplace integral:

$$L[g;\, x] = \int_0^\infty e^{-xu} g(u)\, du. \tag{1.2}$$

As $x \to +\infty$, the kernel e^{-xu} decreases for each fixed positive u, but decreases least rapidly for small u. Consequently the *small-u* behavior of $g(u)$ most influences the large-x behavior of $L[g;\, x]$. If $t = e^u$ and $f(t) = g(u)$ then the Mellin transform becomes a *two*-sided Laplace integral:

$$M[f;\, x] = \int_0^\infty t^{x-1} f(t)\, dt = \int_{-\infty}^\infty e^{xu} g(u)\, du. \tag{1.3}$$

The constant $\int_{-\infty}^0 |g(u)|\, du$ bounds the part $\int_{-\infty}^0 e^{xu} g(u)\, du$, which is a standard one-sided Laplace integral. In the remaining part, $\int_0^\infty e^{xu} g(u)\, du$, as $x \to \infty$ the kernel e^{xu} increases for each fixed positive u, and increases most rapidly for large u. Hence it is the *large-t* behavior of $f(t)$ which most influences the large-x behavior of $M[f;\, x]$.

Since the Mellin transform of a function in M converges in the complex half plane $x > x_0$, it is also natural to ask, in an attempt to extend the validity of (2), whether it is possible for an asymptotic sequence $\{\phi_n(t)\}$ for $t \to \infty$ to induce an asymptotic sequence $\{M[\phi_n;\, z]\}$ as z tends to infinity in some region of the complex plane. By means of a simple example we show that limited results in this direction are possible.

2. A basic inequality. Throughout this paper t is a real variable, z is a complex variable, and we always write $z = x + iy$. The Mellin transform analogue of Watson's lemma will follow from a result relating the relative behavior of two functions at $t = \infty$ to the relative behavior of their Mellin transforms at $x = \infty$. Let the two functions be denoted by $g(t)$ and $h(t)$, and their Mellin transforms by $M[g;\, x]$ and $M[h;\, x]$. The result below is the Mellin transform analogue of Lemma 1 in [2].

LEMMA 1. *Suppose g and h are in M, $h(t) \geqq 0$, and $h(t) > 0$ for $t_1 < t < \infty$ for some $t_1 > 0$. Then*

$$\limsup_{z \to \infty} \frac{|M[g;\, z]|}{M[h;\, x]} \leqq \operatorname{ess\,lim\,sup}_{t \to \infty} \frac{|g(t)|}{h(t)} \tag{2.1}$$

when $z \to \infty$ in such a manner that $x \to +\infty$.

Proof. If g and h are in M, they possess Mellin transforms for $x > x_0$ for some x_0, so that $M[g;\, z]$ and $M[h;\, x]$ exist and $M[h;\, x] > 0$ whenever $x > x_0$. Let $x_1 > x_0$, and assume $\operatorname{Re} z = x \geqq x_1$. For some fixed T, $t_1 < T < \infty$, we write $M[g;\, z] = I_1 + I_2$ where

$$I_1 = \int_0^T t^{z-1} g(t)\, dt, \tag{2.2}$$

and

$$I_2=\int_T^\infty t^{z-1}g(t)\,dt. \tag{2.3}$$

From (2.2) we have

$$|I_1|\leqq\int_0^T t^{x_1-1}|g(t)|t^{x-x_1}\,dt\leqq B(T,x_1)T^{x-x_1}, \tag{2.4}$$

where $B(T,x_1)$ is a constant depending only on T and x_1. This result follows from the absolute convergence of $M[g;\ x]$ for $x\geqq x_1>x_0$.

We set

$$U_T=\operatorname{ess\,sup}\left[\frac{|g(t)|}{h(t)}:T<t<\infty\right] \tag{2.5}$$

and assume that $U_T<\infty$ for some T. (Otherwise the right-hand side of (2.1) is equal to $+\infty$ and there is nothing to prove.) Then $|g(t)|\leqq U_T h(t)$ for almost all $t>T$, and so

$$|I_2|\leqq\int_T^\infty t^{x-1}U_T h(t)\,dt\leqq U_T M[h;\ x]. \tag{2.6}$$

Consequently, for any $t_1<T<\infty$, we have

$$\frac{|M[g;\ z]|}{M[h;\ x]}\leqq U_T+\frac{B(T,x_1)T^{x-x_1}}{M[h;\ x]}. \tag{2.7}$$

Now by the conditions imposed on $h(t)$, for every $\varepsilon>0$ we have

$$M[h;\ x]\geqq\int_{e^\varepsilon T}^\infty t^{x_1-1}h(t)t^{x-x_1}\,dt\geqq e^{\varepsilon(x-x_1)}T^{x-x_1}\int_{e^\varepsilon T}^\infty t^{x_1-1}h(t)\,dt, \tag{2.8}$$

where the last integral in (2.8) is a constant $D(\varepsilon,T,x_1)$ depending only on ε, T and x_1 and is bounded away from zero as $\varepsilon\to 0$. Using (2.8) in (2.7) gives

$$\frac{|M[g;\ z]|}{M[h;\ x]}\leqq U_T+\frac{B(T,x_1)}{D(\varepsilon,T,x_1)}e^{-\varepsilon(x-x_1)}. \tag{2.9}$$

Now let $z\to\infty$ in a manner satisfying the last condition of the lemma, i.e. so that $x\to+\infty$ as well. The second term on the right-hand side of (2.9) then approaches zero and

$$\limsup_{z\to\infty}\frac{|M[g;\ z]|}{M[h;\ x]}\leqq U_T. \tag{2.10}$$

Since the left-hand side of (2.10) is independent of T, and the right-hand side tends to the right hand of (2.1) as $T\to\infty$, this proves the lemma. □

Note that the last condition in the lemma is satisfied if z tends to infinity in the sector $S_\Delta:|\arg z|\leqq\pi/2-\Delta$, $\Delta>0$.

We now show that under certain circumstances (1) holds—that the dominant contribution to $M[h;\ x]$ for $x\to\infty$ comes from the large t-behavior of $h(t)$. This is suggested by Lemma 1, in which only large values of t are important. Specifically, suppose that h_1 and h_2 are in M and $h_1(t)=h_2(t)>0$ for $t_2<t<\infty$ for some $t_2>0$. Then

$$M[h_1;\ x]-M[h_2;\ x]=\int_0^{t_2}t^{x-1}[h_1(t)-h_2(t)]\,dt=O(t_2^x) \tag{2.11}$$

as $x\to\infty$ by an argument similar to that leading to (2.4). Without loss of generality, we shall assume $t_2>1$. Then $\log t_2>0$ and from (2.11) we have for $\log t_2<\delta<\infty$

$$(2.12)\qquad e^{3\delta x/2}(M[h_1;\, x]-M[h_2;\, x])\to 0 \qquad (x\to\infty).$$

Now by the conditions on h_1 and h_2, for $x\geqq x_1>x_0$

$$(2.13)\quad M[h_1;\, x]\geqq -\int_0^{t_2} t^{x_1-1}|h_1(t)|t^{x-x_1}\,dt+\int_{e^{2\delta}t_2}^{\infty} t^{x_1-1}h_1(t)t^{x-x_1}\,dt$$

$$\geqq -t_2^{x-x_1}\int_0^{t_2} t^{x_1-1}|h_1(t)|\,dt+t_2^{x-x_1}e^{2\delta(x-x_1)}\int_{e^{2\delta}t_2}^{\infty} t^{x_1-1}h_1(t)\,dt.$$

From (2.13) it follows that

$$(2.14)\qquad e^{-3\delta x/2}M[h_1;\, x]\to\infty \qquad (x\to\infty),$$

and a similar argument holds for $e^{-3\delta x/2}M[h_2;\, x]$. This result together with (2.12) implies that

$$(2.15)\qquad \frac{M[h_1;\, x]}{M[h_2;\, x]}\to 1 \qquad (x\to\infty).$$

Since t_2 was arbitrary, (2.15) shows that the dominant contribution to $M[h;\, x]$ as $x\to\infty$ for $h(t)>0$ beyond some value of t comes from the large t-behavior of $h(t)$.

3. Asymptotic sequences. The results of the preceding section may be used to show that under certain circumstances the Mellin transforms $\{M[\phi_n;\, x]\}$ of an asymptotic sequence $\{\phi_n(t)\}$ form an asymptotic sequence.

THEOREM 1. (i) *Suppose that $\{\phi_n(t)\}$ is an asymptotic sequence for $t\to\infty$ and for each n, $\phi_n(t)$ is in M, $\phi_n(t)\geqq 0$ and $\phi_n(t)>0$ for $t_n<t<\infty$. Then $\{M[\phi_n;\, x]\}$ is an asymptotic sequence as $x\to\infty$.*

(ii) *If, in addition, there is an unbounded set R in the complex plane such that $x\to\infty$ whenever $z\to\infty$ in R, and if for each n,*

$$(3.1)\qquad M[\phi_n;\, x]=O(M[\phi_n;\, z])$$

as $z\to\infty$ in R, then $\{M[\phi_n;\, z]\}$ is also an asymptotic sequence as $z\to\infty$ in R.

Proof. (i) From Lemma 1, we have

$$(3.2)\qquad \limsup_{x\to+\infty}\frac{M[\phi_{n+1};\, x]}{M[\phi_n;\, x]}\leqq \operatorname{ess}\limsup_{t\to\infty}\frac{\phi_{n+1}(t)}{\phi_n(t)}$$

and the right-hand side is zero since $\{\phi_n(t)\}$ is an asymptotic sequence. Hence $M[\phi_{n+1};\, x]=o(M[\phi_n;\, x])$ as $x\to\infty$, and this proves the first half of the theorem.

(ii) Since $\phi_n(t)\geqq 0$, we have

$$(3.3)\qquad |M[\phi_n;\, z]|\leqq\int_0^\infty t^{x-1}\phi_n(t)\,dt=M[\phi_n;\, x]$$

so that $M[\phi_n;\, z]=O(M[\phi_n;\, x])$ as $z\to\infty$ in R. Since by (3.1) we also have $M[\phi_n;\, x]=O(M[\phi_n;\, z])$, it follows that the two sequences $\{M[\phi_n;\, x]\}$ and $\{M[\phi_n;\, z]\}$ are equivalent as $z\to\infty$ in R. Since by part (i) the first sequence is asymptotic, the second must also be asymptotic. □

Example 1. Consider the sequence $\{\phi_n(t)\}$ with $\phi_n(t)=t^{-\lambda_n}\exp(-\alpha_n t^{\beta_n})$, where λ_n, α_n and β_n are real. The sequence $\{\phi_n(t)\}$ is an asymptotic sequence as $t\to\infty$ in any one of the following cases: for $n > m$

(a) $\beta_n > \beta_m > 0$, and $\alpha_k > 0$, $k = 1,2,3 \cdots$.

(b) $\beta_n \geqq \beta_m > 0$ and $\alpha_k > 0$, $k = 1,2,3 \cdots$; when $\beta_m = \beta_n$, then $\alpha_n > \alpha_m > 0$.

(c) $\beta_n \geqq \beta_m > 0$ and $\alpha_k > 0$, $k = 1,2,3 \cdots$; when $\beta_m = \beta_n$, then $\alpha_n > \alpha_m > 0$; when $\beta_m = \beta_n$ and $\alpha_m = \alpha_n > 0$, then $\lambda_n > \lambda_m$.

Note that each $\phi_n(t)$ is in M if $x > \lambda_n$. The sequence $\{\phi_n(t)\}$ satisfies the conditions of Theorem 1(i), and hence, as is easily checked, the sequence $\{M[\phi_n;\ x]\}$ where

$$(3.4) \qquad M[\phi_n;\ x] = \frac{1}{\beta_n} \alpha_n^{-(x-\lambda_n)/\beta_n} \Gamma\left(\frac{x-\lambda_n}{\beta_n}\right) \qquad (x > \lambda_n)$$

is an asymptotic sequence as $x \to \infty$ under any of the conditions (a), (b), (c). In this example one can also allow α_n to be complex. The sequence $\{\phi_n(t)\}$ then becomes complex valued and Theorem 1 (i) is no longer applicable, but the path of integration in $M[\phi_n;\ x]$ may be rotated and further conditions placed on the sequences $\{|\alpha_n|\}$ and $\{\operatorname{Re} \alpha_n\}$ so that the sequence $\{M[\phi_n;\ x]\}$ is again asymptotic. This particular example was treated by Sidi [7, Thm. 2.1].

The next example illustrates that in certain cases the domain R in Theorem 1 (ii) in which $M[\phi_n;\ x]$ and $M[\phi_n;\ z]$ are equivalent asymptotic sequences may not be the largest domain in the complex plane in which $M[\phi_n;\ z]$ is an asymptotic sequence.

Example 2. Consider the sequence $\{\phi_n(t)\}$ where $\phi_n(t) = e^{-\alpha_n t}$ and $\alpha_{n+1} > \alpha_n > 0$. As $t \to \infty$, $\{\phi_n(t)\}$ is an asymptotic sequence and each $\phi_n(t)$ is in M. Since $M[\phi_n;\ z] = \alpha_n^{-z}\Gamma(z)$, $\operatorname{Re} z > 0$, we have

$$(3.5) \qquad \frac{|M[\phi_n;\ z]|}{M[\phi_n;\ x]} = \frac{|\alpha_n^{-z}\Gamma(z)|}{\alpha_n^{-x}\Gamma(x)} = \frac{|\Gamma(z)|}{\Gamma(x)} \leqq 1 \qquad (\operatorname{Re} z = x > 0);$$

see, for example [5, p. 38]. Therefore $M[\phi_n;\ z] = O(\mathrm{M}[\phi_n;\ x])$, and to apply Theorem 1 (ii), we need to exhibit an unbounded set R with $x \to \infty$ whenever $z \to \infty$ in R, such that, for each n, $M[\phi_n;\ x] = O(M[\phi_n;\ z])$ as $z \to \infty$ in R. Since

$$(3.6) \qquad I = \left(\frac{\Gamma(x)}{|\Gamma(z)|}\right)^2 = \prod_{n=0}^{\infty}\left(1 + \frac{y^2}{(x+n)^2}\right)$$

(see [5, p. 38], for example) this means finding an R such that as $z \to \infty$ in R, $x \to \infty$ and the right-hand side of (3.6) is bounded. From the latter, we have

$$(3.7) \qquad \log I = \sum_{n=0}^{\infty} \log\left(1 + \frac{y^2}{(x+n)^2}\right) = \sum_{n=0}^{\infty}\left(\frac{y^2}{(x+n)^2} + O\left(\frac{y^4}{(x+n)^4}\right)\right)$$
$$= y^2\zeta(2,x) + O(y^4\zeta(4,x))$$

where $\zeta(m,x) = \sum_{n=0}^{\infty}(1/(x+n)^m)$ is the generalized zeta function. For fixed m and large positive x, $\zeta(m,x) = 1/(m-1)x^{1-m} + O(x^{-m})$ (see [4, p. 25]), and using this in (3.7) gives

$$(3.8) \qquad \log I = \frac{y^2}{x} + O\left(\frac{y^2}{x^2}\right).$$

Thus $\log I$, and consequently I, is bounded as $z \to \infty$ in R if R consists of a region of the complex plane in which

$$(3.9) \qquad (\operatorname{Im} z)^2 = O(\operatorname{Re} z) \qquad (z \to \infty, \operatorname{Re} z = x > 0).$$

In this region $x \to \infty$ as $z \to \infty$ and $M[\phi_n; x] = O(M[\phi_n; z])$, so that by Theorem 1 (ii) $\{M[\phi_n; z]\}$ is an asymptotic sequence as $z \to \infty$ in R, defined in (3.9). However,

$$\frac{|M[\phi_{n+1}; z]|}{|M[\phi_n; z]|} = \left(\frac{\alpha_n}{\alpha_{n+1}}\right)^x, \tag{3.10}$$

and the right-hand side of (3.10) tends to zero as $x \to \infty$. Consequently, for example, $M[\phi_n; z]$ is an asymptotic sequence as $z \to \infty$ in the sector $S_\Delta = \{z : |\arg z| \leqq \pi/2 - \Delta, \Delta > 0\}$, a larger region than that in (3.9), which was determined from Theorem 1 (ii).

Example 3. Consider the sequence $\{\phi_n(t)\}$ where

$$\phi_n(t) = \begin{cases} 0, & t < c, \\ (\log t)^{a_n} e^{-b_n t}, & t \geqq c \end{cases} \tag{3.11}$$

where $c > 1$ and $b_n > 0$ for all n. As $t \to \infty$, the sequence $\{\phi_n(t)\}$ is an asymptotic sequence under either of the conditions

(a′) $b_{n+1} > b_n > 0$,

(b′) $b_{n+1} \geqq b_n > 0$; if $b_{n+1} = b_n$ then $a_{n+1} < a_n$.

By Theorem 1, $M[\phi_n; x]$ will be an asymptotic sequence under (a′) or (b′) as $x \to +\infty$. We now confirm this. Note that if a_n is a positive integer, the asymptotic behavior of

$$M[\phi_n; x] = \int_c^\infty t^{x-1} (\log t)^{a_n} e^{-b_n t} dt \tag{3.12}$$

may be obtained by considering the derivative of the incomplete gamma function.

Now write (3.12) as

$$M[\phi_n; x] = I_1 + I_2 + I_3 \tag{3.13}$$

where

$$I_1 = \int_c^{x^{1-\delta}} t^{x-1} (\log t)^{a_n} e^{-b_n t} dt, \tag{3.14}$$

$$I_2 = x^x (\log x)^{a_n} \int_{x^{-\delta}}^{x^\delta} y^{x-1} \left(1 + \frac{\log y}{\log x}\right)^{a_n} e^{-b_n x y} dy \tag{3.15}$$

and

$$I_3 = \int_{x^{1+\delta}}^\infty t^{x-1} (\log t)^{a_n} e^{-b_n t} dt, \tag{3.16}$$

with δ a positive number satisfying $0 < \delta < 1$. We estimate I_1 as follows:

$$I_1 \leqq e^{-b_n c} \left((1-\delta)^{|a_n|} (\log x)^{|a_n|}\right) \int_c^{x^{1-\delta}} t^{x-1} dt. \tag{3.17}$$

Equation (3.17) implies

$$I_1 = O\left((\log x)^{|a_n|} x^x x^{-\delta x - 1}\right) \qquad (x \to +\infty). \tag{3.18}$$

To estimate I_3, note that for any a_n there exists a constant A_n such that

$$(\log t)^{|a_n|} \leqq A_n t, \qquad x^{1+\delta} \leqq t < \infty,$$

and since $b_n t/2x < \exp(b_n t/2x)$ it follows that $t^x < (2x/b_n)^x \exp(b_n t/2)$ for $x^{1+\delta} \leqq t < \infty$. Consequently

$$I_3 \leqq A_n \left(\frac{2x}{b_n}\right)^x \int_{x^{1+\delta}}^\infty e^{-b_n t/2} dt \tag{3.19}$$

or

$$I_3 = O\left(e^{-b_n x^{\delta/2}/2}\right) \qquad (x \to \infty). \tag{3.20}$$

To evaluate I_2, note from (3.15) that within its range of integration $|\log y/\log x| \leqq \delta < 1$, and so using the finite binomial expansion gives

$$\left(1 + \frac{\log y}{\log x}\right)^{a_n} = \sum_{j=0}^{N} \binom{a_n}{j}\left(\frac{\log y}{\log x}\right)^j + R_N, \tag{3.21}$$

where N is a positive integer and

$$|R_N| \leqq K \left|\left(\frac{\log y}{\log x}\right)^{N+1}\right| \tag{3.22}$$

for some fixed $K > 0$. Substituting (3.21) into (3.15) gives

$$I_2 = x^x (\log x)^{a_n} \sum_{j=0}^{N} \binom{a_n}{j} (\log x)^{-j} B_j(x) + r_N(x) \tag{3.23}$$

where

$$B_j(x) = \int_{x^{-\delta}}^{x^{\delta}} e^{-x(b_n y - \log y)} \frac{(\log y)^j}{y}\, dy \tag{3.24}$$

and

$$|r_N(x)| \leqq K x^x (\log x)^{a_n - N - 1} \int_{x^{-\delta}}^{x^{\delta}} y^{x-1} |\log y|^{N+1} e^{-b_n x y}\, dy. \tag{3.25}$$

Since

$$B_j(x) \sim \int_0^\infty e^{-x(b_n y - \log y)} \frac{(\log y)^j}{y}\, dy \tag{3.26}$$

as $x \to +\infty$, by (a modification of) the method of Laplace (see [5]),

$$B_j(x) \sim \sqrt{\frac{2\pi}{x}}\, e^{-x} b_n^{-x} (-\log b_n)^j \qquad (x \to \infty). \tag{3.27}$$

A similar estimate in (3.25) coupled with (3.27) and (3.23) then shows that

$$I_2 \sim x^x (\log x)^{a_n} e^{-x} b_n^{-x} \sqrt{\frac{2\pi}{x}} \qquad (x \to \infty) \tag{3.28}$$

and from (3.13), (3.18) and (3.20) it follows that

$$M[\phi_n;\, x] \sim x^x (\log x)^{a_n} e^{-x} b_n^{-x} \sqrt{\frac{2\pi}{x}} \qquad (x \to \infty). \tag{3.29}$$

Thus, the sequence $\{M[\phi_n;\, x]\}$ is indeed an asymptotic sequence under either conditions (a′) or (b′), in agreement with Theorem 1.

4. Asymptotic expansions. We will now use the previous results to deduce asymptotic expansions of Mellin transforms. Let $\{\psi_n(t)\}$ be an asymptotic sequence as $t \to \infty$ and $f(t)$, $f_k(t)$, $k = 0, 1, \cdots$, functions such that for each nonnegative integer n

$$f(t) = \sum_{k=0}^{n} f_k(t) + o(\psi_n(t)) \qquad (t \to \infty). \tag{4.1}$$

Then we say that $\Sigma f_k(t)$ is a *generalized asymptotic expansion of $f(t)$ with respect to the asymptotic sequence* $\{\psi_n(t)\}$ and write

$$f(t) \sim \sum_{k=0}^{\infty} f_k(t), \qquad \{\psi_k(t)\} \qquad (t \to \infty). \tag{4.2}$$

Our main result is the following analogue of Watson's lemma for Mellin transforms.

THEOREM 2. *Let $\{\psi_n(t)\}$ be an asymptotic sequence for $t \to \infty$ which is equivalent to a sequence $\{\psi_n(t)\}$ satisfying the conditions of Theorem 1. Let R be an unbounded region of the complex plane such that $x \to +\infty$ as $z \to \infty$ in R, and let $\{X_n(z)\}$ be a sequence equivalent to $\{M[\phi_n;\, x]\}$ as $z \to \infty$ in R. If, under these circumstances,*

$$f(t) \sim \sum_{k=0}^{\infty} f_k(t), \qquad \{\psi_k(t)\} \qquad (t \to \infty) \tag{4.3}$$

where $f(t)$ and all the $f_k(t)$ are in M, then

$$M[f, z] \sim \sum_{k=0}^{\infty} M[f_k;\, z], \qquad \{X_k(z)\} \tag{4.4}$$

as $z \to \infty$ in R.

Proof. Fix n and set

$$g(t) = f(t) - \sum_{k=0}^{n} f_k(t), \qquad h(t) = \phi_n(t). \tag{4.5}$$

Then $g(t)$ and $h(t)$ are in M, $\phi_n(t) \geqq 0$ and $\phi_n(t) > 0$ for $t_n < t < \infty$. Since by (4.3) and (4.5), $g(t) = o(\psi_n(t))$ as $t \to \infty$ and $\{\psi_n(t)\}$ and $\{\phi_n(t)\}$ are equivalent asymptotic sequences, we also have $g(t) = o(\phi_n(t))$ as $t \to \infty$. Applying Lemma 1 to $g(t)$ and $h(t)$, it follows that

$$M[g;\, z] = o(M[\phi_n;\, x]) \tag{4.6}$$

as $z \to \infty$ in R. However since $\{M[\phi_n;\, x]\}$ and $\{X_n(z)\}$ are equivalent asymptotic sequences as $z \to \infty$ in R, it is also true that

$$M[g;\, z] = o(X_n(z)) \tag{4.7}$$

as $z \to \infty$ in R, or

$$M[f;\, z] - \sum_{k=0}^{n} M[f_k;\, z] = o(X_n(z)) \tag{4.8}$$

as $z \to \infty$ in R. Since (4.8) holds for each n, we have the conclusion of the theorem. □

Note that in the real case $z = x$ and one may take $X_n(z) = M[\phi_n;\, x]$. Theorem 2 extends Sidi's Theorem 2.1 [7] to a wide class of asymptotic sequences and, in view of a remark following (3.4), covers some of the examples in [7] too.

Example 4. Consider the integral

$$I(x) = \int_0^{\infty} t^{x-1}(1+t)^{-1/2}\left(1 + b\exp\left(a(1+t)^{1/2}\right)\right)^{-\nu} dt \tag{4.9}$$

where a, b and ν are fixed positive constants. The asymptotic expansion of $I(x)$ is not easily obtained from Sidi's theory. However the function

$$f(t)=(1+t)^{-1/2}\left(1+b\exp\left(a(1+t)^{1/2}\right)\right)^{-\nu} \tag{4.10}$$

in the integrand of (4.9) is in M, and

$$f(t)\sim\sum_{k=0}^{\infty}f_k(t),\qquad\{\psi_k(t)\}\qquad(t\to\infty), \tag{4.11}$$

where

$$f_k(t)=(1+t)^{-1/2}\binom{-\nu}{k}b^{-(\nu+k)}\exp\left(-a(k+\nu)(1+t)^{1/2}\right) \tag{4.12}$$

and

$$\psi_k(t)=(1+t)^{-1/2}\exp\left(-a(k+\nu)(1+t)^{1/2}\right) \tag{4.13}$$

both satisfy the requirements of Theorem 2. Consequently, since $I(x)=M[f;\ x]$,

$$I(x)\sim\sum_{k=0}^{\infty}M[f_k;\ x],\qquad\{M[\psi_k;\ x]\} \tag{4.14}$$

as $x\to\infty$. From [5, p. 254], for example, we have

$$M[\psi_k;\ x]=\frac{2}{\pi^{1/2}}\frac{\Gamma(x)K_{x-1/2}(a(k+\nu))}{(a(k+\nu)/2)^{x-1/2}}. \tag{4.15}$$

Note that the asymptotic behavior of $K_x(a)$ for fixed argument and large order is

$$K_x(a)\sim\left(\frac{\pi}{2x}\right)^{1/2}\left(\frac{2x}{ea}\right)^{x}\qquad(a\text{ fixed, }x\to+\infty); \tag{4.16}$$

see, for example [5, p. 328]. Combining the above results and applying Theorem 2, we obtain

(4.17)

$$I(x)\sim\frac{2\Gamma(x)}{\pi^{1/2}}\sum_{k=0}^{\infty}\binom{-\nu}{k}b^{-(\nu+k)}\frac{K_{x-1/2}(a(k+\nu))}{(a(k+\nu)/2)^{x-1/2}},\qquad\left\{\frac{\Gamma(x)K_{x-1/2}(a(k+\nu))}{(a(k+\nu)/2)^{x-1/2}}\right\}$$

as $x\to+\infty$.

Acknowledgments. The author would like to thank Professor R. Wong for informing him about this problem, for his comments on an earlier version of this paper and for his hospitality during the summer of 1984. Also to be thanked are Professors B. Seymour and L. Mysak of the University of British Columbia Institute of Applied Mathematics where this work initially began. Lastly, thanks are due to the referee and Professor F. W. J. Olver, whose comments and suggestions helped to improve this paper.

REFERENCES

[1] G. Doetsch, *Handbuch der Laplace Transformation*, Birkhaüser Verlag, Basel and Stuttgart, 1972.

[2] A. Erdélyi, *General asymptotic expansions of Laplace integrals*, Arch. Rational Mech. Anal., 7 (1961), pp. 1–20.

[3] R. A. Handelsman and J. S. Lew, *Asymptotic expansions of a class of integral transforms via Mellin transforms*, Arch. Rational Mech. Anal., 35 (1969), pp. 382–396.

[4] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

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

[6] E. RIEKSTINS, *Asymptotic Expansions of Integrals*, Izdat. Zinatne, Riga, 1977. (In Russian.)

[7] A. SIDI, *Asymptotic expansions of Mellin transforms and analogues of Watson's lemma*, this Journal, 16 (1985), pp. 896–906.

[8] E. WAGNER, *Taubersche satze fur die Mellin-transformation*, Wiss. Z. Univ. Halle, XV66M, H. 4 pp. 617–624.

[9] R. WONG, *Error bounds for asymptotic expansions of integrals*, SIAM Rev., 22 (1980), pp. 401–435.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

001

POSITIVE SOLUTION OF A PROBLEM OF EMDEN–FOWLER TYPE WITH A FREE BOUNDARY*

GEORGES IFFLAND†

Abstract. A sufficient condition is given for the existence of a solution for a generalized Emden–Fowler problem with a free end point. In the special case of the Emden–Fowler equation, this condition is also necessary; and moreover, a monotone iteration scheme holds for the approximation of a positive solution.

Key words. Emden–Fowler equation, nonlinear boundary value problem, free boundary problem, monotone iteration method

AMS(MOS) subject classifications. 34B15, 34A45

1. Introduction. We are interested in the question of existence of a solution of the following free end point problem. Find $T>0$ and $y\in C[0,T]\cap C^1[0,T)\cap C^2(0,T)$ such that

$$\begin{aligned}&y''(t)+q(t)y(t)^\gamma=0, \qquad t\in(0,T),\\ &y(0)=y(T)=0, \qquad y'(0)=\alpha, \qquad (1.1)\\ &y(t)>0, \qquad t\in(0,T),\end{aligned}$$

where $\gamma\geqq 1$, $\alpha>0$ and q, a positive function, are given. With $q(t):=t^\beta$, (1.1) gives the Emden–Fowler equation [10]. Its origin lies in theories concerning gaseous dynamics in astrophysics around the turn of the century [2]; see [10] for more recent applications. In this special case, (1.1) can be written

$$\begin{aligned}&y''(t)+t^\beta y(t)^\gamma=0, \qquad t\in(0,T),\\ &y(0)=y(T)=0, \qquad y'(0)=\alpha, \qquad (1.2)\\ &y(t)>0, \qquad t\in(0,T).\end{aligned}$$

The question of existence of a solution of (1.1) is related to that of existence of oscillatory solutions of the differential equation

$$y''(t)+q(t)y(t)^\gamma=0. \qquad (1.3)$$

This problem has been profusely studied; see for example [4], [10] or [3]. In this case, γ is supposed to be the quotient of odd, positive integers. For $\gamma:=2n-1$, where $n=2,3,\cdots$, a result of Atkinson [1] allows one to show the existence of a solution of (1.1) if $\int^\infty tq(t)\,dt=\infty$; for problem (1.2), this condition becomes $\beta>-2$. Owing to a theorem of Hinton [5], we are able to give a more general (sufficient) condition for the existence of a solution of (1.1). Moreover, this condition is necessary and sufficient if $q(t):=t^\beta$; it then becomes $\gamma+2\beta+3>0$.

Problem (1.2) can be set in the form of an eigenvalue problem of the following type

$$\begin{aligned}&x''(\tau)+\lambda a(\tau)x(\tau)^\gamma=0, \qquad \tau\in(0,1),\\ &x(0)=x(1)=0, \qquad x'(0)=1, \qquad (1.4)\\ &x(\tau)>0, \qquad \tau\in(0,1).\end{aligned}$$

* Received by the editors December 21, 1984, and in revised form May 15, 1985.

† Département de mathématiques, Ecole polytechnique fédérale, CH-1015 Lausanne, Switzerland.

A monotone iteration method has been proposed by Luning and Perry [7], in order to establish the existence of a solution of (1.4). Their proof needs the assumption $a \in L^1(0, 1)$; for problem (1.2) it becomes $\beta > -1$. Without such a restriction, we show first the convergence of the iterative method to a solution of (1.4), if the existence of a solution is presupposed. The method can then be applied to the Emden–Fowler problem, that is for $a(\tau) := \tau^\beta$, with $\gamma + 2\beta + 3 > 0$.

Let us further point out a problem related to (1.1). We introduce the following initial value problem

$$\begin{aligned} &y''(t) + q(t)y(t)^\gamma = 0, \qquad t \in (0, \infty), \\ &y(0) = 0, \qquad y'(0) = \alpha, \\ &y(t) > 0, \qquad t \in (0, \infty). \end{aligned} \tag{1.5}$$

Under quite general assumptions, (1.3) has a positive solution satisfying the initial condition $y(0) = 0$, $y'(0) = \alpha$. Moreover, y may be chosen such that the following alternative holds: either y is solution of (1.5), or there exists a $T > 0$ such that (T, y) is a solution of (1.1).

Notation and definitions. Let y_1 and y_2 be functions defined on the same interval I (bounded or not). Then $y_1 \leqq y_2$ iff $\forall t \in I$: $y_1(t) \leqq y_2(t)$; $y_1 < y_2$ iff $y_1 \leqq y_2$ and $y_1 \neq y_2$. A function y is *positive* if $y > 0$. A sequence (y_n) of functions is *increasing* iff $y_n \leqq y_{n+1}$ for all $n \in \mathbb{N}$; it is *decreasing* iff $(-y_n)$ is increasing.

2. Existence result for (1.1). Let us introduce the following conditions:

(H1) $\alpha > 0$, $\gamma \geqq 1$.

(H2) $q \in C^2(0, \infty)$, $\forall t > 0$: $q(t) > 0$, $t \mapsto t^\gamma q(t)$ belongs to $L^1(0, 1)$.

(H3) $\eta\eta'' \in L^1(1, \infty)$, where $\eta \in C^2(0, \infty)$ is the function defined by $\eta(t) := [q(t)]^{-1/(\gamma+3)}$.

(H4) $\int_1^\infty [\eta(t)]^{-2}\, dt = \infty$.

(H5) If $\gamma > 1$: $\lim_{t \downarrow 0} t^{-1}\eta(t)[\int_t^\infty \eta(s)|\eta''(s)|\, ds]^{2/(\gamma-1)} = 0$.

The main result is contained in Theorem 2.1; its proof partly reproduces Hinton's [5] and will be given later.

THEOREM 2.1. *Let assumptions* (H1)-(H5) *be satisfied. Then problem* (1.1) *has a unique solution* (T, y), *with* $y \in C[0, T] \cap C^1[0, T) \cap C^2(0, T)$.

LEMMA 2.2. *Let assumptions* (H1) *and* (H2) *be satisfied. Then problem*

$$\begin{aligned} &y''(t) + q(t)y(t)^\gamma = 0, \\ &y(0) = 0, \qquad y'(0) = \alpha \end{aligned} \tag{2.1}$$

has a positive (*local*) *solution.*

Proof. One establishes the existence of a solution by an alternating iteration method, like the one used for example by Stuart [8]. Let us introduce the following iteration scheme

$$y_{n+1}(t) := \alpha t - \int_0^t (t-s)q(s)y_n(s)^\gamma\, ds. \tag{2.2}$$

Set $y_0 := 0$; by (2.2) define y_1: $t \mapsto \alpha t$. Then

$$y_2\colon \quad t \mapsto \alpha t\left(1 - \alpha^{\gamma-1} \int_0^t \left(1 - \frac{s}{t}\right) q(s)s^\gamma\, ds\right).$$

By assumption (H2), a $t^*>0$ can be chosen such that y_2 remains positive on $(0, t^*)$. By means of formula (2.2), an alternating sequence is defined:

$$0=y_0 \leqq y_2 \leqq y_{2n} \leqq y_{2n+1} \leqq y_3 \leqq y_1=\alpha t, \quad \text{on } [0, t^*].$$

Differentiating (2.2):

$$\alpha \geqq y_{n+1}'(t)=\alpha-\int_0^t q(s) y_n(s)^\gamma \, ds \geqq \alpha-\alpha^\gamma \int_0^{t^*} q(s) s^\gamma \, ds, \qquad t \in [0, t^*].$$

The sequence (y_n') being uniformly bounded, the Arzelà-Ascoli theorem implies the existence of a convergent subsequence of (y_n). The subsequence (y_{2n}), increasing and containing a convergent subsequence, is convergent itself to a function $\bar{y}$, uniformly on $[0, t^*]$. Likewise, the decreasing subsequence (y_{2n+1}) converges to a function $\hat{y}$. Moreover

$$\begin{aligned} \bar{y}(t) &= \alpha t-\int_0^t (t-s) q(s) \hat{y}(s)^\gamma \, ds \\ &\leqq \alpha t-\int_0^t (t-s) q(s) \bar{y}(s)^\gamma \, ds=\hat{y}(t), \qquad t \in [0, t^*]. \end{aligned} \tag{2.3}$$

Define $w:=\hat{y}-\bar{y}$. The following inequality holds:

$$\forall t \in [0, t^*]: \quad 0 \leqq \hat{y}(t)^\gamma-\bar{y}(t)^\gamma \leqq \gamma y_1(t)^{\gamma-1} w(t) \leqq C w(t) \tag{2.4}$$

where $C>0$ is a constant. Introduce (2.4) into (2.3):

$$0 \leqq w(t) \leqq C \int_0^t (t^*-s) q(s) w(s) \, ds, \qquad t \in [0, t^*].$$

By Gronwall's lemma, $w=0$. The function $y:=\bar{y}=\hat{y}$ satisfies

$$y(t)=\alpha t-\int_0^t (t-s) q(s) y(s)^\gamma \, ds, \qquad t \in [0, t^*]$$

and is a solution of (2.1) on $[0, t^*]$. □

LEMMA 2.3. *Every positive solution of* (2.1) *may be extended to a function* $\tilde{y}$ *for which the following alternative holds*:

(i) $\tilde{y}$ *is a solution of* (1.5);

(ii) *there exists a* $T>0$ *such that* $(T, \tilde{y})$ *is a solution of* (1.1).

Proof. Let y be a positive solution of (2.1) and $\tilde{y}$ be the maximal positive extension of y. If $\tilde{y}$ is defined for each $t>0$, we are in case (i) and $\tilde{y}(t)>0$ on $(0, \infty)$. Assume now that y is only defined on a bounded interval I. As $\tilde{y}$ is positive and concave, I must be closed; write $I=[0, T]$. If $\tilde{y}(T)>0$, Peano's existence theorem would allow one to extend $\tilde{y}$ at the the right side of I; but it is in contradiction with the definition of $\tilde{y}$. Hence $\tilde{y}(T)=0$. □

LEMMA 2.4. *Suppose* $\gamma>1$. *Assume conditions* (H2), (H3) *and* (H5) *to be satisfied. Let* y *be a positive,* C^1 *function, such that* $y(0)=0$ *and* $y'(0)>0$. *Then for every* $\chi>0$ *there exists a* $t_0>0$ *such that*

$$y(t_0)>\chi \eta(t_0)\left[\int_{t_0}^\infty \eta(s)|\eta''(s)| \, ds\right]^{2/(\gamma-1)}. \tag{2.5}$$

Proof. By contradiction, assume the existence of a $C>0$, such that for every t, in a neighbourhood of zero, the following inequality holds:

$$y(t) \leqq C \eta(t)\left[\int_t^\infty \eta(s)|\eta''(s)| \, ds\right]^{2/(\gamma-1)}.$$

Dividing by t and letting t tend to zero, we get

$$0 < y'(0) \leqq C \lim_{t \downarrow 0} t^{-1}\eta(t)\left[\int_t^\infty \eta(s)|\eta''(s)|\,ds\right]^{2/(\gamma-1)},$$

in contradiction with assumption (H5). □

Proof of Theorem 2.1. (i) Uniqueness is easily established by using an inequality of (2.4) type and Gronwall's lemma. Then owing to Lemmas 2.2 and 2.3, it is sufficient to show the nonexistence of a solution of (1.5). In order to do that, suppose, by way of contradiction, that y is a solution of (1.5). If $\gamma = 1$, choose an arbitrary $t_0 > 0$. If $\gamma > 1$, set $\chi := [2(\gamma+1)]^{1/(\gamma-1)}$; then, by Lemma 2.4, a $t_0 > 0$ can be chosen such that

$$y(t_0) > [2(\gamma+1)K^2]^{1/(\gamma-1)}\eta(t_0) \tag{2.6}$$

where $K := \int_{t_0}^\infty \eta(\tau)|\eta''(\tau)|\,d\tau < \infty$, by (H3).

Define

$$h:\quad t \mapsto \int_{t_0}^t [\eta(\tau)]^{-2}\,d\tau, \qquad t \in [t_0, \infty)$$

and

$$x:\quad s \mapsto (y/\eta)(h^{-1}(s)), \qquad s \in [0, \infty)$$

where h^{-1} is the inverse function of h. Since y is a positive solution of (1.3), x cannot vanish on $(0, \infty)$ and satisfies the following differential equation

$$x''(s) + (\eta^3\eta'')(h^{-1}(s))x(s) + x(s)^\gamma = 0, \qquad s \in (0, \infty). \tag{2.7}$$

Define a function z by

$$z(s) := \frac{1}{2}(x'(s))^2 + \frac{1}{\gamma+1}(x(s))^{\gamma+1}, \qquad s \in [0, \infty). \tag{2.8}$$

(ii) We show that z is bounded on $[0, \infty)$. If not, there exists an increasing sequence $s_1 < s_2 < \cdots$ for which $\lim_{i\to\infty} z(s_i) = \infty$; without loss of generality, the sequence can be taken such that

$$z(s_i) > 1 \quad \text{and} \quad z(s_i) = \max\{z(s);\, 0 \leqq s \leqq s_i\}, \quad i = 1, 2, \cdots. \tag{2.9}$$

By (2.7) and (2.8)

$$\forall s \in (0, \infty), \quad z'(s) = -(\eta^3\eta'')(h^{-1}(s))x(s)x'(s). \tag{2.10}$$

From (2.8) we deduce that

$$\forall s \in [0, \infty), \quad x(s) \leqq [(\gamma+1)z(s)]^{1/(\gamma+1)} \quad \text{and} \quad |x'(s)| \leqq [2z(s)]^{1/2}. \tag{2.11}$$

By assumption (H3), a $t_1 > t_0$ can be chosen such that

$$\sqrt{2}(\gamma+1)^{1/(\gamma+1)} \int_{t_1}^\infty \eta(\tau)|\eta''(\tau)|\,d\tau \leqq \frac{1}{2}. \tag{2.12}$$

Consider the s_i such that $s_i > h(t_1)$ and integrate (2.10)

$$z(s_i) - z(h(t_1)) = -\int_{h(t_1)}^{s_i} (\eta^3\eta'')(h^{-1}(\sigma))x(\sigma)x'(\sigma)\,d\sigma.$$

Taking into account (2.9), (2.11) and (2.12), it follows that

$$\begin{aligned}0&\leqq z(s_i)-z(h(t_1))\\&\leqq\sqrt{2}(\gamma+1)^{1/(\gamma+1)}z(s_i)^{1/2+1/(\gamma+1)}\int_{h(t_1)}^{s_i}|(\eta^3\eta'')(h^{-1}(\sigma))|\,d\sigma\\&\leqq\sqrt{2}(\gamma+1)^{1/(\gamma+1)}z(s_i)\int_{t_1}^{\infty}\eta(\tau)|\eta''(\tau)|\,d\tau\leqq\frac{1}{2}z(s_i).\end{aligned}$$

Whence $z(s_i)\leqq 2z(h(t_1))$, and this is in contradiction to the assumption that $z(s_i)\to\infty$ as $i\to\infty$.

(iii) Let us show that z tends to a limit L as $s\to\infty$. Since z is bounded, we define a constant

$$C:=\sqrt{2}(\gamma+1)^{1/(\gamma+1)}\sup\{z(s)^{1/2+1/(\gamma+1)};\ s\in(0,\infty)\}.$$

Let $\varepsilon>0$ be arbitrary. By assumption (H3), a $t_2>t_0$ can be chosen such that

$$\int_{t_2}^{\infty}\eta(\tau)|\eta''(\tau)|\,d\tau<\frac{\varepsilon}{C}.$$

Integrate (2.10) with $s\geqq h(t_2)$

$$\begin{aligned}|z(s)-z(h(t_2))|&\leqq\int_{h(t_2)}^{s}|(\eta^3\eta'')(h^{-1}(\sigma))|x(\sigma)|x'(\sigma)|\,d\sigma\\&\leqq C\int_{h(t_2)}^{s}|(\eta^3\eta'')(h^{-1}(\sigma))|\,d\sigma<\varepsilon.\end{aligned}$$

The existence of the limit L follows.

(iv) Let us show that $L>0$. Consider first the case $\gamma=1$. From (2.10)

$$\frac{d}{ds}(\ln z(s))=-(\eta^3\eta'')(h^{-1}(s))x(s)x'(s)(z(s))^{-1},\qquad s\in(0,\infty).$$

Using estimates (2.11)

$$|\ln z(s)-\ln z(0)|\leqq 2\int_{t_0}^{\infty}\eta(\tau)|\eta''(\tau)|\,d\tau=2K<\infty.$$

Thus $\ln z(s)$ is bounded for $s\in[0,\infty)$; hence $L>0$. Consider now the case $\gamma>1$ and, by way of contradiction, suppose $L=0$. The function z being bounded, one can define

$$L_1:=\sup\{|x'(s)|;\ s\in[0,\infty)\}\quad\text{and}\quad L_2:=\sup\{x(s);\ s\in[0,\infty)\}.$$

Putting this in (2.10)

$$z(s)=\int_{s}^{\infty}(\eta^3\eta'')(h^{-1}(\sigma))x(\sigma)x'(\sigma)\,d\sigma\leqq KL_1L_2.$$

From (2.11) we obtain

$$L_2^{\gamma}\leqq(\gamma+1)KL_1\quad\text{and}\quad L_1\leqq 2KL_2.$$

Whence $L_2^{\gamma-1}\leqq 2(\gamma+1)K^2$ and

$$y(t_0)=\eta(t_0)x(0)\leqq\eta(t_0)L_2\leqq\eta(t_0)[2(\gamma+1)K^2]^{1/(\gamma-1)},$$

in contradiction to (2.6).

(v) Since $L>0$, one can find an s_0 such that $z(s) \geqq \frac{1}{2}L$ for all $s \geqq s_0$. First case: x' has an infinite number of zeros $s_1 < s_2 < \cdots$ on $[s_0, \infty)$. On every interval $[s_1, s_i]$, the minimum of x is achieved at a zero s_k of x' and

$$\frac{L}{2} \leqq z(s_k) \leqq \frac{1}{\gamma+1}(x(s_k))^{\gamma+1}.$$

Hence

$$x(s) \geqq x(s_k) \geqq (\tfrac{1}{2}(\gamma+1)L)^{1/(\gamma+1)}, \qquad s \in [s_1, \infty).$$

By (2.7), the following holds on $[s_1, \infty)$:

$$|x'(s)| = \left| \int_{s_1}^{s} x''(\sigma)\, d\sigma \right| \geqq \int_{s_1}^{s} x(\sigma)^\gamma \, d\sigma - \left| \int_{s_1}^{s} (\eta^3 \eta'')(h^{-1}(\sigma)) x(\sigma)\, d\sigma \right|$$
$$\geqq \left(\frac{1}{2}(\gamma+1)L \right)^{\gamma/(\gamma+1)} (s - s_1) - KL_2.$$

It follows that x' is unbounded and therefore we have a contradiction. Second case: x' is monotone on some interval $[s_0, \infty)$. Since x is bounded, $x(s) \to M$ as $s \to \infty$. M cannot be zero, for in this case $|x'(s)| \to \sqrt{2L}$ and x would be unbounded. So assume $M>0$. An s_1 can be found such that $x(s) \geqq M/2$ for all $s \geqq s_1$. As in the first case, we establish

$$|x'(s) - x'(s_1)| \geqq \left(\frac{M}{2} \right)^\gamma (s - s_1) - KL_2, \qquad s \in [s_1, \infty).$$

A contradiction follows again, since x' must be bounded. We conclude that x must vanish, and so must y. □

3. Existence result for (1.5). Let us introduce two additional conditions:

(H6) $\lim_{t \downarrow 0} t^{\gamma+2} q(t) = 0$.

(H7) $\forall t > 0$: $(\gamma+3)q(t) + 2tq'(t) \leqq 0$.

We show first the nonexistence of a solution of (1.1); by Lemmas 2.2 and 2.3, the existence of a (unique) solution of (1.5) follows.

LEMMA 3.1. *Let assumptions* (H1), (H2), (H6) *and* (H7) *be satisfied. Then* (1.1) *has no solution.*

Proof. Every solution of the differential equation (1.3) must satisfy the following identity

$$\frac{d}{dt}\left[t(y'(t))^2 - y(t)y'(t) + \frac{2}{\gamma+1} tq(t) y(t)^{\gamma+1} \right]$$
$$= \frac{1}{\gamma+1}[(\gamma+3)q(t) + 2tq'(t)] y(t)^{\gamma+1}. \tag{3.1}$$

By contradiction, assume (T, y) to be a solution of (1.1). Integrate (3.1) between 0 and T; according to the boundary conditions and hypothesis (H6):

$$T(y'(T))^2 = \frac{1}{\gamma+1} \int_0^T ((\gamma+3)y(\tau) + 2\tau q'(\tau)) y(\tau)^{\gamma+1}\, d\tau.$$

This is in contradiction to the hypothesis (H7), since $y'(T) > 0$. □

THEOREM 3.2. *Let assumptions* (H1), (H2), (H6) *and* (H7) *be satisfied. Then problem* (1.5) *has a solution.*

4. A monotone iteration method. We now develop some results given by Luning and Perry [7]. They apply to the eigenvalue problem (1.4). The relation with the problem (1.2) will be shown in the next section. We admit here any nonnegative value for γ. The following condition is prescribed for the function a:

(H8) $a \in C(0,1), \forall \tau \in (0,1)$: $a(\tau) > 0$, $\tau \mapsto \tau^\gamma(1-\tau)a(\tau)$ belongs to $L^1(0,1)$. Introduce the following iteration scheme for $n = 0, 1, \cdots$:

$$\lambda_n := \left[\int_0^1 (1-\theta)a(\theta)x_n(\theta)^\gamma\,d\theta\right]^{-1},$$

$$x''_{n+1}(\tau) = -\lambda_n a(\tau)x_n(\tau)^\gamma, \qquad \tau \in (0,1), \tag{4.1}$$

$$x_{n+1}(0) = x_{n+1}(1) = 0.$$

Suppose that $\gamma \geqq 0$ and assume (H8) to hold. Setting x_0: $\tau \mapsto \tau$, $\tau \in [0,1]$, one defines by (4.1) two sequences (λ_n) and (x_n); and for each n: $x'_n(0) = 1$. Let us recall a result drawn from [7].

LEMMA 4.1. *Suppose $\gamma \geqq 0$ and let a satisfy the condition* (H8). *Then the sequence (λ_n) defined above is increasing, and the sequence (x_n) is decreasing; moreover, $\forall n \in \mathbf{N}$ and $\forall \tau \in (0,1)$: $0 < x_{n+1}(\tau) < x_n(\tau)$.*

The following theorem and its corollary are drawn from [6].

THEOREM 4.2. *Suppose that $\gamma \geqq 0$ and let a satisfy* (H8). *Let (λ_n) and (x_n) be the sequences defined as above. Assume the existence of a solution (λ, x) of* (1.4). *Then the sequence (λ_n) tends to a limit $\bar\lambda$, the sequence (x_n) converges uniformly on $[0,1]$ to a function $\hat x$; $(\bar\lambda, \hat x)$ is a solution of* (1.4), *and $\bar\lambda \leqq \lambda$, $\hat x \geqq x$.*

Proof. (i) Let us show that $\forall n \in \mathbf{N}$: $\lambda_n < \lambda$ and $x_n \geqq x$. Of course $x_0 > x$; moreover for each $\mu \in (0,1)$ there exists a unique $\tau_0(\mu) \in (0,1)$ such that $\mu x_0(\tau_0) = x(\tau_0)$. Make the following inductive hypothesis: $x_n > x$ and for each $\mu \in (0,1)$ there exists at most one $\tau_n(\mu) \in (0,1)$ such that $\mu x_n(\tau_n) = x(\tau_n)$. If this is true, an obvious first consequence is that $\lambda_n < \lambda$. Secondly: $x_{n+1} > x_n$. To show this, set $\mu := (\lambda_n/\lambda)^{1/\gamma} < 1$ and $w := x - x_{n+1}$. Since $x_n(0) = x(0)$ and $x'_n(0) = x'(0) = 1$, there exists a $\delta > 0$ such that

$$w''(\tau) = a(\tau)[\lambda_n x_n(\tau)^\gamma - \lambda x(\tau)^\gamma] = \lambda a(\tau)[\mu^\gamma x_n(\tau)^\gamma - x(\tau)^\gamma] < 0$$

if $\tau \in (0,\delta)$. We obtain $w(0) = w'(0) = w(1) = 0$ and $w''(\tau) < 0$ for $\tau \in (0,\delta)$; by the inductive hypothesis, w'' changes sign at most once in $(0,1)$. We conclude that $w(\tau) < 0$ for $\tau \in (0,1)$. Third consequence: for each $\mu \in (0,1)$, there exists at most one $\tau_{n+1}(\mu) \in (0,1)$ such that $\mu x_{n+1}(\tau_{n+1}) = x(\tau_{n+1})$. Indeed, set $\nu := (\mu\lambda_n/\lambda)^{1/\gamma} < 1$ and $v := x - \mu x_{n+1}$. Since $x_n(0) = x(0)$ and $x'_n(0) = x'(0) = 1$, there exists a $\delta > 0$ such that

$$v''(\tau) = a(\tau)[\mu\lambda_n x_n(\tau)^\gamma - \lambda x(\tau)^\gamma] = \lambda a(\tau)[\nu^\gamma x_n(\tau)^\gamma - x(\tau)^\gamma] < 0$$

if $\tau \in (0,\delta)$. We have $v(0) = v(1) = 0$, $v'(0) = 1 - \mu > 0$ and $v''(\tau) < 0$ for $\tau \in (0,\delta)$; by the inductive hypothesis, v'' vanishes at most once in $(0,1)$. Therefore v cannot vanish more than once in $(0,1)$.

(ii) The sequence (λ_n), being increasing and bounded by λ, tends to a limit $\bar\lambda \leqq \lambda$. The sequence (x_n) is decreasing, and $x_n \geqq x$. For $\tau \in [0,1]$, we get from (4.1)

$$|x'_{n+1}(\tau)| \leqq 1 + \lambda_n \int_0^\tau a(\theta)x_n(\theta)^\gamma\,d\theta$$

$$\leqq 1 + \bar\lambda \int_0^1 a(\theta)x_0(\theta)^\gamma\,d\theta < 1 + \bar\lambda \int_0^1 a(\theta)\theta^\gamma\,d\theta < \infty.$$

By the Arzelà-Ascoli theorem, the decreasing sequence (x_n) converges to a function $\hat{x} \geqq x$ (since it contains a convergent subsequence). From (4.1) again, one has

$$x_{n+1}(\tau) = \tau - \lambda_n \int_0^\tau (\tau - \theta) a(\theta) x_n(\theta)^\gamma \, d\theta, \qquad \tau \in [0, 1].$$

Making n tend to infinity

$$\hat{x}(\tau) = \tau - \bar{\lambda} \int_0^\tau (\tau - \theta) a(\theta) \hat{x}(\theta)^\gamma \, d\theta.$$

It is easily checked that $(\bar{\lambda}, \hat{x})$ is a solution of (1.4). □

COROLLARY 4.3. *Let (λ_n) and (x_n) be the two sequences defined as above. Then the following alternatives hold:*

(i) *The sequence (λ_n) is unbounded. In this case, problem* (1.4) *has no solution and the sequence (x_n) tends to the zero function.*

(ii) *The sequence (λ_n) converges. In this case, problem* (1.4) *has a solution $(\bar{\lambda}, \hat{x})$, such that $\lambda_n \to \bar{\lambda}$ and x_n converges uniformly to $\hat{x}$. If (λ, x) is another solution of* (1.4), *then $\lambda > \bar{\lambda}$ and $x < \hat{x}$.*

5. Emden–Fowler problem (1.2). Consider now the problem (1.2), which corresponds to the special case of (1.1) when $q(t) := t^\beta$.

THEOREM 5.1. *Let $\gamma \geqq 1$ and β be real numbers. Then problem* (1.2) *has a solution if and only if $\gamma + 2\beta + 3 > 0$.*

Proof. (i) Suppose that $\gamma + 2\beta + 3 > 0$. Then $\gamma + \beta + 1 > 0$ (for $\beta \leqq -2$: $\gamma + \beta + 1 = (\gamma + 2\beta + 3) - (\beta + 2) > 0$ and for $\beta > -2$: $\gamma + \beta + 1 = (\gamma - 1) + (\beta + 2) > 0$). Hypothesis (H2) is satisfied iff $\gamma + \beta + 1 > 0$. Conditions (H3) and (H5) hold iff $\gamma + 2\beta + 3 > 0$, and (H4) iff $\gamma + 2\beta + 3 \geqq 0$. Thus Theorem 2.1 applies and (1.2) has a unique solution (T, y), with $y \in C[0, T] \cap C^1[0, T) \cap C^2(0, T)$.

(ii) Notice that if $y \in C^1[0, T)$ satisfies (1.2), then $\gamma + \beta + 1 > 0$. Indeed, for any $t_0 \in (0, \varepsilon)$

$$y'(\varepsilon) = y'(t_0) - \int_{t_0}^{\varepsilon} \tau^\beta y(\tau)^\gamma \, d\tau.$$

By continuity at the origin

$$\int_0^\varepsilon \tau^\beta y(\tau)^\gamma \, d\tau = \alpha - y'(\varepsilon).$$

Therefore $t \mapsto t^\beta y(t)^\gamma$ belongs to $L^1(0, \varepsilon)$. Taking ε small enough and a $\bar{t} \in (0, \varepsilon)$, we have

$$t^\beta y(t)^\gamma = t^{\beta+\gamma} y'(\bar{t})^\gamma \geqq \left(\frac{\alpha}{2}\right)^\gamma t^{\beta+\gamma}, \qquad t \in (0, \varepsilon).$$

Thus $t \mapsto t^{\beta+\gamma}$ is integrable.

(iii) Suppose now that $\gamma + 2\beta + 3 \leqq 0$ and, by way of contradiction, that (T, y) is a solution of (1.2). According to the preceding notice: $\gamma + \beta + 1 > 0$; thus condition (H6) is satisfied. On the other hand, (H7) holds iff $\gamma + 2\beta + 3 \leqq 0$. Therefore (1.2) has no solution, by Lemma 3.1. □

Suppose that $\gamma + 2\beta + 3 > 0$. We establish now the relation between problems (1.2) and (1.4). Let (T, y) be the solution of (1.2) and define $\tau := t/T$, $x : \tau \mapsto (\alpha T)^{-1} y(T\tau)$, for $\tau \in [0, 1]$, and $\lambda := \alpha^{\gamma-1}T^{\gamma+\beta+1}$. Then (λ, x) satisfies

$$\begin{aligned} &x''(\tau)+\lambda\tau^{\beta}x(\tau)^{\gamma}=0, \qquad \tau\in(0,1),\\ &x(0)=x(1)=0, \qquad x'(0)=1, \\ &x(\tau)>0, \qquad \tau\in(0,1). \end{aligned} \tag{5.1}$$

Reciprocally let (λ, x) be a solution of (5.1). Define $T := \lambda^{1/(\gamma+\beta+1)}\alpha^{(\gamma-1)/(\gamma+\beta+1)}$ (note that $\gamma+\beta+1>0$), $t := T\tau$ and $y : t \mapsto \alpha T x(t/T)$, for $t \in [0, T]$. Then (T, y) is a solution of (1.2). One thus gets the uniqueness of the solution of (5.1); this solution may be approximated by the monotone iteration method of the preceding section. Define $x_0(\tau) := \tau$, for $\tau \in [0, 1]$, and

$$\begin{aligned} &\lambda_n := \left[\int_0^1 (1-\theta)\theta^{\beta}x_n(\theta)^{\gamma}\,d\theta\right]^{-1}, \\ &x''_{n+1}(\tau) = -\lambda_n\tau^{\beta}x_n(\tau)^{\gamma}, \qquad \tau\in(0,1), \\ &x_{n+1}(0)=x_{n+1}(1)=0 \quad \text{for } n=0,1,\cdots. \end{aligned} \tag{5.2}$$

The existence of the solution of (1.2), by Theorem 5.1, involves the existence of the solution of (5.1). The condition (H8) being satisfied, Theorem 4.2 yields the convergence of the method (5.2). The sequence (λ_n) converges to λ, the sequence (x_n) converges uniformly to x, and (λ, x) is a solution of (5.1). With the inverse transformation, one gets the solution (T, y) of (1.2).

6. Remarks. (i) In this paper, a sufficient condition for the existence of a solution of (1.1) has been obtained. We do not know any necessary and sufficient condition, except in the special case (1.2).

(ii) The eigenvalue problem (1.4) has a solution if the function a belongs to $L^1(0, 1)$ and satisfies (H8); see [7]. This condition is not optimal, since if $a(\tau) := \tau^{\beta}$, the necessary and sufficient condition for existence is $\gamma+2\beta+3>0$ (thus a is not necessarily integrable).

(iii) For $\gamma<1$, the study of problems (1.1) and (1.4) requires other methods. Let us mention a paper of Taliaferro [9], which is devoted to problem (1.4), with $\gamma<0$. In [7], Luning and Perry studied (1.4) with $\gamma>-1$ and $a\in L^1(0, 1)$.

(iv) More generally, one can consider a differential equation of the type

$$(p(s)u'(s))' + r(s)u(s)^{\gamma} = 0 \tag{6.1}$$

with $p \in C^1(0, \infty)$ and $p(s)>0$ for all $s>0$. Moreover, suppose that $1/p \in L^1_{\text{loc}}(0, \infty)$ or $1/p \in L^1(1, \infty)$; we then introduce one of the following transformations:

If $1/p \in L^1(0, 1)$ and $1/p \notin L^1(1, \infty)$, set

$$t := \int_0^s \frac{d\sigma}{p(\sigma)} \quad \text{and} \quad y(t) := u(s).$$

If $1/p \notin L^1(0, 1)$ and $1/p \in L^1(1, \infty)$, set

$$t := \left[\int_s^{\infty} \frac{d\sigma}{p(\sigma)}\right]^{-1} \quad \text{and} \quad y(t) := tu(s).$$

If $1/p \in L^1(0, \infty)$, set

$$t_0 := \left[\int_0^{\infty} \frac{d\sigma}{p(\sigma)}\right]^{-1}, \quad t := \left[\int_s^{\infty} \frac{d\sigma}{p(\sigma)}\right]^{-1} - t_0 \quad \text{and} \quad y(t) := (t+t_0)u(s).$$

In each case: $0 \leqq t < \infty$ if $0 \leqq s < \infty$, and y is a solution of (1.3) if u is a solution of (6.1).

Acknowledgments. The author wishes to thank Dr. P. Bader, Dr. C. Khanmy, Professor C. Stuart and Professor B. Zwahlen for their kind interest.

REFERENCES

[1] F. V. ATKINSON, *On second-order non-linear oscillations*, Pacific J. Math., 5 (1955), pp. 643-647.

[2] S. CHANDRASEKHAR, *An Introduction to the Study of Stellar Structure*, Dover Reprint, New York, 1967 (1st ed. 1939).

[3] L. H. ERBE, *Oscillation and nonoscillation properties for second order nonlinear differential equations*, in Equadiff 82, H. W. Knobloch and K. Schmitt, eds., Lectures Notes in Mathematics 1017, Springer, Berlin, 1983.

[4] J. W. HEIDEL AND D. B. HINTON, *The existence of oscillatory solutions for a nonlinear differential equation*, this Journal, 3 (1972), pp. 344-351.

[5] D. HINTON, *An oscillation criterion for solutions of* $(ry')'+qy^{\gamma}=0$, Michigan Math. J., 16 (1969), pp. 349-352.

[6] G. IFFLAND, *Itérations monotones dans un espace de Banach ordonné et applications aux équations de Thomas-Fermi et de Lane-Emden-Fowler*, thèse no 500, Ecole polytechnique fédérale, Lausanne, 1983.

[7] C. D. LUNING AND W. L. PERRY, *Positive solutions of negative exponent generalized Emden-Fowler boundary value problems*, this Journal, 12 (1981), pp. 874-879.

[8] C. A. STUART, *Integral equations with decreasing nonlinearities and applications*, J. Differential Equations, 18 (1975), pp. 202-216.

[9] S. D. TALIAFERRO, *A nonlinear singular boundary value problem*, Nonlinear Anal., 3 (1979), pp. 897-904.

[10] J. S. W. WONG, *On the generalized Emden-Fowler equation*, SIAM Rev., 17 (1975), pp. 339-360.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

BEST INTERVAL LENGTHS FOR BOUNDARY VALUE PROBLEMS FOR THIRD ORDER LIPSCHITZ EQUATIONS*

JOHNNY HENDERSON†

Abstract. For the third order differential equation $y''' = f(t, y, y', y'')$, where $|f(t, y_1, y_2, y_3) - f(t, z_1, z_2, z_3)| \leqq \sum_{i=1}^{3} k_i |y_i - z_i|$ on $(a, b) \times \mathbf{R}^3$, subintervals (α, β) of (a, b) of maximal length are characterized, in terms of the Lipschitz coefficients k_i, $i = 1, 2, 3$, for the existence of unique solutions of certain two-point and three-point boundary value problems. The techniques for establishing best interval length involve applications of the Pontryagin Maximum Principle coupled with uniqueness implies existence arguments. For the case $k_i = 1$, $i = 1, 2, 3$, comparisons are made with interval lengths obtained via standard applications of the Contraction Mapping Principle.

Key words. boundary value problem, Lipschitz equation, Pontryagin Maximum Principle, optimal length interval

AMS (MOS) subject classifications. Primary 34B10, 34B15; secondary 49A10, 49A36

1. Introduction.

We shall be concerned with solutions of boundary value problems for the third order differential equation

$$y''' = f(t, y, y', y'') \tag{1}$$

where we assume throughout that

(A) $f(t, y_1, y_2, y_3)$: $(a, b) \times \mathbf{R}^3 \to \mathbf{R}$ is continuous, and

(B) f satisfies the Lipschitz condition

$$|f(t, y_1, y_2, y_3) - f(t, z_1, z_2, z_3)| \leqq \sum_{i=1}^{3} k_i |y_i - z_i|$$

for each (t, y_1, y_2, y_3), $(t, z_1, z_2, z_3) \in (a, b) \times \mathbf{R}^3$.

In particular, we will address the question concerning interval length bounds on subintervals of (a, b), in terms of the Lipschitz coefficients k_i, $i = 1, 2, 3$, on which certain two-point and three-point boundary value problems for (1) have unique solutions. Such questions have been commonly resolved by various applications of the Contraction Mapping Principle; for example, see [1]-[7], [15], [20]. A limitation of the methods using the Contraction Mapping Principle is the fact that often unique solutions of the boundary value problems exist on longer subintervals of (a, b). Recently, for two types of the boundary value problems for (1) that we consider here, Aftabizadeh and Wiener [1] sharpened some of the previous bounds by using the Contraction Mapping Principle, after having transformed their problems into boundary value problems for a second order integro-differential equation. We further mention that in [4], a weight function technique previously used by Collatz was employed in obtaining best possible subinterval lengths in the cases (with one exception) where (1) is independent of y'' or independent of both y' and y''.

The purpose of this paper will be to characterize in terms of the Lipschitz coefficients k_i, $i = 1, 2, 3$, the subintervals (α, β) of (a, b) of *maximal length* for the existence and uniqueness of solutions of certain boundary value problems for (1). We accomplish this by applying techniques from optimal control theory that are motivated

* Received by the editors September 16, 1985; accepted for publication (in revised form) January 22, 1986. This research was supported in part by a grant from the Auburn Research Grant-in-Aid program.

† Department of Mathematics, Auburn University, Auburn, Alabama 36849.

by works of Melentsova and Mil'shtein [18] and Melentsova [19], and most notably by the two works of Jackson [12], [13].

Jackson's [12], [13] papers dealt with the cases of conjugate type boundary value problems and right focal type boundary value problems for nth order Lipschitz equations. In terms of the third order equation (1), his results concerned solutions of the conjugate problems,

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y(t_3)=y_3,\qquad a<t_1=t_2<t_3<b, \tag{2}$$

$$y(t_1)=y_1,\quad y(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1<t_2=t_3<b, \tag{3}$$

$$y(t_1)=y_1,\quad y(t_2)=y_2,\quad y(t_3)=y_3,\qquad a<t_1<t_2<t_3<b, \tag{4}$$

and solutions of the right focal problems,

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y''(t_3)=y_3,\qquad a<t_1=t_2<t_3<b, \tag{5}$$

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y''(t_3)=y_3,\qquad a<t_1<t_2=t_3<b, \tag{6}$$

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y''(t_3)=y_3,\qquad a<t_1<t_2<t_3<b. \tag{7}$$

In [13], Jackson proved the first two theorems we present here.

THEOREM 1.1. *Let $h>0$ be the smallest positive number such that there is a solution $x(t)$ of the boundary value problem*

$$x'''=-k_1x-k_2|x'|-k_3|x''|,$$
$$x(0)=x'(0)=x(h)=0,$$

with $x(t)>0$ on $(0,h)$, or $h=+\infty$ if no such solution exists. Then boundary value problems for (1) *satisfying* (2), (3), *or* (4) *have unique solutions, for any assignment of $y_1, y_2, y_3\in\mathbf{R}$, provided $t_3-t_1<h$. Furthermore, this result is best possible for the class of all differential equations that satisfy the Lipschitz condition* (B).

The use made by Jackson of optimal control was via an application of the Pontryagin Maximum Principle in establishing the uniqueness of solutions of boundary value problems, when solutions exist. In the case of conjugate problems, uniqueness implies existence (see [9], [14]), thus, the existence statement of Theorem 1.1. For right focal problems, the second theorem of Jackson's that we state is a uniqueness result.

THEOREM 1.2. *Let $h=\min\{h_1,h_2\}$, where $h_1>0$ is the smallest positive number such that there is a solution $x(t)$ of the boundary value problem*

$$x'''=-k_1x-k_2|x'|-k_3|x''|,$$
$$x(0)=x'(0)=x''(h_1)=0,$$

with $x(t)>0$ on $(0,h_1]$, or $h_1=+\infty$ if no such solution exists, and where $h_2>0$ is the smallest positive number such that there is a solution $y(t)$ of the boundary value problem

$$x'''=-k_1x-k_2|x'|-k_3|x''|,$$
$$x'(0)=x''(0)=x(h_2)=0,$$

with $y(t)>0$ on $[0,h_2)$, or $h_2=+\infty$ if no such solution exists. Then boundary value problems for (1) *satisfying* (5), (6), *or* (7) *have at most one solution, provided $t_3-t_1<h$, and again, this result is best possible.*

Recently, Henderson [10] proved that uniqueness implies existence for solutions of right focal problems. Thus, we can state an analogue of Theorem 1.1.

THEOREM 1.3. *Let $h>0$ be as in Theorem* 1.2. *Then boundary value problems for* (1) *satisfying* (5), (6), *or* (7) *have unique solutions, for any assignment of* $y_1, y_2, y_3 \in \mathbf{R}$, *provided* $t_3-t_1<h$, *and this result is best possible for the class of all differential equations satisfying* (B).

For our purposes here, we shall consider, (as in [8], [11]), boundary value problems for (1) that are "in between" those of the conjugate type and the right focal type. More specifically, we shall be interested in solutions of (1) satisfying

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1=t_2<t_3<b, \tag{8}$$

$$y(t_1)=y_1,\quad y(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1<t_2=t_3<b, \tag{9}$$

$$y(t_1)=y_1,\quad y(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1<t_2<t_3<b, \tag{10}$$

and in solutions of (1) satisfying

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y'(y_3)=y_3,\qquad a<t_1=t_2<t_3<b, \tag{11}$$

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y''(t_3)=y_3,\qquad a<t_1<t_2=t_3<b, \tag{12}$$

$$y(t_1)=y_1,\quad y'(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1<t_2<t_3<b, \tag{13}$$

and finally, in solutions of (1) satisfying,

$$y(t_1)=y_1,\quad y''(t_2)=y_2,\quad y'(t_3)=y_3,\qquad a<t_1=t_2<t_3<b. \tag{14}$$

In order that this presentation be self-contained, we will briefly state in § 2 the basics involved in applying the Pontryagin Maximum Principle to each of the families of boundary value problems for (1). That discussion is taken from [13]. In §§ 3 and 4, we then apply the Pontryagin Maximum Principle and determine maximal length subintervals on which solutions for each family of "in between" boundary value problems for (1) are unique, when solutions exist. Then, for each case, we will either reference theorems or prove uniqueness implies existence theorems for these boundary value problems. Finally, in § 5, we consider the case where $k_i=1$, $i=1, 2, 3$, and we compute the corresponding best subinterval lengths for subintervals of (a, b), on which there exist unique solutions, for the cases of the conjugate, the right focal, and the "in between" boundary value problems for (1). In some of the cases, we compare the best interval lengths with those obtained by standard applications of the Contraction Mapping Principle; in a couple of cases, we compare the best interval lengths with those obtained by the methods developed in [1].

2. The Pontryagin Maximum Principle. In this section, we will give a brief presentation on the manner in which the Pontryagin Maximum Principle can be applied. Our discussion is taken from Jackson [13]. We formulate this application in terms of nth order differential equations.

Let $k_i>0$, $1\leqq i\leqq n$, be fixed and let

$$U=\{(u_1(t),\cdots,u_n(t))\,|\,u_i(t) \text{ is Lebesgue measurable on } (a,b), \text{and } |u_i(t)|\leqq k_i,\ 1\leqq i\leqq n, \text{ on } (a,b)\}.$$

Let I, J be nonempty subsets of $\{1,\cdots,n\}$ such that $\operatorname{card}(I)+\operatorname{card}(J)=n$, and let I^c, J^c denote the respective complements of I, J in $\{1,\cdots,n\}$.

For fixed sets I, J, consider the boundary value problems

$$x^{(n)}=\sum_{i=1}^{n} u_i(t)x^{(i-1)}, \tag{15}$$

$$x^{(i-1)}(t_1)=0,\qquad i\in I, \tag{16}$$

$$x^{(i-1)}(t_2)=0,\qquad i\in J, \tag{17}$$

where $a < t_1 < t_2 < b$ and $u = (u_1(t), \cdots, u_n(t)) \in U$. Since the $u_i(t)$ are bounded measurable functions, we define here what is meant by a *solution* of (15).

DEFINITION. $x(t)$ is a *solution* of (15), for a control vector $u \in U$, if $x(t)$ is of class $C^{(n-1)}(a, b)$, $x^{(n-1)}(t)$ is absolutely continuous on (a, b), and $x(t)$ satisfies (15) for almost all $t \in (a, b)$.

We assume similar definitions for *solutions* of other differential equations which appear in this paper and which involve the control vectors $u \in U$.

Now, if (15)-(17) has a nontrivial solution for some $t_1 < t_2$ and some $u \in U$, then it follows that there is a boundary value problem in the collection which has a nontrivial optimal solution, (see Lee and Markus [16, Thm. 1, p. 30 or Thm. 4, p. 259]); that is, there exists at least one nontrivial $u^* \in U$ and $t_1 \leqq c < d \leqq t_2$ such that

$$x^{(n)} = \sum_{i=1}^{n} u_i^*(t) x^{(i-1)},$$

$$x^{(i-1)}(c) = 0, \qquad i \in I,$$

$$x^{(i-1)}(d) = 0, \qquad i \in J,$$

has a nontrivial solution $x(t)$ and $d - c$ is a minimum over all such solutions. For this time optimal solution, if $z(t) = (x(t), x'(t), \cdots, x^{(n-1)}(t))^T$, then $z(t)$ is a solution of the corresponding first order system

$$z' = A[u^*(t)]z.$$

By the Pontryagin Maximum Principle [16, Cor. 1, p. 314], the adjoint system

$$\psi' = -A^T[u^*(t)]\psi$$

has a nontrivial solution $\psi(t) = (\psi_1(t), \cdots, \psi_n(t))^T$ such that

(i) $\sum_{i=1}^{n} x^{(i)}(t)\psi_i(t) = \langle z'(t), \psi(t)\rangle = \operatorname*{Max}_{u \in U} \{\langle A[u(t)]z(t), \psi(t)\rangle\}$,
for almost all $t \in [c, d]$, ($\langle \cdot, \cdot \rangle$ denotes inner product);

(ii) $\langle z'(t), \psi(t)\rangle$ is a nonnegative constant for almost all $t \in [c, d]$; and

(iii) $\psi_i(c) = 0, \qquad i \in I^c,$

$\psi_i(d) = 0, \qquad i \in J^c.$

As shown in [13], the maximum condition in (i) can be rewritten as

$$\psi_n(t) \sum_{i=1}^{n} u_i^*(t) x^{(i-1)}(t) = \operatorname*{Max}_{u \in U} \left\{ \psi_n(t) \sum_{i=1}^{n} u_i(t) x^{(i-1)}(t) \right\}, \tag{18}$$

for almost all $t \in [c, d]$, whence it follows that if $\psi_n(t)$ has no zeros on (c, d) and if $x(t) > 0$ on (c, d), then (18) can be used to determine an optimal control $u^*(t)$ for almost all $t \in [c, d]$, (conceivably some derivative of $x(t)$ might be zero at some points).

In particular, if $x(t) > 0$ and $\psi_n(t) < 0$ on (c, d), then the time optimal solution $x(t)$ is a solution of

$$x^{(n)} = -\left[k_1 x + \sum_{i=1}^{n} k_i |x^{(i-1)}| \right], \tag{19}$$

on $[c, d]$. On the other hand, if $x(t) > 0$ and $\psi_n(t) > 0$ on (c, d), then the time optimal solution $x(t)$ is a solution of

$$x^{(n)} = k_1 x + \sum_{i=1}^{n} k_i |x^{(i-1)}|, \tag{20}$$

on $[c, d]$.

In subsequent sections, converse statements concerning the adjoint equation play a major role. If $u \in U$ is such that the boundary value prolem (15)-(17) has a nontrivial solution, then

$$\psi' = -A^T[u(t)]\psi, \tag{21}$$

$$\psi_i(t_1) = 0, \qquad i \in I^c, \tag{22}$$

$$\psi_i(t_2) = 0, \qquad i \in J^c, \tag{23}$$

also has a nontrivial solution; thus, the converse is also true. As a consequence of that, the Pontryagin Maximum Principle associates with a time optimal solution of (15)-(17), a time optimal solution of (21)-(23), and conversely.

In applying our statements in this section to nonlinear equations, consider the nth order differential equation

$$y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}), \tag{24}$$

where f is continuous and satisfies the Lipschitz condition

$$|f(t, y_1, \cdots, y_n) - f(t, z_1, \cdots, z_n)| \leqq \sum_{i=1}^{n} k_i |y_i - z_i| \tag{25}$$

on $(a, b) \times R^n$.

If $y(t)$ and $z(t)$ are distinct solutions of (24) on (a, b), and if $u_i(t)$, $1 \leqq i \leqq n$, is defined by

$$u_i(t) = \begin{cases} \dfrac{f(t, z(t), \cdots, z^{(i-2)}(t), y^{(i-1)}(t), \cdots, y^{(n-1)}(t))}{y^{(i-1)}(t) - z^{(i-1)}(t)} \\ \quad - \dfrac{f(t, z(t), \cdots, z^{(i-1)}(t), y^{(i)}(t), \cdots, y^{(n-1)}(t))}{y^{(i-1)}(t) - z^{(i-1)}(t)} & \text{for } y^{(i-1)}(t) \neq z^{(i-1)}(t), \\ -k_i \quad \text{for } y^{(i-1)}(t) = z^{(i-1)}(t), \end{cases}$$

then $u_i(t)$ is measurable on (a, b) and $|u_i(t)| \leqq k_i$. Let $x(t) = y(t) - z(t)$. Now if $t \in (a, b)$ is such that $y^{(i-1)}(t) = z^{(i-1)}(t)$, $i \in H \subset \{1, \cdots, n\}$, and if $K = H^C$, then invoking the two parts defining $u_i(t)$,

$$x^{(n)}(t) = \sum_{i \in K} u_i(t) x^{(i-1)}(t) + \sum_{i \in H} -k_i(0) = \sum_{i=1}^{n} u_i(t) x^{(i-1)}(t),$$

and consequently, $x(t) = y(t) - z(t)$ is a solution of the linear equation (15).

3. Intervals of existence, I. In this section we will be concerned with determining best possible interval lengths in terms of k_i, $i = 1, 2, 3$, of subintervals of (a, b) on which boundary value problems for the third order equation (1) satisfying (8), (9), or (10) all have unique solutions. First, we will use the results of the preceding section to determine optimal length subintervals on which solutions of each of the above boundary value problems are unique, when solutions exist. We then appeal to uniqueness implies existence theorems for this family of problems.

As we noted in § 2, if a time optimal solution of (15)-(17) and a corresponding solution of the adjoint equation satisfy certain sign conditions, then the time optimal solution is a solution of either (19) or (20). For this section and those that follow (since $n=3$), equations (15), (19), and (20) take the respective forms

$$x'''=u_1(t)x+u_2(t)x'+u_3(t)x'', \tag{26}$$

$$x'''=-k_1x-k_2|x'|-k_3|x''|, \tag{27}$$

$$x'''=k_1x+k_2|x'|+k_3|x''|. \tag{28}$$

THEOREM 3.1. *If there is a vector $u\in U$ such that the corresponding equation* (26) *has a nontrivial solution satisfying*

$$y(t_1)=y'(t_1)=y'(t_2)=0, \qquad a<t_1<t_2<b$$

and if $x(t)$ is a time optimal solution with

$$x(c)=x'(c)=x'(d)=0$$

and with $d-c$ a minimum, then $x(t)$ is a solution of (27) *on* $[c, d]$.

Proof. By the time optimality, it follows that $x'(t)\neq 0$ on (c, d), and thus $x(t)\neq 0$ on $(c, d]$. Without loss of generality we may assume $x''(c)>0$, so that $x(t)>0$ on $(c, d]$.

Now, if $\psi(t)$ is a solution of the adjoint system associated with $x(t)$ by the Pontryagin Maximum Principle, then

$$\psi_3(c)=\psi_1(d)=\psi_3(d)=0,$$

and by its own time optimality, $\psi_3(t)\neq 0$ on (c, d). Hence, $x(t)$ is a solution of (27) or (28) on $[c, d]$. From the nature of these two equations, $x''(t)$ is strictly monotone on $[c, d]$, and since $x'(c)=x'(d)=0$, while $x''(c)>0$, it follows that $x''(d)<0$.

Moreover, from the Maximum Principle, there exists $K\geqq 0$ such that

$$K=\sum_{i=1}^{3} x^{(i)}(t)\psi_i(t)=x''(d)\psi_2(d)=x''(c)\psi_2(c)$$

on $[c, d]$. We conclude $\psi_2(d)<0$, and from the adjoint system

$$\psi_3'(d)=-\psi_2(d)-u_3^*(d)\psi_3(d)=-\psi_2(d)>0.$$

Consequently, $\psi_3(t)<0$ on (c, d), and $x(t)$ is a solution of (27).

The following two theorems concerning uniqueness of solutions and uniqueness implies existence of solutions for (1), (8), (9) and (10) are proven in [8], [11].

THEOREM 3.2. *The boundary value problem* (1), (10) *has at most one solution on* (a, b), *if and only if each of the boundary value problems* (1), (8), *and* (1), (9) *has at most one solution on* (a, b).

THEOREM 3.3. *If* (1), (10) *has at most one solution on* (a, b), *then boundary value problems for* (1) *satisfying* (8), (9), *or* (10) *all have unique solutions on* (a, b).

We can now state the result concerning maximal length subintervals of (a, b) on which the boundary value problems of this section have unique solutions.

THEOREM 3.4. *Let $h>0$ be the smallest positive number such that there is a solution $x(t)$ of the boundary value problem for* (27) *satisfying*

$$x(0)=x'(0)=x'(h)=0,$$

with $x(t)>0$ on $(0, h]$, or $h=+\infty$ if no such solution exists. Then each of the boundary value problems for (1) *satisfying* (8), (9), *or* (10) *has a unique solution, provided $t_3-t_1<h$. Moreover, this result is best possible for the class of all differential equations satisfying the Lipschitz condition* (B).

Proof. We first note that since (27) is autonomous, in applying Theorem 3.1, rather than specifying boundary conditions at $a<c<d<b$, it suffices to consider conditions at 0 and h.

We show that solutions of (1) satisfying (8) or (9) are unique, when they exist, on any subinterval of (a, b) of length less than h. Assume that $y(t)$ and $z(t)$ are distinct solutions of (1) satisfying either (8) or (9) with $t_3-t_1<h$. Then $w(t)=y(t)-z(t)$ is a nontrivial solution of the linear equation (26) for suitable $u=(u_1(t), u_2(t), u_3(t))\in U$, and satisfies either

(i) $w(t_1)=w'(t_1)=w'(t_3)=0$, or

(ii) $w(t_1)=w(t_3)=w'(t_3)=0$.

Case (i) leads to a contradiction of Theorem 3.1. For case (ii), it is shown in Jackson [12, Cor. to Thm. 8] that there is a nontrivial solution $x(t)$ of (28) satisfying $x(0)=x'(0)=x(-k)=0$, for some $0<k<h$. If $v(t)=x(-t)$, then $v(t)$ is a nontrivial solution of (27) and satisfies $v(0)=v'(0)=v(k)=0$. By Rolle's theorem, there exists $0<\eta<k$ such that

$$v(0)=v'(0)=v'(\eta)=0,$$

which again contradicts Theorem 3.1.

Therefore, solutions of boundary value problems for (1) satisfying (8) or (9) are unique, when they exist. That unique solutions of (1) satisfying (8), (9), or (10) exist, provided $t_3-t_1<h$, is immediate from Theorems 3.2 and 3.3. Since (27) is a Lipschitz equation satisfying (B), it also follows that this result is best possible.

4. Intervals of existence, II. We now consider interval lengths of subintervals of (a, b) on which there exist unique solutions of boundary value problems for (1) satisfying (11), (12), or (13) (note that (11) is the same as (8)). By further restricting our subinterval length for the above three problems, we will also consider existence and uniqueness for problems studied by Aftabizadeh and Wiener [1] satisfying (1), (14).

THEOREM 4.1. *Assume that for all vectors $u\in U$, the corresponding linear equation* (26) *has only the trivial solution satisfying*

$$x(t_1)=x'(t_1)=x'(t_2)=0, \qquad a<t_1<t_2<b.$$

If there is a control vector $u\in U$ such that the corresponding equation (26) *has a nontrivial solution satisfying*

$$y(t_1)=y'(t_2)=y''(t_2)=0, \qquad a<t_1<t_2<b$$

and if $x(t)$ is a time optimal solution with

$$x(c)=x'(d)=x''(d)=0$$

and with $d-c$ a minimum, then $x(t)$ is a solution of (28) *on* $[c, d]$.

Proof. By the time optimality of $x(t)$ and uniqueness of solutions of initial value problems, $x(t)\neq 0$ on $(c, d]$. Also, we may assume that $x'(c)>0$ so that $x(t)>0$ on $(c, d]$.

If $\psi(t)$ is a solution of the adjoint system associated with $x(t)$ by the Maximum Principle, then

$$\psi_2(c)=\psi_3(c)=\psi_1(d)=0$$

and $\psi_1(t)\neq 0$ on $[c, d)$.

In order to apply the results of § 2, we need to show that $\psi_3(t)\neq 0$ on (c, d). To this end, let $y(t)=(y_1(t), y_2(t), y_3(t))^T$, where $y_1(t)=\psi_3(t)$, $y_2(t)=\psi_2(t)$, and $y_3(t)=\psi_1(t)$, so that $y(t)$ is a solution of

$$y'=B[u^*(t)]y \tag{29}$$

where

$$B[u(t)] = \begin{bmatrix} -u_3(t) & -1 & 0 \\ -u_2(t) & 0 & -1 \\ -u_1(t) & 0 & 0 \end{bmatrix}$$

and $y_1(c) = y_2(c) = y_3(d) = 0$. Our argument will now be concerned with showing $y_1(t) = \psi_3(t) \neq 0$ on (c, d).

We remark that from the optimality of $x(t)$ and $\psi(t)$ and from the hypotheses, $\psi_3(d) = y_1(d) \neq 0$. For $j = 1, 2, 3$, let

$$y^j(t) = (y_1^j(t), y_2^j(t), y_3^j(t))^T$$

denote the solution of the initial value problem for (29) satisfying

$$y_i^j(d) = \delta_{ij}, \qquad i = 1, 2, 3.$$

It follows that

$$y(t) = C_1 y^1(t) + C_2 y^2(t),$$

for some C_1, C_2. Moreover, since $y_1(d) \neq 0$, we have $C_1 \neq 0$.

Now, it is also the case that $y_1^2(t) \neq 0$ on (a, d). To see this, asume there exists $\tau \in (a, d)$ so that $y_1^2(\tau) = 0$. Since $y^2(t)$ is a solution of (29), it follows that the adjoint system $\psi' = -A[u^*(t)]\psi$ has a solution $\eta(t) = (\eta_1(t), \eta_2(t), \eta_3(t))^T$, where $\eta_1 = y_3^2$, $\eta_2 = y_2^2$, $\eta_3 = y_1^2$, satisfying

$$\eta_3(\tau) = \eta_1(d) = \eta_3(d) = 0.$$

This, in turn, implies there exists an optimal such solution to the adjoint equation, for some $u^{**} \in U$, and then by the Pontryagin Maximum Principle, there is an optimal solution $v(t)$ of (26), for $u^{**} \in U$, such that

$$v(\tau_1) = v'(\tau_1) = v'(\tau_2) = 0,$$

where $\tau \leqq \tau_1 < \tau_2 \leqq d$; this is a contradiction to the hypotheses of the theorem. Thus, $y_1^2(t) \neq 0$ on (a, d).

Now, if $y_1(t_0) = \psi_3(t_0) = 0$, for some $t_0 \in (c, d)$, since

$$-W(y^2(t), y(t)) = [y_1^2(t)]^2 (y_1(t)/y_1^2(t))',$$

(where $W(\cdot, \cdot)$ denotes the Wronskian), and since $y_1(c) = 0$ and $y_1^2(c) \neq 0$, it follows from Rolle's theorem that $W(y^2(t_1), y(t_1)) = 0$, for some $c < t_1 < t_0$. But $W(y^2(t_1), y(t_1)) = C_1 W(y^1(t_1), y^2(t_1))$, and since $C_1 \neq 0$, there are constants r_1, r_2 such that the solution

$$w(t) \equiv r_1 y^1(t) + r_2 y^2(t)$$

of (29) satisfies $w_1(t_1) = w_2(t_1) = w_3(d) = 0$. This implies the adjoint system has a solution $\beta(t)$ such that $\beta_2(t_1) = \beta_3(t_1) = \beta_1(d) = 0$, where $c < t_1 < d$. This contradicts the optimality of $d - c$. Therefore $y_1(t) = \psi_3(t) \neq 0$ on (c, d).

From our assumption that $x(t) > 0$, it follows that $x(t)$ is a solution of (27) or (28) on $[c, d]$. It follows, in turn, from the constancy in sign of $x'''(t)$ on $[c, d]$ and from the boundary conditions $x(c) = x'(d) = x''(d) = 0$, that $x'''(t) > 0$ on $[c, d]$, so that $x(t)$ is a solution of (28) on $[c, d]$.

For our remaining existence considerations, we appeal to results much like those in the previous section. The next two theorems are proven in [8], [11].

THEOREM 4.2. *If there is at most one solution of the boundary value problem* (1), (13) *on* (a, b), *then each of the boundary value problems* (1), (11), *and* (1), (12) *has at most one solution on* (a, b).

THEOREM 4.3. *If* (1), (13) *has at most one solution on* (a, b), *then boundary value problems for* (1) *satisfying* (11), (12), *or* (13) *all have unique solutions on* (a, b).

Although Theorem 4.2 is not the analogue of Theorem 3.2, the converse of Theorem 4.2 is true for linear equations.

LEMMA. *Let Ly be a linear third order differential operator on an interval I. If the only solution of $Ly = 0$ satisfying either*

$$y(t_1) = y'(t_2) = y'(t_3) = 0, \qquad t_1 = t_2 < t_3,$$

or

$$y(t_1) = y'(t_2) = y''(t_3) = 0, \qquad t_1 < t_2 = t_3$$

is the trivial solution, then the only solution of $Ly = 0$ satisfying

$$y(t_1) = y'(t_2) = y'(t_3) = 0, \qquad t_1 < t_2 < t_3$$

is $y(t) = 0$.

Proof. Assume to the contrary that there is a nontrivial solution $y(t)$ of $Ly = 0$ satisfying $y(t_1) = y'(t_2) = y'(t_3) = 0$, for some $t_1 < t_2 < t_3$ belonging to I. By the hypotheses, $y'(t_1), y''(t_2), y''(t_3) \neq 0$.

Now, by the hypotheses, there exists a unique solution $z(t)$ of $Ly = 0$ satisfying $z(t_1) = z'(t_1) = 0$ and $z'(t_2) = +1$. Moreover, $z'(t) \neq 0$, for all $t \in [t_2, t_3]$. It follows that for some nonzero $\alpha \in \mathbf{R}$ and $\tau \in (t_2, t_3)$, the solution $w(t) \equiv y(t) + \alpha z(t)$ of $Ly = 0$ satisfies $w(t_1) = w'(\tau) = w''(\tau) = 0$; see [17]. This contradicts the hypotheses of the lemma, and we conclude no such nontrivial solution $y(t)$ exists.

For the case of boundary value problems (1), (11), (12), and (13), we can now determine our optimal length subintervals on which all have unique solutions.

THEOREM 4.4. *Let $k = \min\{h, r\}$, where h is the number obtained in Theorem* 3.4, *and $r > 0$ is the smallest positive number such that there is a solution $x(t)$ of the boundary value problem for* (28) *satisfying*

$$x(0) = x'(r) = x''(r) = 0,$$

with $x(t) > 0$ on $(0, r]$, or $r = +\infty$ if no such solution exists. Then each of the boundary value problems for (1) *satisfying* (11), (12), *or* (13) *has a unique solution, provided $t_3 - t_1 < k$. Again, this result is best possible.*

Proof. We will show that solutions for (1), (13) are unique, when they exist. For the purpose of contradiction, assume there are distinct solutions $y(t)$ and $z(t)$ of (1), (13), where $t_3 - t_1 < k$. Then $w(t) = y(t) - z(t)$ is a nontrivial solution of linear equation (26), for a suitable $u^0 \in U$, and satisfies $w(t_1) = w'(t_2) = w'(t_3) = 0$. The above lemma coupled with Theorem 4.2 implies (26), for $u^0 \in U$, has a solution $\beta(t)$ satisfying either

(i) $\beta(\tau_1) = \beta'(\tau_1) = \beta'(\tau_2) = 0$, or

(ii) $\beta(\tau_1) = \beta'(\tau_2) = \beta''(\tau_2) = 0$

where $t_1 \leqq \tau_1 < \tau_2 \leqq t_3$; consequently, there is a nontrivial optimal solution $x(t)$ satisfying either (27) or (28) and boundary conditions of the type respectively given in (i) or (ii). Since $t_3 - t_1 < k$, this is a contradiction.

Therefore, solutions of (1), (13) are unique, when they exist, and the conclusion follows from Theorem 4.3.

We conclude this section by further restricting our above subinterval length k and obtaining, in addition, subintervals on which there unique solutions of (1), (14). Such a problem was considered in [1], by first transforming the problem into a boundary value problem for a second order integro-differential equation. Then successive approximations were used in obtaining length bounds on intervals for existence of unique solutions; these bounds were somewhat better than those obtained using standard Contraction Mapping arguments. However, we will determine best length subintervals for the existence of unique solutions of (1), (14), when the problems (1), (11), (12), and (13) all have unique solutions.

THEOREM 4.5. *Assume that for all vectors $u \in U$, the corresponding linear equation* (26) *has only the trivial solution satisfying*

$$x(t_1)=x'(t_1)=x'(t_2)=0, \qquad a<t_1<t_2<b.$$

If there is a control vector $u \in U$ such that the corresponding equation (26) *has a nontrivial solution satisfying*

$$y(t_1)=y''(t_1)=y'(t_2)=0, \qquad a<t_1<t_2<b$$

and if $x(t)$ is a time optimal solution with

$$x(c)=x''(c)=x'(d)=0$$

and with $d-c$ a minimum, then $x(t)$ is a solution of (27) *on* $[c, d]$.

Proof. The proof proceeds much like those before. By optimality $x'(t)\neq 0$ on $[c, d)$, and thus by Rolle's theorem, $x(t)\neq 0$ on $(c, d]$. We may assume $x(t)>0$ on $(c, d]$. Now if $\psi(t)$ is the corresponding optimal solution of the adjoint system, then

$$\psi_2(c)=\psi_1(d)=\psi_3(d)=0.$$

Now, if $\psi_3(t_0)=0$, for some $t_0\in[c, d)$, since $\psi_1(d)=\psi_3(d)=0$, there is an optimal solution $z(t)$ of (26) satisfying

$$z(\tau_1)=z'(\tau_1)=z'(\tau_2)=0,$$

for some $t_0\leqq \tau_1<\tau_2\leqq d$, a contradiction. Therefore, $\psi_3(t)\neq 0$ on $[c, d)$, and $x(t)$ is a solution of (27) or (28).

Since $x'''(t)$ is of constant sign on [c, d], from the conditions $x(c)=x''(c)=x'(d)=0$, we conclude $x'''(t)<0$ on $[c, d]$, and $x(t)$ is a solution of (27) on $[c, d]$.

For uniqueness implies existence we have the following.

LEMMA. *Assume that solutions of* (1), (13), *and* (1), (14) *are unique, when they exist on* (a, b). *Then boundary value problems for* (1) *satisfying* (11), (12), (13), *or* (14) *all have unique solutions on* (a, b).

Proof. In light of Theorem 4.3, we need only prove the statement for (1), (14). For this, let $a<t_1<t_2<b$ and $y_1, y_2, y_3\in \mathbf{R}$ be given, and let $y(t)$ be the solution of the initial value problem for (1) satisfying

$$y(t_1)=y_1, \quad y'(t_1)=0, \quad y''(t_1)=y_2.$$

Now define

$$S\equiv\{z'(t_2)\,|\,z(t) \text{ is a solution of (1) and } z(t_1)=y(t_1),\ z''(t_1)=y''(t_1)\}.$$

Using continuous dependence and the fact that solutions of (1), (11) and (1), (12) exist, it can be shown by fairly standard arguments that $S=\mathbf{R}$; see [10]. Thus, by choosing $y_3\in S$, the corresponding solution $z(t)$ of (1) satisfies (14).

THEOREM 4.6. *Let $l=\min\{k, s\}$, where k is the number obtained in Theorem* 4.4, *and $s>0$ is the smallest positive number such that there is a solution $x(t)$ of the boundary problem for* (27) *satisfying*

$$x(0)=x''(0)=x'(s)=0,$$

with $x(t)>0$ *on* $(0, s]$, *or* $s=+\infty$ *if no such solution exists. Then each of the boundary value problems for* (1) *satisfying* (11), (12), (13), *or* (14) *has a unique solution, provided* $t_3-t_1<l$. *Again, this is best possible.*

Proof. The proof is immediate from the last lemma and Theorem 4.4.

5. Best interval lengths for $k_1=k_2=k_3=1$. In this section, we employ the results of the previous sections in computing the best possible interval lengths of subintervals of (a, b) on which the boundary value problems for (1) have unique solutions for the case where the Lipschitz coefficients satisfy $k_1=k_2=k_3=1$. For some of the problems, we compare our results with interval length bounds obtained by standard Contraction Mapping methods. In a couple of cases, we will also compare our results with those obtained in [1].

For each of the following cases, we let $k_1=k_3=k_3=1$.

(i) For our first case, we are concerned with solutions of conjugate boundary value problems for (1); that is, conditions (2), (3), and (4). As a corollary to Theorem 1.1, Jackson [12] showed that the boundary value problems for (1) satisfying (2), (3), or (4) all have unique solutions on any open subinterval of length less than h, where $x(t)$ is the solution of (27) satisfying the initial conditions

$$x(0)=x'(0)=0, \qquad x''(0)=1$$

and $h>0$ is the first positive number such that $x(h)=0$. In this case, Jackson obtained the best possible result to be $h=2.7353$. On the other hand, from Contraction Mapping bounds, interval lengths of only 1.1284 result.

(ii) In this case, we are concerned with right focal boundary value problems for (1) satisfying conditions (5), (6), and (7). From Theorem 1.3, boundary value problems for (1) satisfying (5), (6), or (7) all have unique solutions on open subintervals of (a, b) of length less than $k=\min\{r_1, r_2\}$, where $x(t)$ is the solution of (27) satisfying the initial conditions

$$x(0)=x'(0)=0, \qquad x''(0)=1$$

and $r_1>0$ is the first positive number such that $x''(r_1)=0$, and $y(t)$ is the solution of (27) satisfying

$$y(0)=1, \qquad y'(0)=y''(0)=0$$

and $r_2>0$ is the first positive number such that $y(r_2)=0$. For this case, we obtain the best subinterval length as $k=r_1=1.03842$. For comparison, using Contraction Mapping, interval lengths are bounded by .672496, whereas Aftabizadeh and Wiener's techniques yielded lengths of .896861.

(iii) For this case, we consider the problems of § 3. From Theorem 3.4, boundary value problems for (1) satisfying (8), (9), or (10) have unique solutions on open subintervals of (a, b) of length less than η, where $x(t)$ is the solution of (27) and satisfies

$$x(0)=x'(0)=0, \qquad x''(0)=1,$$

and $\eta>0$ is the first positive number such that $x'(\eta)=0$. We find here that best subinterval lengths are $\eta=1.923239$. Aftabizadeh and Wiener's techniques yield interval lengths bounded by 1.59542.

(iv) For this case, we are concerned with subinterval lengths for the problems of § 4. For the case of problems (1)-(11), (12), or (13), all have unique solutions on subintervals of (a, b) of length less than $l=\min\{l_1, l_2\}$, where $x(t)$ is the solution (27) satisfying

$$x(0)=x'(0)=0, \qquad x''(0)=1$$

and $l_1 > 0$ is the first number such that $x(l_1) = 0$, and $y(t)$ is the solution of (27) satisfying

$$y(0) = 1, \qquad y'(0) = y''(0) = 0$$

and $l_2 > 0$ is the first positive number such that $y(l_2) = 0$ (to see this is the appropriate initial value problem for l_2, replace the solution $x(t)$ in Theorem 4.4 by $x(-t+r)$). We note that $l_1 = \eta$ from (iii) and $l_2 = r_2$ from (ii), and it follows that the best subinterval length is $l = l_2 = 1.52374$.

As we proved in Theorem 4.6, by restricting the subinterval lengths further, (1), (14) also has a unique solution. In fact, for this special case of Lipschitz coefficients, we have unique solutions of (1), (14) on intervals of length less than m, where $x(t)$ is the solution of (27) satisfying

$$x(0) = x''(0) = 0, \qquad x'(0) = 1$$

and $m > 0$ is the first number such that $x'(m) = 0$. For the best possible interval length, $m = 1.10647$.

Remark. For the calculations discussed in (ii)-(iv), Runge-Kutta methods were used. However, in some cases, such as (ii), it is elementary to explicitly solve the specified initial value problem; then applying Newton's method to the appropriate derivative of the solution, the optimal interval length can be determined.

REFERENCES

[1] R. Aftabizadeh and J. Wiener, *On the solutions of third order nonlinear boundary value problems*, in Proc. 6th International Conference on Nonlinear Analysis, V. Lakshmikantham, ed., North-Holland, Amsterdam-New York, 1985, pp. 1-6.

[2] R. Agarwal, *Boundary value problems for higher order integro-differential equations*, Nonlinear Anal., 7 (1983), pp. 259-270.

[3] R. Agarwal and P. Krishnamoorthy, *On the uniqueness of solutions of nonlinear boundary value problems*, J. Math. Phys. Sci., 10 (1976), pp. 17-31.

[4] ———, *Boundary value problems for nth order ordinary differential equations*, Bull. Inst. Math. Acad. Sinica, 7 (1979), pp. 211-230.

[5] P. Bailey, L. Shampine and P. Waltman, *Nonlinear Two Point Boundary Value Problems*, Academic Press, New York, 1968.

[6] D. Barr and T. Sherman, *Existence and uniqueness of solutions of three-point boundary value problems*, J. Differential Equations, 13 (1973), pp. 197-212.

[7] K. Das and B. Lalli, *Boundary value problem for $y''' = f(x, y, y', y'')$*, J. Math. Anal. Appl., 81 (1981), pp. 300-307.

[8] D. Goecke and J. Henderson, *Uniqueness of solutions of right focal problems for third order differential equations*, Nonlinear Anal., 8 (1984), pp. 253-259.

[9] P. Hartman, *On n-parameter families and interpolation problems for nonlinear ordinary differential equations*, Trans. Amer. Math. Soc., 154 (1971), pp. 201-226.

[10] J. Henderson, *Existence of solutions of right focal point boundary value problems for ordinary differential equations*, Nonlinear Anal., 5 (1981), pp. 989-1002.

[11] ———, *Right $(m_1; \cdots; m_l)$ focal boundary value problems for third order differential equations*, J. Math. Phys. Sci., 18 (1984), pp. 405-413.

[12] L. Jackson, *Existence and uniqueness of solutions of boundary value problems for Lipschitz equations*, J. Differential Equations, 32 (1979), pp. 76-90.

[13] ———, *Boundary value problems for Lipschitz equations*, in Differential Equations, S. Ahmad, M. Keener and A. C. Lazer, eds., Academic Press, New York, 1980, pp. 31-50.

[14] G. Klaasen, *Existence theorems for boundary value problems for nth order ordinary differential equations*, Rocky Mountain J. Math., 3 (1973), pp. 457-472.

[15] P. Krishnamoorthy, *Three-point nonlinear boundary value problems*, Proc. 16th Anniversary Symposium, Inst. Math. Sci., Madras, 1978.

[16] E. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[17] W. Leighton and Z. Nehari, *On the oscillation of solutions of self-adjoint linear differential equations of the fourth order*, Trans. Amer. Math. Soc., 89 (1958), pp. 325-37.

[18] Yu. Melentsova and G. Mil'shtein, *An optimal estimate of the interval on which a multipoint boundary value problem possesses a solution*, Differencial'nye Uravrnenija USSR, 10 (1974), pp. 1630–1641. (In Russian.) English translation in Differential Equations, 10 (1974), pp. 1257–1265.

[19] Yu. Melentsova, *A best possible estimate of the nonoscillation interval for a linear differential equation with coefficients bounded in L_r*, Differencial'nye Uravrnenija USSR, 13 (1977), pp. 1776–1786. (In Russian.) English translation in Differential Equations, 13 (1977), pp. 1236–1244.

[20] V. Moorti and J. Garner, *Existence and uniqueness theorems for three-point boundary value problems for third order differential equations*, J. Math. Anal. Appl., 70 (1979), pp. 370–385.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

003

ON THE RATE OF CONVERGENCE OF VISCOSITY SOLUTIONS FOR BOUNDARY VALUE PROBLEMS*

JENS LORENZ† AND RICHARD SANDERS‡

Abstract. A class of singularly perturbed boundary value problems is considered for viscosity tending to zero. From compactness arguments it is known that the solutions converge to a limit function characterized by an entropy inequality. We formulate an approximate entropy inequality (AEI) and use it to obtain the order of convergence. The AEI is also used to obtain the order of convergence for monotone difference schemes.

Key words. singular perturbations, entropy inequality, bounded variation functions

AMS(MOS) subject classifications. Primary 34E99, 65L10

1. Introduction. In this paper we establish a minimal rate of convergence theorem for solutions to the singularly perturbed boundary value problem

$$\begin{aligned} &-\varepsilon \frac{d^2}{dx^2} u_\varepsilon + \frac{d}{dx} f(x, u_\varepsilon) + b(x, u_\varepsilon) = 0, \\ &u_\varepsilon(0) = \gamma_0, \qquad u_\varepsilon(0) = \gamma_1, \end{aligned} \tag{1.1}$$

as the parameter $\varepsilon > 0$ tends to zero. Throughout we impose no special conditions on $f(x, u)$ or $b(x, u)$ other than that they are smooth and satisfy

$$\frac{\partial}{\partial u} b(x, u) - \left| \frac{\partial^2}{\partial u \partial x} f(x, u) \right| \geqq \mu > 0 \tag{1.2}$$

for all $(x, u) \in [0, 1] \times I$ where I is an a priori interval determined from the maximum principle.[1]

It is well known that for positive ε, condition (1.2) implies that the boundary value problem (1.1) has a unique smooth solution; see [8], [9] for results in this direction.

As ε tends to zero, solutions to (1.1) need not converge to a continuous function. Therefore, it is natural to seek a rate of convergence result in an integral sense. Below we show that there exists a function $\bar{u} \in BV$ such that for sufficiently small ε

$$\int_0^1 |u_\varepsilon - \bar{u}| \, dx \leqq \frac{C_\gamma}{\mu} \sqrt{\varepsilon}, \tag{1.3}$$

where the constant C_γ depends on the boundary data γ_0 and γ_1. In general the rate above is not valid *unless* condition (1.2) is imposed. That is to say there are examples

* Received by the editors January 13, 1986; accepted for publication May 8, 1986.

† Applied Mathematics 217-50, California Institute of Technology, Pasadena, California 91125. The research of this author was supported by National Science Foundation grant DMS 83-12264 and Department of Energy grant DE/AT/03/76ER/72012.

‡ Department of Mathematics, University of Houston, Houston, Texas 77004. The research of this author was supported by National Science Foundation grant DMS 8505422.

[1] The interested reader can greatly relax the smoothness assumption we make on the coefficients of (1.1). We consider these details uninteresting however, and we take $f(x, u)$ and $b(x, u)$ as smooth as the specified number of derivatives indicate.

of boundary value problems of the form (1.1) that violate (1.2) (they must satisfy an estimate like (1.2) with $\mu = 0$) and satisfy

$$\int_0^1 |u_\varepsilon - \bar{u}|\, dx \geqq \text{const. } \varepsilon^q$$

for any $\frac{1}{2} \geqq q > 0$. Moreover, given that condition (1.2) *is* satisfied, the L^1 rate result (1.3) cannot be improved unless further conditions are imposed. That is, there are examples that satisfy condition (1.2) and satisfy

$$\int_0^1 |u_\varepsilon - \bar{u}|\, dx \geqq \text{const.} \sqrt{\varepsilon}.$$

Some examples that demonstrate the sharpness of our rate result are presented at the end of this section.

The main contribution of this paper is to extend the techniques developed in [6], [13] to include problems with Dirichlet boundary conditions. The notion of an "approximate entropy inequality (AEI)," first introduced in the study of single conservation laws, is suitably modified to include boundary value problems of the type studied here. Specifically, what we show is that if a parameterized family of functions, say $\{v_h\}_{h>0}$, satisfies the uniform estimate

$$\text{var}\,(v_h) \leqq \text{const.},$$

together with an h-dependent AEI, then v_h satisfies (1.3) with h taking the place of ε. We have intentionally been vague about the precise definition of the family $\{v_h\}_{h>0}$ since it is shown below that besides representing the family of solutions to (1.1) it can also represent a family of certain numerical approximations. In the application to numerical approximations v_h denotes an interpolation of grid values generated by a finite difference scheme, and h denotes a measure of grid refinement.

In § 2 the characterizing "entropy inequality" for the limit of solutions to (1.1) is stated; see [2], [3], [5], [14] for a thorough development of these ideas. The "approximate entropy inequality" is also defined in this section, and solutions of (1.1) are shown to satisfy it. The abstract rate of convergence theorem implied by the AEI is also stated in § 2. In § 3 the abstract rate of convergence theorem, stated in § 2, is proved. Finally in § 4 the rate of convergence theorem is applied to numerical approximations generated by certain types of finite difference schemes.

We should mention that most of the results of this paper can be routinely extended to quasilinear Dirichlet problems in many space dimensions. This will be the topic of future work; see [12], where somewhat parallel techniques are applied to nonlinear problems with boundary conditions of Neumann type.

We conclude this section by constructing some nontrivial examples of the type mentioned above. First we show that if (1.2) is violated then an arbitrarily slow L^1 rate of convergence is possible. To this end, consider the boundary value problem

$$\begin{gathered} -\varepsilon \frac{d^2}{dx^2} u_\varepsilon + \frac{d}{dx}(\gamma - u_\varepsilon)^{2p} = 0, \\ u_\varepsilon(0) = 0, \qquad u_\varepsilon(1) = \gamma > 0, \end{gathered} \tag{1.4}$$

where we take $p > 1$. Clearly (1.4) violates (1.2). By [8, Thm. 4] the solutions of (1.4) tend uniformly to γ on any interval $[\delta, 1]$, $1 > \delta > 0$, as ε tends to zero. Therefore, we wish to examine

$$\int_0^1 |u_\varepsilon - \gamma|\, dx. \tag{1.5}$$

With this end in mind we first consider

$$-\varepsilon\frac{d^2}{dx^2}v_\varepsilon+\frac{d}{dx}(\gamma-v_\varepsilon)^{2p}=0, \tag{1.6}$$
$$v_\varepsilon(0)=0, \qquad v_\varepsilon(\infty)=\gamma,$$

which can be integrated exactly, giving

$$v_\varepsilon(x)=\Phi^{-1}(x/\varepsilon), \tag{1.7}$$

where

$$\Phi(w)=\int_0^w (\gamma-s)^{-2p}\,ds.$$

Since $v_\varepsilon(x)+\gamma-v_\varepsilon(1)$ is an upper solution for (1.4), it follows that this function is $\geqq u_\varepsilon(x)$; thus

$$\gamma-u_\varepsilon(x)\geqq v_\varepsilon(1)-v_\varepsilon(x)\geqq 0,$$

which implies

$$\int_0^1 |u_\varepsilon-\gamma|\,dx\geqq\int_0^1 (v_\varepsilon(1)-v_\varepsilon(x))\,dx. \tag{1.8}$$

Interchanging the order of integration and using (1.7), we have that the right-hand side of (1.8) is given by

$$\varepsilon\int_0^{v_\varepsilon(1)}\Phi(w)\,dw.$$

Finally, a simple calculation will reveal that

$$\varepsilon\int_0^{v_\varepsilon(1)}\Phi(w)\,dw=\frac{1}{2p-2}\left(\frac{\varepsilon}{2p-1}\right)^{1/2p-1}+O(\varepsilon),$$

which therefore shows that

$$\int_0^1 |u_\varepsilon-\gamma|\,dx\geqq C_p(\varepsilon^{1/2p-1}),$$

as ε tends to zero.

To establish the fact, given only condition (1.2), our rate result (1.3) is the best possible, we note that the trivial example

$$-\varepsilon\frac{d^2}{dx^2}u_\varepsilon+u_\varepsilon=0,$$
$$u_\varepsilon(0)=0, \qquad u_\varepsilon(1)=1,$$

satisfies the $\sqrt{\varepsilon}L^1$ rate of convergence exactly. A less trivial example is given by

$$-\varepsilon\frac{d^2}{dx^2}u_\varepsilon+\frac{d}{dx}((1-x)u_\varepsilon)+2u_\varepsilon=0, \tag{1.9}$$
$$u_\varepsilon(0)=0, \qquad u_\varepsilon(1)=1.$$

To obtain the sharp $\sqrt{\varepsilon}$ rate for example (1.9) we apply the "shooting method." (Although this method is less general than the techniques we present in the following sections, it takes into account specific properties of $f(x, u)$ and often leads to sharper results than ours; see [4] for applications of differential inequalities.) From our results below, we expect that $\lim_{\varepsilon\downarrow 0} u_\varepsilon = 0$. The maximum principle together with integrating (1.9) gives us that

$$\int_0^1 |u_\varepsilon - 0|\, dx = \frac{\varepsilon}{2}\left[\frac{d}{dx}u_\varepsilon(1) - \frac{d}{dx}u_\varepsilon(0)\right]. \tag{1.10}$$

By applying the shooting method, it is easy to conclude that for all $\varepsilon > 0$ sufficiently small, we have

$$\frac{1}{2\sqrt{\varepsilon}} \leqq \frac{d}{dx}u_\varepsilon(1) \leqq \frac{3}{\sqrt{\varepsilon}},$$

and

$$0 \leqq \frac{d}{dx}u_\varepsilon(0) \leqq 1.$$

Inserting these inequalities into (1.10) we finally get

$$\frac{\sqrt{\varepsilon}}{4} - O(\varepsilon) \leqq \int_0^1 |u_\varepsilon - 0|\, dx \leqq \frac{3}{2}\sqrt{\varepsilon},$$

which establishes our L^1 rate for example (1.9).

2. The approximate entropy inequality. Throughout this paper the following notation is used:

(i) BV denotes the space of functions $u: [0, 1] \to \mathbb{R}$ of bounded variation.

(ii) $\operatorname{var}(u)$ is the total variation of $u \in BV$.

(iii) C_+^∞ denotes the space of functions $\phi: \mathbb{R} \to \mathbb{R}$ which are infinitely differentiable and nonnegative.

The sign-function is defined by

$$\operatorname{sgn}(u) = \begin{cases} -1 & \text{if } u < 0, \\ 0 & \text{if } u = 0, \\ 1 & \text{if } u > 0, \end{cases}$$

and

$$\operatorname{sgn}_\delta(u) = \begin{cases} u/|u| & \text{for } |u| \geqq \delta, \\ u/\delta & \text{for } |u| < \delta \end{cases}$$

denotes a Lipschitz continuous approximation to $\operatorname{sgn}(u)$ for $\delta > 0$. We furthermore use

$$\|u\|_\infty = \sup\{|u(x)|: 0 \leqq x \leqq 1\} \quad \text{for } u \in L^\infty[0, 1],$$

$$\|u\|_1 = \int_0^1 |u(x)|\, dx \quad \text{for } u \in L^1[0, 1].$$

For simplicity we assume that $f(x, u) \in C^2([0, 1] \times \mathbb{R})$ and $b(x, u) \in C^1([0, 1] \times \mathbb{R})$ although we recognize that weaker conditions are sufficient. The essential assumption is nevertheless condition (1.2), and it will be assumed throughout.

In the next proposition we state some known results concerning the second order boundary value problem (1.1); see [2], [9]. These facts are relevant in what follows.

PROPOSITION 2.1. *For all $\varepsilon > 0$, (1.1) has a unique smooth solution u_ε. Moreover, there exists a constant c, not depending on ε, such that*

$$\|u_\varepsilon\|_\infty + \operatorname{var}(u_\varepsilon) \leqq c.$$

Finally, there exists a unique (a.e.) function $\bar{u} \in BV$ such that

$$\|u_\varepsilon - \bar{u}\|_1 \to 0,$$

as $\varepsilon \downarrow 0$.

It is known that the limit $\bar{u} \in BV$ is the only (a.e.) BV function satisfying the following so-called "entropy inequality":

For all $k \in \mathbb{R}$ and all $\phi \in C_+^\infty$

$$\begin{aligned}\int_0^1 \operatorname{sgn}(\bar{u}-k)\left\{-(f(x,\bar{u})-f(x,k))\phi_x + \left(b(x,\bar{u}) + \frac{\partial}{\partial x} f(x,k)\right)\phi\right\} dx \\ + \operatorname{sgn}(\gamma_1 - k)\{f(1,\bar{u}(1-)) - f(1,k)\}\phi(1) \\ - \operatorname{sgn}(\gamma_0 - k)\{f(0,\bar{u}(0+)) - f(0,k)\}\phi(0) \leqq 0.\end{aligned}$$

We now state what we call the "approximate entropy inequality" (or AEI) for a parameterized family of BV functions $\{v_h\}_{0<h\leqq h_0}$. Below the family of solutions to (1.1) are shown to satisfy the AEI, and in § 4 certain numerical approximations are shown to satisfy the AEI as well.

DEFINITION 2.1. A family of BV functions $\{v_h\}_{0<h\leqq h_0}$ is said to satisfy the AEI if there exists nonnegative functions

$$R_h^0, \quad R_h^1, \quad R_h^2 \in BV, \qquad 0 < h \leqq h_0,$$

with the properties that

A. There exists a constant c independent of $0 < h \leqq h_0$ and $0 \leqq \delta \leqq 1$ such that

A0. $$\int_0^\delta R_h^0(s)\, ds \leqq c(h + \delta^2),$$

A1. $$\int_{1-\delta}^1 R_h^1(s)\, ds \leqq c(h + \delta^2),$$

A2. $$\int_0^1 R_h^2(s)\, ds \leqq ch;$$

B. For all $k \in \mathbb{R}$, all $\phi \in C_+^\infty$ and almost every $\alpha, \beta \in (0,1)$ we have

$$\begin{aligned}\int_0^1 \operatorname{sgn}(v_h - k)\left\{-(f(x,v_h) - f(x,k))\phi_x + \left(b(x,v_h) + \frac{\partial}{\partial x} f(x,k)\right)\phi\right\} dx \\ + \operatorname{sgn}(\gamma_1 - k)\{f(\beta, v_h(\beta)) - f(1,k)\}\phi(1) \\ - \operatorname{sgn}(\gamma_0 - k)\{f(\alpha, v_h(\alpha)) - f(0,k)\}\phi(0) \\ \leqq R_h^1(\beta)\phi(1) + R_h^0(\alpha)\phi(0) + \int_0^1 R_h^2 |\phi_x|\, dx.\end{aligned}$$

In § 3 we prove the following theorem.

THEOREM 2.1. *Let $\{v_h\}$ satisfy the AEI and assume that there exists a constant C_1 independent of $0 < h \leqq h_0$ such that $\|v_h\|_\infty + \operatorname{var}(v_h) \leqq C_1$. Then, there exists a constant C_2 independent of $0 < h \leqq h_0$ such that*

$$\|v_h - \bar{u}\|_1 \leqq C_2\sqrt{h},$$

where $\bar{u} = \lim_{\varepsilon\downarrow 0} u_\varepsilon$.

The result of Theorem 2.1 applies directly to the family of solutions to (1.1).

COROLLARY. *The family* $\{u_\varepsilon\}_{\varepsilon>0}$ *of solutions to* (1.1) *satisfies the AEI*; *consequently we have the estimate*

$$\|u_\varepsilon - \bar{u}\| \leqq C\sqrt{\varepsilon}.$$

Proof. Multiply the identity

$$0 = -\varepsilon \frac{d^2}{dx^2} u_\varepsilon + \frac{d}{dx}(f(x, u_\varepsilon) - f(x, k)) + \left(b(x, u_\varepsilon) + \frac{\partial}{\partial x} f(x, k) \right)$$

by $\operatorname{sgn}_\delta (u_\varepsilon - k)\phi$, $\phi \in C_+^\infty$, integrate over $0 \leqq x \leqq 1$, apply integration by parts, use the estimate

$$\int_0^1 \left(\frac{d}{dx} \operatorname{sgn}_\delta (u_\varepsilon - k) \frac{d}{dx} u_\varepsilon \right) \phi \, dx \geqq 0$$

and let $\delta \downarrow 0$ to find with Lebesgue's dominated convergence theorem that

$$\begin{aligned} \int_0^1 \operatorname{sgn}(u_\varepsilon - k) &\left\{ -(f(x, u_\varepsilon) - f(x, k))\phi_x + \left(b(x, u_\varepsilon) + \frac{\partial}{\partial x} f(x, k) \right) \phi \right\} dx \\ &+ \operatorname{sgn}(\gamma_1 - k) \left\{ f(1, \gamma_1) - f(1, k) - \varepsilon \frac{d}{dx} u_\varepsilon(1) \right\} \phi(1) \\ &- \operatorname{sgn}(\gamma_0 - k) \left\{ f(0, \gamma_0) - f(0, k) - \varepsilon \frac{d}{dx} u_\varepsilon(0) \right\} \phi(0) \\ \leqq -\varepsilon &\int_0^1 \operatorname{sgn}(u_\varepsilon - k) \left(\frac{d}{dx} u_\varepsilon \right) \phi_x \, dx. \end{aligned}$$

Integrating the differential equation (1.1) we make it evident that the boundary terms above can be written as

$$\operatorname{sgn}(\gamma_1 - k) \left\{ (f(\beta, u_\varepsilon(\beta)) - f(1, k)) - \varepsilon \frac{du_\varepsilon}{dx}(\beta) - \int_\beta^1 b(x, u_\varepsilon) \, dx \right\} \phi(1),$$

and

$$\operatorname{sgn}(\gamma_0 - k) \left\{ (f(\alpha, u_\varepsilon(\alpha)) - f(0, k)) - \varepsilon \frac{du_\varepsilon}{dx}(\alpha) + \int_0^\alpha b(x, u_\varepsilon) \, dx \right\} \phi(0).$$

Inserting these identities into the inequality above we easily find that u_ε satisfies the AEI with

$$R_\varepsilon^0(\alpha) = \varepsilon \left| \frac{d}{dx} u_\varepsilon(\alpha) \right| + \int_0^\alpha |b(x, u_\varepsilon)| \, dx,$$

$$R_\varepsilon^1(\beta) = \varepsilon \left| \frac{d}{dx} u_\varepsilon(\beta) \right| + \int_\beta^1 |b(x, u_\varepsilon)| \, dx,$$

$$R_\varepsilon^2(x) = \varepsilon \left| \frac{d}{dx} u_\varepsilon(x) \right|.$$

Using the result of Proposition 2.1 we finally conclude that R_ε^0, R_ε^1 and R_ε^2 above satisfy properties A0, A1 and A2 of Definition 2.1 with ε taking the role of h.

3. Proof of Theorem 2.1. Let $\{u_\varepsilon\}_{\varepsilon>0}$ denote the family of solutions to (1.1), and let $\{v_h\}_{h>0}$ denote a family of functions which have uniformly bounded variation and which satisfy the AEI. Moreover, let R^j_ε, (resp. R^j_h), with $j=0,1,2$, represent the estimating functions of the AEI of Definition 2.1 for $\{u_\varepsilon\}_{\varepsilon>0}$, (resp. $\{v_h\}_{h>0}$). The proof of Theorem 2.1 is essentially a test function argument with a particular family of test functions ϕ_δ of the form

$$\phi_\delta(x)=\frac{1}{\delta}\phi\left(\frac{x}{\delta}\right),$$

where $\phi\in C^\infty_+$ is symmetric, $\int_{-\infty}^{\infty}\phi(x)\,dx=1$ and supp $(\phi)\subset(-1,1)$. Now consider the AEI of Definition 2.1 applied to $u_\varepsilon(x)$ with $k=v_h(y)$ and the test function ϕ replaced by $\phi(x)=\phi_\delta(x-y)$ and with $\alpha=\beta=y$ in R^j_h, $j=0,1$. Integrate the resulting inequality from $y=0$ to $y=1$. Since v_h also satisfies the AEI, the same procedure can be done as above with the roles of v_h and u_ε reversed. Adding both resulting inequalities together we obtain

$$\begin{aligned}\int_0^1\int_0^1 &\operatorname{sgn}(u_\varepsilon(x)-v_h(y))\Big\{(b(x,u_\varepsilon(x))-b(y,v_h(y)))\\ &\qquad+\left(\frac{\partial}{\partial x}f(x,v_h(y))-\frac{\partial}{\partial y}f(y,u_\varepsilon(x))\right)\Big\}\phi_\delta(x-y)\,dx\,dy\\ &\qquad+\int_0^1\int_0^1\operatorname{sgn}((u_\varepsilon(x)-v_h(y))\{(f(x,v_h(y))-f(y,v_h(y)))\\ &\qquad+(f(y,u_\varepsilon(x))-f(x,u_\varepsilon(x)))\}\frac{d}{dx}\phi_\delta(x-y)\,dx\,dy+T_1(\varepsilon,h,\delta)-T_0(\varepsilon,h,\delta)\\ &\leqq P^0_h(\delta)+P^1_h(\delta)+P^2_h(\delta)+P^0_\varepsilon(\delta)+P^1_\varepsilon(\delta)+P^2_\varepsilon(\delta),\end{aligned}\tag{3.1}$$

where

$$\begin{aligned}T_1(\varepsilon,h,\delta)=&\int_0^1\operatorname{sgn}(\gamma_1-v_h(y))\{f(y,u_\varepsilon(y))-f(1,v_h(y))\}\phi_\delta(1-y)\,dy\\ &+\int_0^1\operatorname{sgn}(\gamma_1-u_\varepsilon(x))\{f(x,v_h(x))-f(1,u_\varepsilon(x))\}\phi_\delta(1-x)\,dx,\end{aligned}\tag{3.2}$$

$$\begin{aligned}T_0(\varepsilon,h,\delta)=&\int_0^1\operatorname{sgn}(\gamma_0-v_h(y))\{f(y,u_\varepsilon(y))-f(0,v_h(y))\}\phi_\delta(y)\,dy\\ &+\int_0^1\operatorname{sgn}(\gamma_0-u_\varepsilon(x))\{f(x,v_h(x))-f(0,u_\varepsilon(x))\}\phi_\delta(x)\,dx,\end{aligned}\tag{3.3}$$

and

$$P^0_h(\delta)=\int_0^1 R^0_h(x)\phi_\delta(x)\,dx,\tag{3.4}$$

$$P^1_h(\delta)=\int_0^1 R^1_h(x)\phi_\delta(1-x)\,dx,\tag{3.5}$$

$$P^2_h(\delta)=\int_0^1\int_0^1 R^2_h(y)\left|\frac{d}{dx}\phi_\delta(x-y)\right|\,dx\,dy,\tag{3.6}$$

and a similar expression for $P^j_\varepsilon(\delta)$. The proof is divided into basically the following four lemmas.

LEMMA 3.1. *There is a constant c independent of $\varepsilon > 0$, $h > 0$, and $\delta > 0$ such that we have*

$$r.h.s. \leqq c\left(\frac{\varepsilon}{\delta}+\frac{h}{\delta}+\delta\right),$$

where r.h.s. is the right-hand side of (3.1).

LEMMA 3.2. *There is another constant c as above such that we have*

$$T_0(\varepsilon, h, \delta) \leqq c\left(\frac{\varepsilon}{\delta}+\frac{h}{\delta}+\delta\right), \qquad -T_1(\varepsilon, h, \delta) \leqq c\left(\frac{\varepsilon}{\delta}+\frac{h}{\delta}+\delta\right),$$

where T_0 and T_1 are defined in (3.2), (3.3).

LEMMA 3.3. *There is a constant c independent of positive ε, h and δ such that*

$$\left|\int_0^1 \int_0^1 \operatorname{sgn}(u_\varepsilon(x) - v_h(y))\{(f(x, v_h(y)) - f(y, v_h(y))) + (f(y, u_\varepsilon(x)) - f(x, u_\varepsilon(x)))\} \frac{d}{dx}\phi_\delta(x-y)\, dx\, dy\right| \leqq c\delta.$$

LEMMA 3.4. *There is a constant c independent of positive ε, h, $\delta \leqq 1/2$ such that*

$$\begin{aligned}\int_0^1 &\left|\left(b(x, u_\varepsilon(x)) - \frac{\partial}{\partial x} f(x, u_\varepsilon(x))\right) - \left(b(x, v_h(x)) - \frac{\partial}{\partial x} f(x, v_h(x))\right)\right| dx \\ &\leqq 2\int_0^1\int_0^1 \operatorname{sgn}(u_\varepsilon(x) - v_h(y))\left\{(b(x, u_\varepsilon(x)) - b(y, v_h(y))) \right.\\ &\qquad \left. + \left(\frac{\partial}{\partial x} f(x, v_h(y)) - \frac{\partial}{\partial y} f(y, u_\varepsilon(x))\right)\right\} \phi_\delta(x-y)\, dx\, dy + c\delta.\end{aligned}$$

Given the results above, the final result follows by first noting that, along with condition (1.2), they imply

$$\begin{aligned}\mu &\int_0^1 |u_\varepsilon(x) - v_h(x)|\, dx \\ &\leqq \int_0^1 \left|\left(b(x, u_\varepsilon(x)) - \frac{\partial}{\partial x} f(x, u_\varepsilon(x))\right) - \left(b(x, v_h(x)) - \frac{\partial}{\partial x} f(x, v_h(x))\right)\right| dx \\ &\leqq \hat{c}\left(\frac{\varepsilon}{\delta}+\frac{h}{\delta}+\delta\right),\end{aligned}$$

with $\hat{c}$ independent of positive ε, h and $\delta \leqq \frac{1}{2}$. Sending ε to zero, we conclude that

$$\|\bar{u} - v_h\|_1 \leqq \frac{\hat{c}}{\mu}\left(\frac{h}{\delta}+\delta\right),$$

and choosing $\delta = \sqrt{h}$ proves the theorem.

Proof of Lemma 3.1. The terms to be estimated are given in (3.4)-(3.6). Note that

$$P_h^0(\delta) = \int_0^\delta R_h^0(x)\phi_\delta(x)\, dx \leqq \frac{C}{\delta}(h+\delta^2)$$

and

$$P_h^2(\delta) = \int_0^1 R_h^2(y)\left\{\int_0^1 \left|\frac{d}{dx}\phi_\delta(y-x)\right| dx\right\} dy \leqq C\frac{h}{\delta},$$

where the first inequality above follows from the definition $R_h^0(x)$ and $\phi_\delta(x)$, and the second follows from the definition of $R_h^2(y)$ and the fact that

$$\int_0^1 \left|\frac{d}{dx}\phi_\delta(y-x)\right| dx \leqq \frac{1}{\delta^2}\int_{-\infty}^{\infty} \left|\phi'\left(\frac{z}{\delta}\right)\right| dz \leqq C/\delta.$$

Similar estimates hold for the remaining terms on the right-hand side of (3.1).

Proof of Lemma 3.2. We only estimate T_0 since $-T_1$ can be treated similarly. Note that T_0 can be written as

$$\begin{aligned} T_0(\varepsilon, h, \delta) = \int_0^1 & \{\operatorname{sgn}(\gamma_0 - v_h) - \operatorname{sgn}(\gamma_0 - u_\varepsilon)\} \\ & \cdot \{(f(x, u_\varepsilon) - f(x, \gamma_0)) + (f(x, \gamma_0) - f(x, v_h))\}\phi_\delta(x)\, dx \\ & + \int_0^1 \operatorname{sgn}(\gamma_0 - v_h)\{f(x, v_h) - f(0, v_h)\}\phi_\delta(x)\, dx \\ & + \int_0^1 \operatorname{sgn}(\gamma_0 - u_\varepsilon)\{f(x, u_\varepsilon) - f(0, u_\varepsilon)\}\phi_\delta(x)\, dx. \end{aligned}$$

Clearly, the second two terms above can be bounded above by $C\delta$. The first term above is bounded above by

$$\begin{aligned} & \int_0^1 [|f(x, u_\varepsilon) - f(x, \gamma_0)| + \operatorname{sgn}(u_\varepsilon - \gamma_0)(f(x, u_\varepsilon) - f(x, \gamma_0))]\phi_\delta(x)\, dx \\ & \quad + \int_0^1 [|f(x, v_h) - f(x, \gamma_0)| + \operatorname{sgn}(v_h - \gamma_0)(f(x, v_h) - f(x, \gamma_0))]\phi_\delta(x)\, dx. \end{aligned} \tag{3.7}$$

We estimate the first integral above only since the second integral can be estimated similarly. Return now to the AEI applied to u_ε, and set $k = \gamma_0$ and suppose there we replace the test function $\phi(x)$ with a test function approaching

$$H(x - x_0) = \begin{cases} 1, & x < x_0, \\ 0, & x \geqq x_0. \end{cases}$$

Doing so we find that for almost every $x_0 \in (0, 1)$ the AEI implies that

$$\begin{aligned} \operatorname{sgn}(u_\varepsilon(x_0) - \gamma_0)\{f(x_0, u_\varepsilon(x_0)) - f(x_0, \gamma_0)\} & \leqq R_\varepsilon^0(x_0) + R_\varepsilon^2(x_0) + cx_0 \\ & \equiv e_\varepsilon(x_0). \end{aligned} \tag{3.8}$$

To find an estimate for the integral in question, we note the obvious implication: If $Q \leqq e$ then $|Q| + Q \leqq 2e$. Therefore, applying (3.8) (and the analogous estimate for v_h) to (3.7) shows that (3.7) is bounded above by

$$2\int_0^1 [e_\varepsilon(x) + e_h(x)]\phi_\delta(x)\, dx.$$

Finally, applying the simple estimates of the previous lemma completes the proof of the present lemma.

The proof of Lemma 3.3 is routine and is left to the reader.

Proof of Lemma 3.4. The proof of this lemma can be given for fixed $\varepsilon > 0$, $h > 0$, using only condition (1.2) and the uniform estimate

$$\|u_\varepsilon\|_\infty + \int_0^1 \left|\frac{d}{dx}u_\varepsilon\right| dx \leqq C_0.$$

For convenience we omit the subscripts ε and h. Condition (1.2) and a simple rearrangement gives us that

$$\begin{aligned}\left|(b(y,u(y))-\frac{\partial}{\partial y}f(y,u(y)))-\left(b(y,v(y))-\frac{\partial}{\partial y}f(y,v(y))\right)\right|&\\ \leqq \operatorname{sgn}(u(x)-v(y))\Big\{(b(x,u(x))-b(y,v(y)))&\\ +\left(\frac{\partial}{\partial x}f(x,v(y))-\frac{\partial}{\partial y}f(y,u(x))\right)\Big\}&\\ +L_1|x-y|+L_2|u(x)-u(y)|,&\end{aligned}$$

where

$$L_1=\max\left\{\left|\frac{\partial}{\partial x}b(x,u)\right|+\left|\frac{\partial^2}{\partial x^2}f(x,u)\right|, u\in I, x\in[0,1]\right\},$$

$$L_2=\max\left\{\left|\frac{\partial}{\partial u}b(x,u)\right|+\left|\frac{\partial^2}{\partial x\partial u}f(x,u)\right|, u\in I, x\in[0,1]\right\}.$$

Furthermore, it is obvious that for all $0\leqq y\leqq 1$ and all $\frac{1}{2}\geqq\delta>0$

$$\int_0^1 \phi_\delta(x-y)\,dx\geqq\tfrac{1}{2}.$$

Therefore

$$\begin{aligned}&\int_0^1\left|\left(b(y,u(y))-\frac{\partial}{\partial y}f(y,u(y))\right)-\left(b(y,v(y))-\frac{\partial}{\partial y}f(y,v(y))\right)\right|dy\\ &\leqq 2L_1\int_0^1\int_0^1|x-y|\phi_\delta(x-y)\,dx\,dy+2L_2\int_0^1\int_0^1|u(x)-u(y)|\phi_\delta(x-y)\,dx\,dy\\ &+2\int_0^1\int_0^1\operatorname{sgn}(u(x)-v(y))\\ &\cdot\left\{(b(x,u(x))-b(y,v(y)))+\left(\frac{\partial}{\partial x}f(x,v(y))-\frac{\partial}{\partial y}f(y,u(x))\right)\right\}\phi_\delta(x-y)\,dx\,dy.\end{aligned}$$

To complete the proof we need only estimate the second term on the right-hand side above since the other terms have been dealt with already. To see that

$$\int_0^1\int_0^1|u(x)-u(y)|\phi_\delta(x-y)\,dx\,dy\leqq C\delta,$$

extend the smooth function $u(x)$ to the whole real line by $u(x)=u(0)$ for $x<0$ and $u(x)=u(1)$ for $x>1$. Then for each $0\leqq y\leqq 1$ we have

$$\begin{aligned}\int_0^1|u(x)-u(y)|\phi_\delta(x-y)\,dx&\leqq\int_{y-\delta}^{y+\delta}\int_{y-\delta}^{y+\delta}|u'(s)|\phi_\delta(x-y)\,dx\,ds\\ &=\int_{-\delta}^{\delta}|u'(y+s)|\,ds.\end{aligned}$$

Integrating this inequality with respect to y and interchanging the order of integration makes the desired result obvious.

4. Application to difference schemes. In this section we give another application of the AEI. We show that certain finite difference schemes yield approximations that satisfy the AEI; hence, according to Theorem 2.1, they satisfy the $\sqrt{h}L^1$ rate of convergence. We begin with a few preliminaries.

Partition the interval $[0, 1]$ into subintervals $I_j=[x_j, x_{j+1}]$, $0\leqq j\leqq J-1$, with $x_0=0$ and $x_J=1$, and define $\Delta x_j=(x_{j+1}-x_j)$. Define the approximate solution v_h by

$$v_h(x)=\sum_{j=0}^{J-1} u_j\chi_{I_j}(x),$$

where χ_{I_j} is the characteristic function of the interval I_j. We shall consider a class of finite difference schemes of the form

$$\begin{aligned}&\Delta^+F(x_j, u_j, u_{j-1})+\Delta x_jB(x_j, x_{j+1}, u_j)=0, \qquad 0\leqq j\leqq J-1,\\ &u_{-1}=\gamma_0, \qquad u_J=\gamma_1,\end{aligned} \tag{4.1}$$

where the forward difference operator Δ^+ is defined by $\Delta^+\alpha_j=\alpha_{j+1}-\alpha_j$ and $B(x_j, x_{j+1}, u_j)$ is given by

$$B(x_j, x_{j+1}, u)=b(x_j, u)+\frac{\partial}{\partial x}f(x_j, u)-\frac{1}{\Delta x_j}(f(x_{j+1}, u)-f(x_j, u)). \tag{4.2}$$

The numerical flux function $F(\cdot,\cdot,\cdot)$ of (4.1) is assumed throughout to satisfy the following properties:

F1. $F(x, u, u)=f(x, u)$.
F2a. $u\to F(x, u, v)$ is nonincreasing for all $x\in[0, 1]$, $v, u\in\mathbb{R}$.
F2b. $v\to F(x, u, v)$ is nondecreasing for all $x\in[0, 1]$, $u, v\in\mathbb{R}$.
F3. $F(x, u, v)$ is Lipschitz continuous in x, u, v.[2]

We now give three examples of numerical flux functions that satisfy the properties above.

1) Lax-Friedrichs [7]:

$$F(x, u, v)=\tfrac{1}{2}\{f(x, u)+f(x, v)-\lambda(u-v)\},$$

where $\lambda\geqq\|(\partial/\partial u)f(x, u)\|_\infty$.

2) Godunov [10]:

$$F(x, u, v)=\begin{cases}\max\limits_{u\leqq s\leqq v} f(x, s) & \text{if } u\leqq v,\\ \min\limits_{v\leqq s\leqq u} f(x, s) & \text{if } v\leqq u.\end{cases}$$

3) Engquist-Osher [1], [11].

$$F(x, u, v)=\int_0^u \min\left(\frac{\partial}{\partial s}f(x, s), 0\right) ds+\int_0^v \max\left(\frac{\partial}{\partial s}f(x, s), 0\right) ds+f(x, 0).$$

The following theorem is a straightforward extension of known results [1], [8].

[2] Assumption F3 need only be valid in the a priori interval determined by the maximum principle.

THEOREM 4.1. *Under properties* F1, F2, *and* F3, *and condition* (1.2) *of* § 1, *the difference scheme* (4.1) *has a unique solution for every grid. Moreover, there exists a constant c independent of the grid such that*

$$\|v_h\|_\infty + \operatorname{var}(v_h) \leqq c.$$

By using the results of Theorem 4.1 we next prove the following.

THEOREM 4.2. *Under the conditions of the previous theorem, the family of approximations* $\{v_h\}$ *satisfy the AEI; consequently by Theorem* 2.1 *they satisfy the rate of convergence*

$$\|v_h - \bar{u}\|_1 \leqq c\sqrt{h},$$

where $h = \max \Delta x_j$, $\bar{u} = \lim_{\varepsilon \downarrow 0} u_\varepsilon$ *and for some constant c which does not depend on h.*

Proof of Theorem 4.2. For arbitrary $k \in \mathbb{R}$, $\phi \in C_+^\infty$ and $0 \leqq \alpha$, $\beta \leqq 1$, we estimate the quantity

$$\begin{aligned}&\int_0^1 \operatorname{sgn}(v_h - k)\left\{-(f(x, v_h) - f(x, k))\phi_x + (b(x, v_h) + \frac{\partial}{\partial x} f(x, k))\phi\right\} dx \\ &\quad + \operatorname{sgn}(\gamma_1 - k)\{f(\beta, v_h(\beta)) - f(1, k)\}\phi(1) \\ &\quad - \operatorname{sgn}(\gamma_0 - k)\{f(\alpha, v_h(\alpha)) - f(0, k)\}\phi(0),\end{aligned} \tag{4.3}$$

where v_h is the piecewise constant interpolation of grid values generated by (4.1). Using the explicit form of v_h and integration by parts we find that the integral term in (4.3) is given by

$$\begin{aligned}\sum_{j=0}^{J-1} \operatorname{sgn}(u_j - k)\Big\{&-(f(x_{j+1}, u_j) - f(x_{j+1}, k))\phi(x_{j+1}) + (f(x_j, u_j) - f(x_j, k))\phi(x_j) \\ &+ \int_{x_j}^{x_{j+1}} \left(\frac{\partial}{\partial x} f(x, u_j) + b(x, u_j)\right)\phi(x)\, dx\Big\}.\end{aligned} \tag{4.4}$$

Rearranging terms and then adding and subtracting $F(x_{j+1}, u_{j+1}, u_j)$ and $F(x_j, u_j, u_{j-1})$ into this result we get that (4.4) equals

$$\begin{aligned}&\sum_{j=1}^{J} \operatorname{sgn}(u_{j-1} - k)[-(F(x_j, u_j, u_{j-1}) - f(x_j, k))]\phi(x_j) \\ &\quad + \sum_{j=0}^{J-1} \operatorname{sgn}(u_j - k)[(F(x_j, u_j, u_{j-1}) - f(x_j, k))]\phi(x_j) \\ &\quad + \sum_{j=0}^{J-1} \operatorname{sgn}(u_j - k)[(F(x_{j+1}, u_{j+1}, u_j) - f(x_{j+1}, u_j))]\phi(x_{j+1}) \\ &\quad + \sum_{j=0}^{J-1} \operatorname{sgn}(u_j - k)[-(F(x_j, u_j, u_{j-1}) - f(x_j, u_j))]\phi(x_j) \\ &\quad + \sum_{j=0}^{J-1} \operatorname{sgn}(u_j - k) \int_{x_j}^{x_{j+1}} \left(\frac{\partial}{\partial x} f(x, u_j) + b(x, u_j)\right)\phi(x)\, dx \\ &\quad = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}.\end{aligned} \tag{4.5}$$

Before proceeding we give two simple lemmas.

LEMMA 4.1. *For any three numbers a, b,* $k \in \mathbb{R}$ *and* $x \in [0, 1]$ *we have*

$$\{\operatorname{sgn}(b - k) - \operatorname{sgn}(a - k)\}\{F(x, b, a) - f(x, k)\} \leqq 0.$$

Proof. The quantity above can be written as

$$\{\cdots\}\{F(x, b, a) - F(x, b, k)\} + \{\cdots\}\{F(x, b, k) - F(x, k, k)\},$$

where $\{\cdots\}=\{\operatorname{sgn}(b-k)-\operatorname{sgn}(a-k)\}$ and where we have used property F1 to write $f(x,k)=F(x,k,k)$. We can now bound this above by

$$|F(x,b,a)-F(x,b,k)|-\operatorname{sgn}(a-k)\{F(x,b,a)-F(x,b,k)\}$$
$$+\operatorname{sgn}(b-k)\{F(x,b,k)-F(x,k,k)\}+|F(x,b,k)-F(x,k,k)|.$$

Finally, using properties F2a and F2b shows us that this quantity is equal to zero.

LEMMA 4.2. *If $H(x)\in C^1$, $\phi(x)\in C^1$, then*

$$\left|\int_{x_j}^{x_{j+1}} H(x)\phi(x)\,dx-\Delta x_j H(x_j)\phi(x_j)\right|$$
$$\leqq \Delta x_j\left[\left\|\frac{d}{dx}H\right\|_\infty \int_{x_j}^{x_{j+1}}|\phi(s)|\,ds+\|H\|_\infty\int_{x_j}^{x_{j+1}}\left|\frac{d}{ds}\phi(s)\right|\,ds\right].$$

Proof. The left-hand side of the inequality above can be written as

$$\left|\int_{x_j}^{x_{j+1}}[(H(x)-H(x_j))\phi(x)+H(x_j)(\phi(x)-\phi(x_j))]\,dx\right|,$$

which is bounded above by

$$\Delta x_j\left\|\frac{d}{dx}H\right\|_\infty\int_{x_j}^{x_{j+1}}|\phi(x)|\,dx+\|H\|_\infty\left|\int_{x_j}^{x_{j+1}}\int_{x_j}^{x}\frac{d}{ds}\phi(s)\,ds\,dx\right|.$$

The final estimate is now obvious.

Continuing the proof of the theorem, it is clear with the result of Lemma 4.1 that the sums of terms I and II of (4.5) is bounded above by

$$\mathrm{I}+\mathrm{II}\leqq -\operatorname{sgn}(u_{J-1}-k)\{F(1,\gamma_1,u_{J-1})-f(1,k)\}\phi(1)$$
$$+\operatorname{sgn}(u_0-k)\{F(0,u_0,\gamma_0)-f(0,k)\}\phi(0).$$

Moreover, the sum of III and IV can be rewritten as

$$\sum_{j=0}^{J-1}\operatorname{sgn}(u_j-k)\{F(x_{j+1},u_{j+1},u_j)-f(x_{j+1},u_j)\}(\phi(x_{j+1})-\phi(x_j))$$
$$+\sum_{j=0}^{J-1}\operatorname{sgn}(u_j-k)\{(F(x_{j+1},u_{j+1},u_j)-F(x_j,u_j,u_{j-1}))$$
$$-(f(x_{j+1},u_j)-f(x_j,u_j))\}\phi(x_j),$$

and using Lemma 4.2 we see that term V of (4.5) can be bounded above by

$$\sum_{j=0}^{J-1}\operatorname{sgn}(u_j-k)\left\{\left(\frac{\partial}{\partial x}f(x_j,u_j)+b(x_j,u_j)\right)\Delta x_j\right\}$$
$$+h\left(\left\|\frac{d}{dx}H(x,v_h)\right\|_\infty+\|H(x,v_h)\|_\infty\right)\int_0^1\left(|\phi(x)|+\left|\frac{d}{dx}\phi(x)\right|\right)dx,$$

where $H(x,u)=(\partial/\partial x)f(x,u)+b(x,u)$ and $h=\max\Delta x_j$. Combining these estimates and using the difference scheme (4.1) we conclude that

$$\int_0^1\operatorname{sgn}(v_h-k)\left\{-(f(x,v_h)-f(x,k))\phi_x+(b(x,v_h)+\frac{\partial}{\partial x}f(x,k))\phi\right\}dx$$
$$\leqq -\operatorname{sgn}(u_{J-1}-k)\{F(1,\gamma_1,u_{J-1})-f(1,k)\}\phi(1)$$
$$+\operatorname{sgn}(u_0-k)\{F(0,u_0,\gamma_0)-f(0,k)\}\phi(0) \tag{4.6}$$
$$+\sum_{j=0}^{J-1}L|u_{j+1}-u_j|\int_{x_j}^{x_{j+1}}|\phi_x|\,dx+h\text{ const.}\left\{\int_0^1(|\phi|+|\phi_x|)\,dx\right\},$$

where L is the Lipschitz constant, given by

$$L=\sup\left\{\frac{|F(x,u,v)-F(x,v,v)|}{|u-v|},\ u,v\in I,\ x\in[0,1]\right\}.$$

Next we include the boundary terms of the AEI into the inequality above. Using the fact that Lemma 4.1 implies that

$$-\operatorname{sgn}(u_{J-1}-k)\{F(1,\gamma_1,u_{J-1})-f(1,k)\}\leqq-\operatorname{sgn}(\gamma_1-k)\{F(1,\gamma_1,u_{J-1})-f(1,k)\}$$

and

$$\operatorname{sgn}(u_0-k)\{F(0,u_0,\gamma_0)-f(0,k)\}\leqq\operatorname{sgn}(\gamma_0-k)\{F(0,u_0,\gamma_0)-f(0,k)\},$$

we find that for $\alpha,\beta\in(0,1)$

$$\begin{aligned}(4.7)\qquad &\int_0^1\operatorname{sgn}(v_h-k)\left\{-(f(x,v_h)-f(x,k))\phi_x+\left(b(x,v_h)+\frac{\partial}{\partial x}f(x,k)\right)\phi\right\}dx\\&+\operatorname{sgn}(\gamma_1-k)\{f(\beta,v_h(\beta))-f(1,k)\}\phi(1)\\&-\operatorname{sgn}(\gamma_0-k)\{f(\alpha,v_h(\alpha))-f(0,k)\}\phi(0)\end{aligned}$$

is bounded above by

$$\begin{aligned}(4.8)\qquad &\operatorname{sgn}(\gamma_1-k)\{f(\beta,v_h(\beta))-F(1,\gamma_1,u_{J-1})\}\phi(1)\\&-\operatorname{sgn}(\gamma_0-k)\{f(\alpha,v_h(\alpha))-F(0,u_0,\gamma_0)\}\phi(0)\\&+L\sum_{j=0}^{J-1}|u_{j+1}-u_j|\int_{x_j}^{x_{j+1}}|\phi_x|\,dx+h\text{ const.}\left\{\int_0^1(|\phi|+|\phi_x|)\,dx\right\}.\end{aligned}$$

We estimate the first boundary term above only since the second can be estimated in a similar way. Sum (4.1) from $j=j_1$ to $J-1$ where j_1 is chosen so that $x_{j_1}\leqq\beta<x_{j_1+1}$. Doing so, we substitute the result into the first term of (4.8) and find that

$$\begin{aligned}&\operatorname{sgn}(\gamma_1-k)\{f(\beta,v_h(\beta))-F(1,\gamma_1,u_{J-1})\}\\&\qquad\leqq L|u_{j_1}-u_{j_1-1}|+\left\|\frac{\partial}{\partial x}f(x,u)\right\|_\infty|\beta-x_{j_1}|+\sum_{j=j_1}^{j-1}|B(x_j,x_{j+1},u_j)|\Delta x_j\\&\qquad\leqq L|u_{j_1}-u_{j_1-1}|+\text{const.}\,(h+1-\beta).\end{aligned}$$

Now define

$$H_h(x)=\sum_{j=0}^{J-1}\frac{|u_{j+1}-u_j|}{\Delta x_j}\chi_{I_j}(x),$$

and note that

$$\int_0^1H_h(x)\,dx=\sum_{j=0}^{J-1}|u_{j+1}-u_j|\leqq\operatorname{var}(v_h).$$

Inserting this and the estimate above into (4.8) we have that (4.7) is bounded above by

$$\text{const.}\left\{(hH_h(\beta)+1-\beta+h)\phi(1)+(hH_h(\alpha)+\alpha+h)\phi(0)+h\int_0^1(H_h(x)+1)|\phi_x|\,dx\right\}.$$

Note that above we have used the fact that

$$\int_0^1|\phi(x)|\,dx\leqq\int_0^1|\phi_x(x)|\,dx+\phi(0).$$

Reading off the terms $R_h^0(\alpha)$, $R_h^1(\beta)$ and $R_h^2(x)$ from above one easily establishes that they satisfy properties A0, A1 and A2 of Definition 2.1. Therefore the families of approximate solutions generated by finite difference schemes of the form (4.1) satisfy the AEI of Definition 2.1. This completes the proof of Theorem 4.2.

REFERENCES

[1] L. ABRAHAMSSON AND S. OSHER, *Monotone difference schemes for singular perturbation problems*, SIAM J. Numer. Anal., 19 (1982), pp. 979-992.

[2] C. BARDOS, A. Y. LEROUX AND J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017-1034.

[3] P. BÉNILAN AND H. TOURÉ, *Sur l'équation génerale* $u_t = \phi(u)_{xx} - \psi(u)_x + v$, C.R. Acad. Sci., Paris, Sér. I Math., 18 (1984), pp. 919-922.

[4] F. A. HOWES, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203 (1978).

[5] S. N. KRUZKOV, *First order quasi-linear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217-243.

[6] N. N. KUZNETSOV, *On stable methods for nonlinear first order partial differential equations in the class of discontinuous functions*, Proc. Roy. Irish Acad. Conf. Numer. Anal. (1976), pp. 183-197.

[7] P. D. LAX, *Weak solutions of nonlinear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math., 7 (1954), pp. 159-193.

[8] J. LORENZ, *Nonlinear boundary value problems with turning points and properties of difference schemes*, in Lecture Notes in Mathematics 942, W. Eckhaus and E. M. de Jager, eds., Springer-Verlag, Berlin, 1982.

[9] J. LORENZ AND R. SANDERS, *Second order nonlinear singular perturbation problems with boundary conditions of mixed type*, this Journal, 17 (1986), pp. 580-594.

[10] O. A. OLEINIK, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl., 26 (1962), pp. 95-172.

[11] S. OSHER, *Nonlinear singular perturbation problems and one-sided difference schemes*, SIAM J. Numer. Anal., 18 (1981), pp. 129-144.

[12] B. PERTHAME AND R. SANDERS, *The Neumann problem for nonlinear second order singular perturbation problems*, this Journal, to appear.

[13] R. SANDERS, *On convergence of monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 91-106.

[14] A. I. VOL'PERT, *The spaces BV and quasilinear equations*, Math. USSR-Sb., 2 (1967), pp. 225-267.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

004

NONMONOTONE INTERIOR LAYER THEORY FOR SOME SINGULARLY PERTURBED QUASILINEAR BOUNDARY VALUE PROBLEMS WITH TURNING POINTS*

ALBERT J. DESANTI†

Abstract. The quasilinear singular perturbation problem $\varepsilon y''=f(x,y)y'+g(x,y)$, $y(-1,\varepsilon)=A$, $y(1,\varepsilon)=B$ is studied under the principal assumption that $f(0,y)=0$ for all y, i.e., that $x=0$ is a turning point for the function f. Under explicit conditions on f, g, A and B, solutions are shown to exhibit one of two types of nonmonotone interior layer behavior: (i) spike layer behavior or (ii) nonmonotone transition layer behavior. The results are obtained using a method based on the theory of differential inequalities. Applications and examples are discussed in detail.

Key words. singular perturbation, quasilinear boundary value problems, turning points, differential inequalities, spike layer behavior, nonmonotone transition layer behavior

AMS(MOS) subject classifications. 34E20, 34A40

1. Introduction. In this paper we study the asymptotic behavior (as $\varepsilon\to 0$) of solutions of the singularly perturbed quasilinear boundary value problem

$$(\mathrm{P}_1)\qquad \begin{aligned}&\varepsilon y''=f(x,y)y'+g(x,y), \qquad -1<x<1,\\ &y(-1,\varepsilon)=A, \qquad y(1,\varepsilon)=B\end{aligned}$$

where ε is a small parameter, f and g are smooth functions, $f(0,y)=0$ for all y (i.e., $x=0$ is a turning point), and A and B are prescribed. In particular, we are interested in establishing the existence, for ε sufficiently small, of a solution of problem (P_1) which exhibits spike layer behavior and a solution which exhibits nonmonotone transition layer behavior at $x=0$. We say that a solution $y=y(x,\varepsilon)$ of problem (P_1) exhibits spike layer behavior at $x=0$ if

$$\lim_{\varepsilon\to 0} y(x,\varepsilon)=\begin{cases} u_L(x) & \text{for } -1<x<0,\ 0<x<1,\\ s & \text{for } x=0\end{cases}$$

where $u_L=u_L(x)$ is a certain solution of the reduced problem

$$\begin{aligned}&0=f(x,u)u'+g(x,u), \qquad -1<x<1,\\ &u(-1)=A \qquad (\text{or } u(1)=B),\end{aligned}$$

and $s\neq u_L(0)$. Similarly, we say that $y(x,\varepsilon)$ exhibits nonmonotone transition layer behavior if

$$\lim_{\varepsilon\to 0} y(x,\varepsilon)=\begin{cases} u_L(x) & \text{for } -1<x<0,\\ s & \text{for } x=0,\\ u_R(x) & \text{for } 0<x<1\end{cases}$$

where u_L and u_R are certain solutions of the reduced problems

$$\begin{aligned}&0=f(x,u)u'+g(x,u), \qquad u(-1)=A,\\ &0=f(x,u)u'+g(x,u), \qquad u(1)=B,\end{aligned}$$

respectively, and s is such that $s>\max\{u_L(0),u_R(0)\}$ or $s<\min\{u_L(0),u_R(0)\}$.

* Received by the editors July 8, 1985; accepted for publication (in revised form) February 19, 1986.

† Center For Naval Analyses, 2000 N. Beauregard Street, Alexandria, Virginia 22311. Present address, Systems Analysis Branch, Code 3196, Naval Weapons Center, China Lake, California 93555.

An example of a function exhibiting spike layer behavior is $y_1(x, \varepsilon) = \operatorname{sech}^2(x/\varepsilon)$. We note that $y_1(x, \varepsilon) \to 0$ for $x \neq 0$ and $y_1(0, \varepsilon) = 1$. The function y_1 is sketched in Fig. 1.

Interior nonmonotone transition layer behavior is exemplified by the function $y_2(x, \varepsilon) = \frac{1}{2}\tanh(x/\varepsilon) + \operatorname{sech}^2(x/\varepsilon)$. We note that $y_2(x, \varepsilon) \to -\frac{1}{2}$ for $x < 0$, $y_2(x, \varepsilon) \to \frac{1}{2}$ for $x > 0$, and $y_2(0, \varepsilon) = 1$. The function y_2 is sketched in Fig. 2.

The existence of a solution of problem (P_1) exhibiting interior layer behavior depends upon the behavior of the function f near the turning point $x = 0$. In other words, interior layer behavior is possible only if the function f changes its algebraic sign in passing through zero. Howes [8] showed that this is true for shock layer behavior, and we shall show that this is true for spike and nonmonotone transition layer behavior. If f does not change sign across zero, then Howes has shown that problem (P_1) admits only solutions which exhibit boundary layer behavior at one or both of the endpoints.

Physically, problem (P_1) may be interpreted as a model for a one-dimensional, steady-state, reaction-diffusion-convection system. In this connection, ε is a measure of the diffusivity of the medium, the function f is a measure of convection, and the

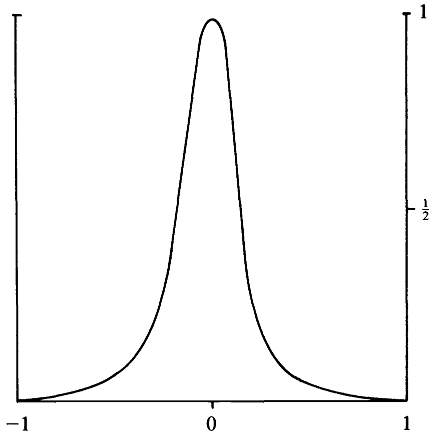


FIG. 1. *Graph of* $y_1(x) = \operatorname{sech}^2(x/\varepsilon)$.

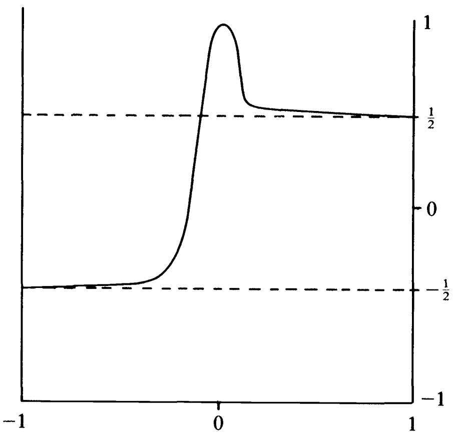


FIG. 2. *Graph of* $y_2(x) = \frac{1}{2}\tanh(x/\varepsilon) + \operatorname{sech}^2(x/\varepsilon)$.

function g is a measure of the effect of reaction and of sources in the medium. Under this interpretation, spike layer solutions of problem (P_1) are reasonable. Since the convection term is zero at $x=0$, the dominant mechanism of transport of substance or heat away from $x=0$ is diffusion, which is a slow process since the diffusivity of the medium is small. Hence, if there is a source or a reaction at $x=0$, a buildup of substance or of heat is likely.

Problem (P_1) may also be interpreted as a model for a nonlinear mass-spring system, or, more generally, simply as a formulation of Newton's Second Law of Motion. In this connection, x is time, ε is the mass of the object, f is a measure of the damping effect of the medium and g is a nonlinear restoring force.

Nonmonotone transition layer behavior has been observed in some problems arising in gas dynamics. Indeed, Zel'dovich and Raizer [16] have observed such behavior in their studies of entropy changes across shock fronts in gases. Majda [12] has also found such behavior in solutions of a quasilinear equation modeling dynamic combustion in gases.

A number of authors have studied problem (P_1) under various assumptions. O'Malley [14] and Ackerberg and O'Malley [1] have provided a theory for the boundary and shock layer behavior of solutions of problem (P_1) for the case in which f is independent of y and g is a linear function of y. Dorr [6] has extended the results of the linear theory to quasilinear equations of the form

$$\varepsilon y''=x^nF(x,y)y', \qquad -1<x<1,$$

where $n\geqq 1$. Most recently, Howes [8], [9] has provided a theory for the boundary and shock layer behavior of solutions of the full quasilinear problem (P_1).

The only published studies of the spike and nonmonotone transition layer behavior of solutions of problem (P_1) concern the case $f(x,y)\equiv 0$ for all x in $[-1,1]$ and for all y. In particular, O'Malley [15] has studied the autonomous problem

$$\varepsilon y''=g(y), \qquad y(-1,\varepsilon)=A, \quad y(1,\varepsilon)=B,$$

using a phase-plane argument. In his work, O'Malley made the astonishing discovery that this problem has solutions that exhibit spike layer behavior at each rational point in $(-1,1)$. DeSanti [5] has generalized the results of O'Malley to the nonautonomous case, i.e., $g=g(x,y)$, and has given a method for determining the location of the spike layer as well as the height of the spike. We make extensive use of these results in this paper.

To study the asymptotic behavior of solutions of problem (P_1), we use a method based on the theory of differential inequalities. This theory is due originally to Nagumo [13] and Brish [3], but most recently has been described by Jackson [11] and Bernfeld and Lakshmikantham [2]. We describe the method in the next section.

2. Mathematical preliminaries. Throughout the paper we use standard asymptotic terminology. We say that the function $h(x,\varepsilon)$ is $O(\varepsilon^n)$ if $\lim_{\varepsilon\to 0} h(x,\varepsilon)/\varepsilon^n$ exists. In other words, $h(x,\varepsilon)$ is $O(\varepsilon^n)$ if h behaves like ε^n as $\varepsilon\to 0$. We say that the function $h(x,\varepsilon)$ is transcendentally small if $h(x,\varepsilon)$ is $O(\varepsilon^n)$ for every n. Thus a transcendentally small term (abbreviated T.S.T.) behaves like $\exp(-k/\varepsilon^\nu)$ for some $k,\nu>0$ as $\varepsilon\to 0$. Finally, we say that a function $h(x,\varepsilon)$ approaches zero exponentially as $\varepsilon\to 0$ if $|h(x,\varepsilon)|<M\exp(-\eta(x,\varepsilon))$ where M is a positive constant and $\lim_{\varepsilon\to 0}\eta(x,\varepsilon)=+\infty$. In a region in which η is bounded away from zero, a term that decays exponentially to zero as $\varepsilon\to 0$ is also a transcendentally small term.

The basic mathematical tool we use in our study of problem (P$_1$) is a theorem due to Nagumo [13] concerning the boundary value problem

$$(\mathrm{P}_2)\qquad \begin{aligned} &y''=f(x,y)y'+g(x,y), \quad a<x<b,\\ &y(a)=A, \qquad y(b)=B, \end{aligned}$$

and generalizations of this problem.

THEOREM 2.1 (Nagumo [13]). *Let f and g be of class $C^{(1)}$ on $(-1,1)\times R$. Suppose that there exist functions $\alpha=\alpha(x)$ and $\beta=\beta(x)$ of class $C^{(2)}$ on (a,b) such that*

$$\begin{aligned} &\alpha''\geqq f(x,\alpha)\alpha'+g(x,\alpha), \qquad a<x<b,\\ &\alpha(a)\leqq A, \qquad \alpha(b)\leqq B,\\ &\beta''\leqq f(x,\beta)\beta'+g(x,\beta), \qquad a<x<b,\\ &\beta(a)\geqq A, \qquad \beta(b)\geqq B,\\ &\alpha(x)\leqq\beta(x), \qquad a\leqq x\leqq b. \end{aligned}$$

Then there exists a solution $y=y(x)$ of (P$_2$) *such that $\alpha(x)\leqq y(x)\leqq\beta(x)$ for $a\leqq x\leqq b$.*

Theorem 2.1 is valid if α and β are not differentiable at a finite number of points in (a,b), provided that α and β behave appropriately near these points. The behavior required at a point x_0 of nondifferentiability is that the inequalities $D_l\alpha(x_0)\leqq D_r\alpha(x_0)$ and $D_l\beta(x_0)\geqq D_r\beta(x_0)$ are satisfied, where D_l and D_r denote differentiation on the left and right, respectively. It is this stronger version of Nagumo's Theorem, proved in Jackson [11] and in Bernfeld and Lakshmikantham [2], that we use in the following sections.

3. Spike layer theory. In this section we use the method of differential inequalities described in the previous section to deduce the existence of a spike layer solution of the problem

$$(\mathrm{P}_1)\qquad \begin{aligned} &\varepsilon y''=f(x,y)y'+g(x,y),\\ &y(-1,\varepsilon)=A, \qquad y(1,\varepsilon)=B \end{aligned}$$

when f has a certain type of turning point at $x=0$. Our method allows the asymptotic determination of the spike height as well as the construction of $O(\varepsilon)$—approximate solutions that serve as upper and lower bounds on the exact solutions of problem (P$_1$). The principal result is the following.

THEOREM 3.1. *Assume*

(a) *the functions f and g are of class $C^{(1)}$ on $[-1,1]\times R$, and $f(0,y)\equiv 0$ for all y*;

(b) *there exists a function $u_L=u_L(x)$ of class $C^{(2)}$ on $[-1,1]$ satisfying the reduced problem*

$$\begin{aligned} &0=f(x,u)u'+g(x,u),\\ &u(-1)=A, \qquad u(-1)>B; \end{aligned}$$

(c) $f_y(x,u_L(x))u_L'(x)+g_y(x,u_L(x))>K_1>0$ *for some* $K_1>0$ *and for all x in* $(-1,1)$;

(d) $g(0,u_L(0))=0$;

(e) *there exists a number $s\neq u_L(0)$ such that $[s-u_L(0)]g(0,s)<0$, $J_L(s)=0$ and $[s-u_L(0)]J_L(z)>0$ for $u_L(0)<z<s$ or $s<z<u_L(0)$, where $J_L(z)=\int_{u_L(0)}^z g(0,u)\,du$*;

(f) $f_y(1,y)u_L'(1)+g_y(1,y)>K_2$ *for some* $K_2>0$ *and for* $B\leqq y\leqq u_L(1)$;

(g) $f_x(0,y)<0$ *for* $u_L(0)\leqq y\leqq s$ *or* $s\leqq y\leqq u_L(0)$;

(h) $f(1,y)<0$ *for* $B\leqq y\leqq u_L(1)$.

Then, there exists a solution $y=y(x,\varepsilon)$ of problem (P_1) *for ε sufficiently small, say $0<\varepsilon\leqq\varepsilon_1$, that exhibits spike layer behavior at $x=0$. More precisely, $y(x,\varepsilon)\to u_L(x)$ for x in $(-1,1)-\{0\}$ as $\varepsilon\to 0$ and $y(0,\varepsilon)\to s$ as $\varepsilon\to 0$.*

Proof. We consider only the case $u_L(0)<s$. The proof for the case $u_L(0)>s$ is similar.

To prove the theorem, we construct lower and upper solutions α and β, respectively, as in Theorem 2.1. These functions are taken to be straightforward modifications of the solution u_L of the reduced problem.

We define α and β as follows:

$$\alpha(x,\varepsilon)=u_L(x)+v_1(x,\varepsilon)+v_b(x,\varepsilon)-\gamma\varepsilon,$$
$$\beta(x,\varepsilon)=u_L(x)+v_2(x,\varepsilon)+\gamma\varepsilon,$$

where γ is such that $\gamma\min\{K_1,K_2\}>\max_{-1\leqq x\leqq 1}|u_L''(x)|$, and where v_1, v_2 and v_b satisfy, respectively,

$$(\mathrm{C}_1)\quad\begin{cases}\varepsilon v_1''>g(0,u_L(0)+v_1)\text{ in }(-1,1),\\ v_1(0,\varepsilon)=s-u_L(0),\\ v_1>0\text{ for all }x\text{ in }(-1,1),\\ v_1'(0,\varepsilon)=0,\\ v_1'>0\text{ for }x<0,\ v_1'<0\text{ for }x>0,\\ v_1'(x,\varepsilon)=(x/\sqrt{\varepsilon})p(x,\varepsilon),\text{ where }p\to 0\text{ exponentially as }\varepsilon\to 0\text{ for }x\neq 0,\\ v_1\to 0\text{ exponentially as }\varepsilon\to 0\text{ for }x\neq 0,\end{cases}$$

$$(\mathrm{C}_2)\quad\begin{cases}\varepsilon v_2''<g(0,u_L(0)+v_2)\text{ in }(-1,1)-\{0\},\\ v_2(0,\varepsilon)=s-u_L(0),\\ v_2>0\text{ for all }x\text{ in }(-1,1),\\ v_2'>0\text{ for }x<0,\ v_2'<0\text{ for }x>0,\\ v_2\to 0\text{ exponentially as }\varepsilon\to 0\text{ for }x\neq 0,\end{cases}$$

$$(\mathrm{C}_b)\quad\begin{cases}\varepsilon v_b''>f(1,u_L(1)+v_b)v_b'\text{ in }(-1,1),\\ v_b(1,\varepsilon)=B-u_L(1),\\ v_b<0\text{ for all }x\text{ in }(-1,1),\\ v_b'<0\text{ for }x<1,\\ v_b\to 0\text{ exponentially as }\varepsilon\to 0\text{ for }x\neq 1.\end{cases}$$

The existence of a function v_1 satisfying conditions (C_1) has been proved by DeSanti [5] under hypotheses (d) and (e). The existence of a piece-wise smooth function v_2 satisfying conditions (C_2) has been established by Fife [7]. Finally, the existence of a function v_b satisfying conditions (C_b) has been proved by Howes [10] using a result of Coddington and Levinson [4] and hypothesis (h). The functions α and β are sketched in Fig. 3.

We consider first the proof that α is a lower solution on the interval $(-1,1)$. The proof is divided into three stages. First we verify that α is a lower solution for x near 1, then for x near 0, and finally for x bounded away from 0 and 1.

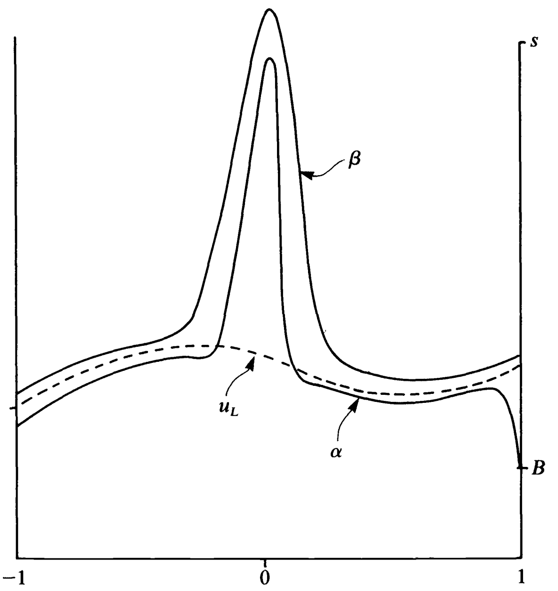


FIG. 3. *Graphs of the lower solution α and the upper solution β.*

Let us first examine the behavior of α in the vicinity of $x=1$. From the Mean Value theorem we have

$$(*)\qquad \begin{aligned} \varepsilon\alpha''-f(x,\alpha)\alpha'-g(x,\alpha)&=\varepsilon v_b''-f(x,u_L(x)+v_b+\gamma\varepsilon)\\ &\quad-[f_y(x,u_L(x)+\theta_1(x))u_L'+g_y(x,u_L(x)+\theta_2(x))]v_b \end{aligned}$$

where $u_L(x)\leqq\theta_1$, $\theta_2<v_b(x,\varepsilon)+\gamma\varepsilon$. From conditions C_b and hypotheses (c) and (f), we see that for x near 1, say in the interval $(1-\rho,1)$, where ρ is a sufficiently small positive number, the right-hand side of equation $(*)$ is positive for ε sufficiently small. Thus, α is a lower solution on the interval $(1-\rho,1)$.

We turn now to an examination of α in the vicinity of $x=0$. Let $S(\varepsilon)$ denote the set

$$\{x\colon x=\delta\varepsilon^{\nu}, \text{for all } \delta \text{ such that } -1\leqq\delta\leqq1, \text{ and for some } \nu, \tfrac{1}{4}<\nu<\tfrac{1}{2}\}.$$

The set $S(\varepsilon)$ is an $O(\varepsilon^{\nu})$-neighborhood of $x=0$. Taylor expanding about $x=0$, and making use of the fact that $v_1'=(x/\sqrt{\varepsilon})p(x,\varepsilon)$, where $p\to0$ exponentially as $\varepsilon\to0$, we have, for x in $S(\varepsilon)$,

$$\varepsilon\alpha''-f(x,\alpha)\alpha'-g(x,\alpha)=\varepsilon v_1''-g(0,u_L(0)+v_1)+O(\varepsilon^{2\nu-1/2}).$$

By virtue of the nature of v_1, the right-hand side of this equation is positive for ε sufficiently small. Thus, α is a lower solution for x in $S(\varepsilon)$.

Let us finally consider the region

$$I^0=(-1,1)-S(\varepsilon)-(1-\rho,1).$$

In this region, v_1 and v_b are both transcendentally small terms (T.S.T.). Thus, for x in I^0, we have

$$\begin{aligned} &\varepsilon\alpha''-f(x,\alpha)\alpha'-g(x,\alpha)\\ &\qquad=\varepsilon[u_L''(x)+\gamma(f_y(x,u_L(x))u_L'(x)+g_y(x,u_L(x))]+\text{T.S.T.}\\ &\qquad\geqq\varepsilon[u_L''+\gamma K_1]. \end{aligned}$$

For ε sufficiently small the right-hand side of this inequality is positive by virtue of the definition of γ. Thus, α is a lower solution for x in the region I°. We conclude that α is a lower solution in the sense of Theorem 2.1 on all of the interval $(-1, 1)$.

We consider now the proof that β is an upper solution on the interval $(-1, 1)$. We note first of all that β is not differentiable at $x=0$. However, since $v_2(0, \varepsilon) = s - u_L(0)$, $v_2' > 0$ for $x<0$ and $v_2'<0$ for $x>0$, and since $v_2 \to 0$ exponentially as $\varepsilon \to 0$ for $x \neq 0$, it follows that

$$\lim_{\varepsilon\to 0} D_l v_2(0, \varepsilon) = +\infty \quad \text{and} \quad \lim_{\varepsilon\to 0} D_r v_2(0, \varepsilon) = -\infty.$$

Thus, for ε sufficiently small we must have

$$D_l\beta(0, \varepsilon) = u_L'(0) + D_l v_2(0, \varepsilon) \geqq D_r\beta(0, \varepsilon) = u_L'(0) + D_r v_2(0, \varepsilon).$$

We see that the behavior of β at $x=0$ is correct in view of the extended version of Theorem 2.1.

Let us now consider the verification that β is an upper solution in the vicinity of $x=0$. Recalling that $\varepsilon v_2'' < g(0, u_L(0)+v_2)$, it follows that $\varepsilon\beta'' < g(x, u_L(x)+v_2+\gamma\varepsilon) \equiv g(x, \beta)$ for ε sufficiently small and for x sufficiently close to zero, say x in the interval $(-d, d)$ for some $d>0$. Thus, we have, for x in $(-d, d)$,

$$\begin{aligned} &\varepsilon\beta'' - f(x, \beta)\beta' - g(x, \beta) \\ (**) \qquad &= \varepsilon\beta'' - g(x, \beta) - f(x, \beta)u_L'(x) - f(x, \beta)v_2' + O(\varepsilon) \\ &= \varepsilon\beta'' - g(x, \beta) - f(x, \beta)v_2' + O(d) \end{aligned}$$

where we have made use of the fact that $f(0, y) = 0$ for all y to assert that the term $f(x, \beta)u_L'(x)$ is $O(d)$ for x in $(-d, d)$. Since $f(x, y)>0$ for $x<0$ and $f(x, y)<0$ for $x>0$, and since $v_2'>0$ for $x<0$ and $v_2'<0$ for $x>0$, it follows that $f(x, \beta)v_2' \geqq 0$ for all x near zero. Hence, the right-hand side of equation (**) is negative for ε and d sufficiently small. We conclude that β is an upper solution on the interval $(-d, d)$.

In the region $(-1, 1)-(-d, d)$, the verification that β is an upper solution proceeds much like the verification that α is a lower solution. In this region, v_2 and v_b are both transcendentally small terms (T.S.T.), so that

$$\begin{aligned} \varepsilon\beta'' - f(x, \beta)\beta' - g(x, \beta) &= \varepsilon[u_L'' - \gamma(f_y(x, u_L)u_L' + g_y(x, u_L))] + \text{T.S.T.} \\ &\leqq \varepsilon[u_L'' - \gamma K_1]. \end{aligned}$$

The right-hand side of this inequality is negative by virtue of the definition of γ. Thus, β is an upper solution for x in the region $(-1, 1)-(-d, d)$. Putting everything together, we see that β is an upper solution in the sense of Theorem 2.1 on all of the interval $(-1, 1)$.

We have thus far shown that α and β satisfy the appropriate differential inequalities of Theorem 2.1. Since v_1 and v_2 are transcendentally small terms for $x=-1$ and $x=1$, and since $v_b>0$, it follows that for ε sufficiently small we have $\alpha(-1, \varepsilon) \leqq A \leqq \beta(1, \varepsilon)$ and $\alpha(1, \varepsilon) \leqq \beta \leqq \beta(1, \varepsilon)$. Furthermore, since $\alpha(0, \varepsilon) < \beta(0, \varepsilon)$ and since v_1 and v_2 both converge to zero exponentially (meaning that v_1 and v_2 are transcendentally small terms for x away from zero), it follows that $\alpha(x, \varepsilon) \leqq \beta(x, \varepsilon)$ on all of $(-1, 1)$. Thus we see that α and β satisfy all the conditions of Theorem 2.1. We conclude, therefore, that there is a solution $y = y(x, \varepsilon)$ of problem (P_1) such that $\alpha(x, \varepsilon) \leqq y(x, \varepsilon) \leqq \beta(x, \varepsilon)$ for all x in $[-1, 1]$. Since α and β both converge to u_L for x in $(-1, 1)-\{0\}$ and to s for $x=0$, the solution $y(x, \varepsilon)$ must behave in the same way. This completes the proof of Theorem 3.1.

Remark 3.1.1. The conclusion of Theorem 3.1 obviously remains valid if f, g, A and B are allowed to depend smoothly on ε.

Remark 3.1.2. The reduced solution u_L of condition (b) could be replaced by the solution u_R of the problem

$$0=f(x,u)u'+g(x,u), \qquad u(1)=B.$$

In this case we would require $u(-1)>A$ and $f(-1,y)>0$ for $A\leqq y\leqq u_R(-1)$. All other conditions of the theorem remain unchanged.

4. Nonmonotone transition layer theory. In this section we again consider the singularly perturbed boundary value problem

$$(\mathrm{P}_1) \qquad \begin{gathered} \varepsilon y''=f(x,y)y'+g(x,y), \\ y(-1,\varepsilon)=A, \qquad y(1,\varepsilon)=B. \end{gathered}$$

Our principal goal is to deduce the existence of a solution of (P_1) that exhibits interior nonmonotone transition layer behavior at the turning point $x=0$. We make extensive use of the results of the previous sections. Our principal result is the following:

THEOREM 4.1. *Assume*

(a) *the functions f and g are of class $C^{(1)}$ on $[-1,1]\times R$, and $f(0,y)\equiv 0$ for all y*;

(b) *there exist functions $u_L=u_L(x)$ and $u_R=u_R(x)$ of class $C^{(2)}$ on $[-1,1]$ satisfying, respectively, the reduced problems*

$$0=f(x,u)u'+g(x,u), \qquad u(-1)=A,$$

$$0=f(x,u)u'+g(x,u), \qquad u(1)=B;$$

(c) *$f_y(x,u)u'+g_y(x,u)>K$ for some $K>0$, for $u=u_L$, u_R, and for all x in $(-1,1)$*;

(d) *$g(0,u_L(0))=0$*;

(e) *$u_L(0)<u_R(0)$ and either*

(i) *there exists a number $s>u_R(0)$ such that $g(0,s)<0$, $J_L(s)=0$, $J_L(z)>0$ for $u_L(0)\leqq z<s$, and $J_R(z)>0$ for $u_R(0)<z\leqq s$, or*

(ii) *there exists a number $s<u_L(0)$ such that $g(0,s)>0$, $J_L(s)=0$, $J_R(z)<0$ for $s<z<u_R(0)$, and $J_L(z)<0$ for $s\leqq z<u_R(0)$, where $J_L(z)=\int_{u_{L(0)}}^{z} g(0,u)\,du$ and $J_R(z)=\int_{u_{R(0)}}^{z} g(0,u)\,du$*;

(f) *$f_x(0,y)<0$ for $u_L(0)\leqq y\leqq s$ or $s\leqq y\leqq u_R(0)$.*

Then, there exists a solution $y=y(x,\varepsilon)$ of (P_1) for ε sufficiently small, say $0<\varepsilon\leqq\varepsilon_2$, that exhibits interior nonmonotone transition layer behavior at $x=0$. More precisely, $y(x,\varepsilon)\to u_L(x)$ for x in $(-1,0)$, $y(x,\varepsilon)\to u_R(x)$ for x in $(0,1)$, and $y(0,\varepsilon)\to s$ as $\varepsilon\to 0$.

Proof. We consider only the case $s>u_R(0)$. The proof for the remaining case is similar.

As in the proof of Theorem 3.1, the proof of this theorem inolves the construction of lower and upper solutions which satisfy the conditions of Theorem 2.1. In this construction, we make extensive use of the lower and upper solutions α and β constructed in the previous section.

We define new lower and upper solutions $\bar{\alpha}$ and $\bar{\beta}$ as follows:

$$\bar{\alpha}(x,\varepsilon)=\begin{cases} \alpha(x,\varepsilon) & \text{for } x \text{ in } [-1,0], \\ \max\{\alpha(x,\varepsilon),u_R(x)-\lambda\varepsilon\} & \text{for } x \text{ in } (0,1]. \end{cases}$$

$$\bar{\beta}(x,\varepsilon)=\begin{cases} \beta(x,\varepsilon) & \text{for } x \text{ in } [-1,0], \\ u_R(x)+\bar{v}(x,\varepsilon)+\lambda\varepsilon & \text{for } x \text{ in } (0,1], \end{cases}$$

where λ is such that $\lambda K \max_{-1 \leqq x \leqq 1} |u_R''(x)|$, and where the function $\bar{v}$ satisfies

$$(\bar{\mathrm{C}}) \qquad \begin{aligned} &\varepsilon \bar{v}'' < g(0, u_R(0) + \bar{v}), \\ &\bar{v} > 0,\ \bar{v}' < 0 \text{ for } x \text{ in } [0, 1], \\ &\bar{v}(0, \varepsilon) = s - u_R(0), \\ &\bar{v} \to 0 \text{ exponentially as } \varepsilon \to 0^+ \text{ for } x \neq 0. \end{aligned}$$

The existence of a function $\bar{v}$ satisfying conditions ($\bar{\mathrm{C}}$) follows from hypotheses (d) and (e) and Lemma 2.1 of Fife [7]. The functions $\bar{\alpha}$ and $\bar{\beta}$ are sketched in Fig. 4.

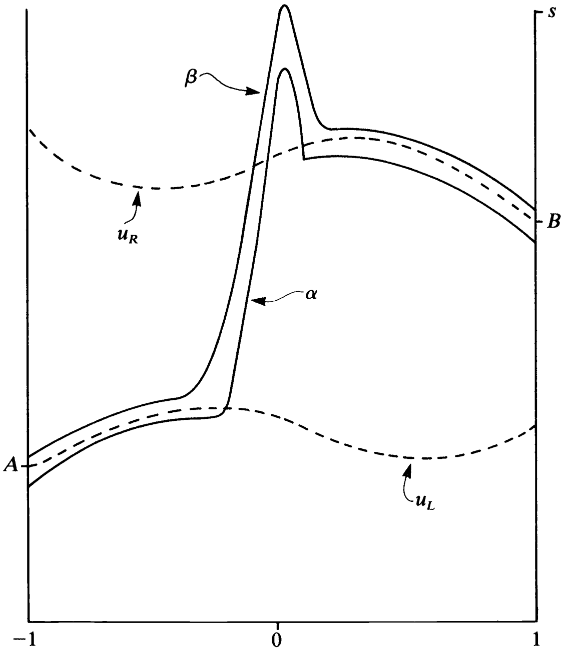


FIG. 4. *Graphs of the lower solution $\bar{\alpha}$ and the upper solution $\bar{\beta}$.*

That $\bar{\alpha}$ and $\bar{\beta}$ are lower and upper solutions, respectively, on the interval $[-1, 0]$ is a consequence of the proof of Theorem 3.1. Moreover, if we simply replace u_L with u_R in the proof of Theorem 3.1, then it is clear that $u_R(x) - \lambda\varepsilon$ is a lower solution on $(0, 1]$ and that $\bar{\beta}(x, \varepsilon) = u_R(x) + \bar{v}(x, \varepsilon) + \lambda\varepsilon$ is a lower solution on $(0, 1]$. At the point of intersection $\bar{x}$ of $\alpha(x, \varepsilon)$ and $u_R(x) - \lambda\varepsilon$, the function $\bar{\alpha}$ is not differentiable. However, since $\lim_{\varepsilon \to 0} v_1'(\bar{x}, \varepsilon) = -\infty$ we have

$$D_l \bar{\alpha}(\bar{x}, \varepsilon) = \alpha'(\bar{x}, \varepsilon) = u_L'(\bar{x}) + v_1'(\bar{x}, \varepsilon) \leqq D_r \bar{\alpha}(\bar{x}, \varepsilon) = u_R'(\bar{x})$$

for ε sufficiently small. Thus, $\bar{\alpha}$ and $\bar{\beta}$ satisfy all the appropriate differential inequalities in the extended version of Theorem 2.1.

From the construction of $\bar{\alpha}$ and $\bar{\beta}$ it is clear that $\bar{\alpha}(-1, \varepsilon) \leqq A \leqq \bar{\beta}(-1, \varepsilon)$ and $\bar{\alpha}(1, \varepsilon) \leqq \beta \leqq \bar{\beta}(1, \varepsilon)$. Moreover, since $\alpha(x, \varepsilon) \leqq \beta(1, \varepsilon)$ and $\bar{v}(x, \varepsilon) > 0$, we have $\bar{\alpha}(x, \varepsilon) \leqq \bar{\beta}(x, \varepsilon)$ for all x in $[-1, 1]$. Thus, $\bar{\alpha}$ and $\bar{\beta}$ satisfy all the conditions of the extended version of Theorem 2.1. We conclude, therefore, that there exists a solution $y(x, \varepsilon)$ of problem (P_1) such that $\bar{\alpha}(x, \varepsilon) \leqq y(x, \varepsilon) \leqq \bar{\beta}(x, \varepsilon)$. Since $\bar{\alpha}$ and $\bar{\beta}$ both converge to u_L for x in $(-1, 0)$, to u_R for x in $(0, 1)$, and to s for $x = 0$ as $\varepsilon \to 0$, the function y must behave in the same way. This completes the proof of Theorem 4.1.

Remark 4.1.1. In condition (e) of Theorem 4.1 we could have $u_L(0) > u_R(0)$. In this case, Theorem 4.1 remains valid with u_L replaced by u_R and J_L by J_R.

Remark 4.1.2. As in Theorem 3.1, Theorem 4.1 remains valid if f, g, A and B are allowed to depend smoothly on ε.

Remark 4.1.3. Majda [12] has considered a time-dependent version of problem (P_1) as a model for dynamic combustion in gases. A simplified version of this combustion model is

$$(P_3) \qquad \mu(\varepsilon)u_t = \varepsilon u_{xx} - f(x, t, u)u_x - g(x, t, u),$$

where ε is a lumped parameter representing the effects of diffusion and heat conduction, and where μ is a measure of combustion speed. This simplified model is appropriate for the case in which the mass fraction of unburnt gas in the system is constant, or is at least a slowly varying function of time.

The nonmonotone interior layer Theorem 4.1 is applicable to problem (P_3) provided that $\mu(\varepsilon) \to 0$ as $\varepsilon \to 0$. Let $\bar{\alpha} = \bar{\alpha}(x, t, \varepsilon)$ and $\bar{\beta} = \bar{\beta}(x, t, \varepsilon)$ be functions analogous to those given in the proof of Theorem 4.1; i.e., suppose that

$$\varepsilon\bar{\alpha}_{xx} > f(x, t, \bar{\alpha})\bar{\alpha}_x + g(x, t, \bar{\alpha}),$$

and

$$\varepsilon\bar{\beta}_{xx} < f(x, t, \bar{\beta})\bar{\beta}_x + g(x, t, \bar{\beta}),$$

where f and g satisfy the conditions of Theorem 4.1 with t considered to be a parameter. Then, because the inequalities are sharp, they hold if the terms $\mu\bar{\alpha}_t$ and $\mu\bar{\beta}_t$ are added to the appropriate sides of the inequalities, provided that ε is sufficiently small. Thus, in agreement with the qualitative theory of Majda, nonmonotone interior layer behavior is possible for solutions of problem (P_3).

5. Examples. In this section we give some examples of the application of Theorems 3.1 and 4.1.

Example 5.1. Consider the problem

$$(E_1) \qquad \begin{aligned} &\varepsilon y'' = -xy' + y(1-y), \qquad x \text{ in } (-1, 1),\\ &y(-1, \varepsilon) = 0, \qquad y(1, \varepsilon) = -1. \end{aligned}$$

The reduced problem is

$$0 = -xu' + u(1-u), \qquad u(-1) = 0.$$

This reduced problem has the solution $u_L(x) = 0$. Now, $f_y(x, u_L)u'_L + g_y(x, u_L) = -u_L^2 - u_L + 1 = 1 > 0$ and $f_y(1, y)u'_L + g_y(1, y) = -y^2 - y + 1 > 0$ for $-1 \leqq y \leqq 0$ since $\min_{-1 \leqq y \leqq 0}[-y^2 - y + 1] = 1$. Moreover, $J_L(z) = z^2(\frac{1}{2} - \frac{1}{3}z)$. The unique positive zero of J_L is $s = \frac{3}{2} > u_L(0) = 0$. We note that $g(0, \frac{3}{2}) < 0$ and $z^2(\frac{1}{2} - \frac{1}{3}z) > 0$ for $0 \leqq z \leqq s$. Finally, we have $f_x(0, y) = -1 < 0$ and $f(1, y) = -1 < 0$. Thus, all the conditions of Theorem 3.1 are satisfied. We conclude that the boundary value problem (E_1) has a solution $y = y(x, \varepsilon)$, for ε sufficiently small, such that $y(x, \varepsilon) \to u_L(x) \equiv 0$ for x in $(-1, 1) - \{0\}$ and $y(0, \varepsilon) \to s = \frac{3}{2}$ as $\varepsilon \to 0$.

Example 5.2. Consider now the problem

$$(E_2) \qquad \begin{aligned} &\varepsilon y'' = -xy' - y(y^2 - 1)(2 - y), \qquad x \text{ in } (-1, 1),\\ &y(-1, \varepsilon) = -1, \qquad y(1, \varepsilon) = 1. \end{aligned}$$

The reduced problems are

$$\begin{aligned} 0 &= -xu' - u(u^2 - 1)(2 - u), \qquad u(-1) = -1,\\ 0 &= -xu' - u(u^2 - 1)(2 - u), \qquad u(1) = 1. \end{aligned}$$

These problems have the solutions $u_L(x)\equiv -1$ and $u_R(x)\equiv 1$ respectively. We have $u_L(0)<u_R(0)$, $f_y(x, u_L)u'_L+g_y(x, u_L)=6>0$ and $f_y(x, u_R)u'_R+g_y(x, u_R)=2>0$. Furthermore, we have $J_L(z)=\frac{1}{60}[-12z^5+15z^4+20z^3-60z^2+53]$ and $J_R(z)=\frac{1}{60}[-12z^5+15z^4+20z^3-60z^2+77]$. The polynomial $J_L(z)$ has a unique zero at a point $s>1$ in the interval $(-1, 2)$. Moreover, $J_R(z)>0$ for $u_R(0)=1<z<s$. Finally, we note that $g(0, s)<0$ since $s>1$ and $f_x(0, y)=-1<0$ for all y. Thus, all conditions of Theorem 4.1 are satisfied. We conclude that (E_2) has a solution $y=y(x, \varepsilon)$, for ε sufficiently small, such that $y(x, \varepsilon)\to u_L(x)=-1$ for x in $(-1, 0)$, $y(x, \varepsilon)\to u_R(x)=1$ for x in $(0, 1)$, and $y(0, \varepsilon)\to s>1$ as $\varepsilon\to 0$, where, again, s is the unique root of the polynomial equation $J_L(z)=0$ in the interval $-1<z<2$.

Acknowledgment. The author would like to thank the referee for pointing out the work of A. Majda.

REFERENCES

[1] R. C. Ackerberg and R. E. O'Malley, Jr., *Boundary layer problems exhibiting resonance*, Stud. Appl. Math., 49 (1970), pp. 277-295.

[2] S. Bernfeld and V. Lakshmikantham, *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1974.

[3] N. I. Brish, *On boundary value problems for the equation $\varepsilon y''=f(x, y, y')$ for small ε*, Dokl. Akad. Nauk SSSR, 95 (1954), pp. 429-432.

[4] E. A. Coddington and N. Levinson, *A boundary value problem for a non-linear differential equation with a small parameter*, Proc. Amer. Math. Soc., 3 (1952), pp. 73-81.

[5] A. J. DeSanti, *Boundary and interior layer behavior of solutions of some singularly perturbed semilinear elliptic boundary value problems*, J. Math. Pures Appl., to appear.

[6] F. W. Dorr, *Some examples of singular perturbation problems with turning points*, this Journal, 1 (1970), pp. 141-146.

[7] P. Fife, *Semilinear elliptic boundary value problems with small parameters*, Arch. Rational Mech. Anal., 52 (1973), pp. 205-232.

[8] F. A. Howes, *Singularly perturbed nonlinear boundary value problems with turning points*, this Journal, 6 (1975), pp. 644-660.

[9] ———, *Singularly perturbed nonlinear boundary value problems with turning points*, II, this Journal, 9 (1978), pp. 250-271.

[10] ———, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203 (1978).

[11] L. K. Jackson, *Subfunctions and second order ordinary differential inequalities*, Adv. in Math., 2 (1968), pp. 307-363.

[12] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70-93.

[13] M. Nagumo, *Über die Differentialgleichung $y''=f(x, y, y')$*, Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861-866.

[14] R. E. O'Malley, Jr., *On boundary value problems for a singularly perturbed differential equation with a turning point*, this Journal, 1 (1970), pp. 479-490.

[15] ———, *Phase-plane solutions to some singular perturbation problems*, J. Math. Anal. Appl., 54 (1976), pp. 449-466.

[16] Ya. B. Zel'dovich and Yu. P. Raizer, *Elements of Gasdynamics and the Classical Theory of Shock Waves*, Academic Press, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

THE EXISTENCE OF BOUNDED SOLUTIONS OF A SEMILINEAR HEAT EQUATION*

WILLIAM C. TROY†

Abstract. We investigate the existence of bounded solutions of $\Delta w - y \cdot \nabla w/2 - (1/(p-1))w + |w|^{p-1}w = 0$, $y \in R^N$, for $N > 2$ and $p > (N+2)/(N-2)$. We show that if $N = 3$ and $6 \leqq p \leqq 12$ then there are infinitely many positive, bounded, radially symmetric solutions $w(r)$, $r = |y|$, such that $\lim_{r\to\infty} w(r) = 0$.

Key words. backward self-similar solutions, radially symmetric solutions, globally bounded solutions

Introduction. We investigate the existence of nonconstant, bounded solutions of the equation

$$\Delta w - y \cdot \nabla w/2 + |w|^{p-1}w - w/(p-1) = 0 \tag{1}$$

in R^N, $N > 2$, where $p > (N+2)/(N-2)$, and $y \cdot \nabla \equiv \sum_{j=1}^N y_j \cdot \partial/\partial y_j$. Equation (1) is derived from the semilinear heat equation

$$u_t - \Delta u - |u|^{p-1}u = 0. \tag{2}$$

Weissler [7] has shown that (2) has solutions which blow up at $(x, t) = (0, 0)$. Giga and Kohn [5] prove that the asymptotic behavior near the blow up time is described by special solutions of (2) called "backward self-similar solutions," i.e., solutions of the form

$$u(x, t) = (-t)^{1/(1-p)} w(y) \tag{3}$$

where $y = x/(-t)^{1/2}$ and $t < 0$. Substitution of (3) into (2) yields (1). Their analysis demonstrates that (1) has no globally bounded solution for $N = 1$ and 2, nor for $N > 2$ and $p \leqq (N+2)/(N-2)$. However, Giga [4] has recently shown that (1) does have radially symmetric, bounded solutions if the term $w/(p-1)$ is replaced by αw, $\alpha > 1/p - 1$, $N > 2$ and $p < (N+2)/(N-2)$. In addition, he shows that if $\alpha \leqq 1/(p-1)$ and $p < (N+2)/(N-2)$ then there are no radially symmetric solutions.

In this paper we consider the parameter range $N > 2$ and $p > (N+2)/(N-2)$, and investigate (1) for the existence of nonconstant, globally bounded solutions. For simplicity we restrict our attention to the case $N = 3$ and therefore $p > 5$. We look for radially symmetric solutions $w = w(r)$, $r = |y|$ so (1) becomes

$$w'' + \left(\frac{2}{r} - \frac{r}{2}\right)w' + |w|^{p-1}w - \frac{w}{(p-1)} = 0. \tag{4}$$

A bounded solution of (4) must satisfy

$$w(0) = \alpha \in R, \qquad w'(0) = 0. \tag{5}$$

We can assume throughout that on each compact interval $[0, L] \subseteq [0, \infty]$, the solution of (4)-(5) exists, is unique and depends continuously on initial values (see, for example, Haraux and Weissler [6, Thm. 1]). Our main result is:

* Received by the editors January 27, 1986; accepted for publication (in revised form) June 3, 1986. This work was supported in part by National Science Foundation grant MCS-83-01085.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

THEOREM. *Let $N=3$ and $6 \leqq p \leqq 12$. There is an unbounded, increasing, positive sequence $\{\alpha_L\}_{L \geqq N}$ such that for each L, if $w(0)=\alpha_L$ and $w'(0)=0$ then $w(r)>0$ for every $r>0$ and $\lim_{r\to\infty} w(r)=0$. Furthermore, $0<w(r)<[(2p-6)/(p-1)^2]^{1/(p-1)}r^{-2/(p-1)}$ for all large r.*

Remarks. 1) It will become clear during the course of our proof that the theorem can be extended to a larger range of parameters than $N=3$ and $6 \leqq p \leqq 12$. However, the details become more complicated. Thus, for the sake of simplicity we restrict our attention to the values given.

2) Results similar to those stated above have recently been given for a combustion model similar to (2), namely

$$u_t - \Delta u - e^u = 0.$$

Solutions of this equation also blow up at $(x,t)=(0,0)$. Again, a similarity form of solution and the assumption of radial symmetry lead to an equation like (4),

$$w'' + \left(\frac{N-1}{r} - \frac{r}{2}\right)w' + e^w - 1 = 0. \tag{6}$$

The appropriate boundary conditions are

$$w(0) \in R, \quad w'(0)=0 \quad \text{and} \quad \lim_{r\to\infty} \frac{w}{2\ln(r)} = -1. \tag{7}$$

For $N=1$ Bebernes and Troy [1] have shown that there are no solutions of (6)-(7). Subsequently, Eberly [2] proved that (6)-(7) has no solution if $N=2$. However, if $N \in (2,9)$, Eberly and Troy [3] proved that the problem (6)-(7) has an infinite number of solutions.

Outline of Proof. The proof of our theorem uses a shooting argument. First, we note that (4) has two particular solutions. One of these is the constant solution $w=\beta^\beta$ where $\beta = 1/(p-1)$. The other, a nonconstant solution, is given by

$$w_0(r) = [(2p-6)/(p-1)^2]^{1/(p-1)} r^{-2/(p-1)}.$$

Define the auxiliary function $h = w - w_0$ where w solves (4)-(5). We prove that if $\alpha - \beta^\beta > 0$ is small then h has at most two zeros before $w=0$. Also, for any given integer $L \geqq 1$, if α is sufficiently large then h has at least $2L+2$ zeros before $w=0$. This leads us to define the set

$$A_{2L} = \{\alpha > \beta^\beta \mid h \text{ has at least } 2L+2 \text{ zeros before } w=0\}. \tag{8}$$

We show that for each $L \geqq 0$, the set A_{2L} is open, nonempty and unbounded above. The remainder of the proof is devoted to showing that there exists $\alpha_L \in (\beta^\beta, \inf A_{2L}]$ such that if $w(0)=\alpha_L$ and $w'(0)=0$ then h has exactly $2L$ zeros and $w(r)>0$ on $(0,\infty)$, with $\lim_{r\to\infty} w(r)=0$.

Proof of Theorem. We assume that $\alpha>0$. Thus, for $r>0$, as long as $w>0$ we observe that w satisfies

$$w'' + \left(\frac{2}{r} - \frac{r}{2}\right)w' + w^p - \frac{w}{(p-1)} = 0 \tag{9}$$

with initial values

$$w(0)=\alpha, \qquad w'(0)=0. \tag{10}$$

The first step of our proof is to show that h has at most two zeros before $w=0$ if $\alpha-\beta^{\beta}>0$ is sufficiently small. This is done in Lemma 3. However, we first need two technical lemmas before proceeding with the proof of Lemma 3.

LEMMA 1. *For each $p>5$, $w''<0$ for $r>0$ as long as $w(r)\leqq w_0(r)$.*

Proof. If there is a first $\bar{r}>0$ for which $w''(\bar{r})=0$ then

$$w'''(\bar{r})\geqq 0. \tag{11}$$

Suppose that $w(\bar{r})<w_0(\bar{r})$. Then (9) and the definition of $w_0(r)$ lead to $w'''(\bar{r})<w'(\bar{r})(2p+2)/(p-1)^2\bar{r}^2<0$, contradicting (11).

Next, recall that $h=w-w_0$.

LEMMA 2. *For each $\alpha>\beta^{\beta}$ there is a first $r_1=r_1(\alpha)>0$ for which $h(r_1)=0$, and*

$$0<r_1(\alpha)<[(2p-6)/(p-1)^2]^{1/2}[(p+1)/(\alpha(p-1))]^{(p-1)/2}. \tag{12}$$

Proof. If there were an $\alpha>\beta^{\beta}$ for which $r_1(\alpha)$ does not exist, then, by Lemma 1, $w''<0$ and $w'<0$ until $w(a)=\beta^{\beta}$ at some $a\in(0,((2p-6)/(p-1))^{1/2})$. But then it follows from (9) that $w''(a)>0$, a contradiction. Thus $r_1(\alpha)$ exists for every $\alpha>\beta^{\beta}$. Further, from Lemma 1, $w''<0$ on $[0,r_1(\alpha)]$ and therefore $w_0'(r_1)<\Delta w/\Delta r=(w_0(r_1)-\alpha)/r_1$. Since $w_0'(r_1)=-2w_0(r_1)/(p-1)r_1$ this becomes $(-2/(p-1))\times(w_0(r_1)/r_1)<(w_0(r_1)-\alpha)/r_1$ and (12) follows.

LEMMA 3. *If $\alpha-\beta^{\beta}>0$ is sufficiently small then h has at most two zeros before $w=0$.*

Proof. The function h satisfies

$$h''=\left(\frac{r}{2}-\frac{2}{r}\right)h'+h\left(\frac{1}{(p-1)}-g(r)\right) \tag{13}$$

where

$$g(r)\equiv w^{p-1}+w^{p-2}w_0+w^{p-3}w_0^2+\cdots+w_0^{p-1}. \tag{14}$$

We observe that

$$\frac{1}{p-1}-pw_0^{(p-1)}>0 \quad \text{for } r>r_p\equiv\left(\frac{p(2p-6)}{p-1}\right)^{1/2}. \tag{15}$$

Since $w\equiv\beta^{\beta}$ satisfies (9) it follows from continuity that if $\alpha-\beta^{\beta}>0$ is sufficiently small then $r_1(\alpha)$ is the only zero of h on $(0,r_p+1)$, and $w(r_p+1)>w_0(r_p+1)$. If h has another zero, $r_2(\alpha)$, on (r_p+1,∞) then $h'(r_2)<0$ and $h''(r_2)<0$. For $r>r_2$, as long as $0<w<w_0$ then $g(r)<pw_0^{p-1}$ and it follows from (13) and (15) that $h''<0$ and $h'<0$ at least until $w\leqq 0$. This completes the proof of Lemma 3.

The following lemma plays a key role in our shooting method.

LEMMA 4. *Let $L\geqq 1$. Then A_{2L} is nonempty, open and unbounded.*

Proof. Continuity and uniqueness imply that A_{2L} is open. To prove that $A_{2L}\neq\varnothing$ we use a comparison method. First, let $v=hre^{-r^2/8}$. Then (13) becomes

$$v''+f(r)v=0 \tag{16}$$

where

$$f(r)\equiv g(r)+\frac{3}{4}-\frac{1}{p-1}-\frac{r^2}{16}. \tag{17}$$

Next, consider the auxiliary equation

$$u''+\frac{.26u}{r^2}=0 \tag{18a}$$

and define

$$R=\min\left((2(p-5)/(p-1))^{1/2},(2p-6)/(p-1)^3)^{1/2},4(\tfrac{3}{4}-1/(p-1))^{1/2}\right).$$

It is well known that (18a) has a solution $u_0(r)$ which has an infinite number of zeros accumulating at $r=0$. Since $\lim_{\alpha\to\infty} r_1(\alpha)=0$, we choose $\alpha^*>\bar{\alpha}$ such that for all $\alpha>\alpha^*$, $u_0(r)$ has at least $2L+4$ zeros in $(r_1(\alpha), R)$. With $\alpha>\alpha^*$, it follows from the Sturm Comparison Theorem (see Simmons [3]) that if $f(r)\geqq .26/r^2$ on $(r_1(\alpha), R)$, then $v(r)$ and $h(r)$ have at least $2L+2$ zeros in $(r_1(\alpha), R)$. Since $\frac{3}{4}-(1/(p-1))-(r^2/16)>0$ for all $p\geqq 5$ and $r\in[0, R]$, it suffices to show that $g(r)\geqq .26/r^2$ on $[r_1(\alpha), R]$ for each $\alpha\geqq\alpha^*$. We note that $w(r)\geqq w_0(r)$ whenever $h(r)\geqq 0$ and therefore $g(r)\geqq 2p(p-3)/(r^2(p-1)^2)\geqq 1/r^2$, $p\geqq 5$. Thus it remains to determine a lower bound on $g(r)$ when $h<0$. For this we make the transformation $w=sw_0$ and $z=\ln(r)$. Then (9) becomes

$$\ddot{s}+\left(\frac{p-5}{p-1}-\frac{e^{2z}}{2}\right)\dot{s}+\frac{(2p-6)}{(p-1)^2}(s^p-s)=0 \tag{18b}$$

where $\dot{s}\equiv ds/dz$. Let $z_1=\ln(r_1)$ and $z_R=\ln(R)$. The definition of R implies that

$$\frac{p-5}{p-1}-\frac{e^{2z}}{2}>0 \quad \forall z\in(z_1, z_R). \tag{19}$$

Further, it follows from Lemma 1 that $w'(r_1)<0$; hence $h'(r_1)\leqq -w_0'(r_1)=2w_0(r_1)/(p-1)r_1$. Therefore

$$\dot{s}(z_1)\leqq\frac{2}{p-1}. \tag{20}$$

Next, we multiply both sides of (18b) by $\dot{s}$, integrate and obtain

$$\frac{(\dot{s})^2}{2}\leqq\frac{2}{(p-1)^2}+\frac{(2p-6)}{(p-1)^2}\left(\frac{1}{p+1}-\frac{1}{2}+\frac{s^2}{2}-\frac{s^{p+1}}{p+1}\right), \tag{21}$$

$z\in[z_1, z_R]$. It follows from (21) that the right-hand side of (21) is negative if

$$6\leqq p\leqq 7 \quad\text{and}\quad s\leqq .21 \tag{22}$$

or

$$7\leqq p\leqq 13.8 \quad\text{and}\quad s\leqq\tfrac{1}{2}. \tag{23}$$

Thus, $s>.2$ on $[z_1, z_R]$ for $6\leqq p\leqq 7$, and $s>.5$ on $[z_1, z_R]$ if $7\leqq p\leqq 13.8$. Consequently, if $6\leqq p\leqq 7$ then it follows that

$$g(r)\geqq\left(\frac{1-.2^p}{.80}\right)\left(\frac{2p-6}{(p-1)^2}\right)\frac{1}{r^2}\geqq\frac{.27}{t^2}\quad\text{for all } r\in[r_1, R].$$

Similarly, from (23), $g(r)\geqq .26/r^2$ for all $r\in[r_1, R]$ if $7\leqq p\leqq 12.6$. Thus, from this analysis and the Sturm Comparison Theorem it follows that $w>0$ on $[r_1, R]$, and h has at least $2L+2$ zeros on $[r_1, R]$.

Remark. It is clear from the proof that we can vastly expand the range of values of p for which the appropriate estimates hold. For the sake of brevity and simplicity we have chosen not to do so.

For each $L\geqq 1$ it follows from Lemmas 3 and 4 that the set A_{2L} is bounded below with

$$\gamma_L\equiv\inf A_{2L}>\beta^\beta. \tag{24}$$

We fix $L \geqq 1$ and consider the solution of (9) with $w(0) = \gamma_L$ and $w'(0) = 0$. If $h = w - w_0$ has exactly $2L$ zeros and $w > 0$ on $(0, \infty)$ with $\lim_{r\to\infty} w(r) = 0$ then $\alpha_L \equiv \gamma_L$ and the proof of the theorem is complete. Otherwise there are several cases to consider:

(i) h has at least $2L+2$ zeros before $w = 0$. Then continuity implies the same is true if $\gamma_L - \alpha > 0$ is sufficiently small, contradicting the definition of γ_L. Therefore h has at most $2L+1$ zeros before $w = 0$;

(ii) h has at most $2L+1$ zeros followed by a finite value of r for which $h = 0$. Again, continuity implies the same is true if $\alpha - \gamma_L > 0$ is sufficiently small, contradicting the definition of γ_L;

(iii) h has less than $2L$ zeros and $w > 0$ on $(0, \infty)$. Continuity implies that if $\alpha - \gamma_L > 0$ is sufficiently small then h has at most $2L-1$ zeros on $(0, r_p+1)$ as well as at least three more, say at a_1, a_2, a_3 on (r_p+1, ∞) before $w = 0$. Thus, either $h'(a_1) < 0$ or $h'(a_2) < 0$. It suffices to consider the case that $h'(a_1) < 0$. Then $h''(a_1) < 0$ since $a_1 > r_p$ and it follows as in the proof of Lemma 3 that $h'' < 0$ for $r > a_2$ until $w = 0$ at some finite value of r. This contradicts the assumption that h has at least $2L+2$ zeros before $w = 0$ for $\alpha - \gamma_L > 0$ sufficiently small. We conclude from (i)-(iii) that h has at least $2L$ zeros and $w > 0$ for every $r > 0$. If h has exactly $2L$ zeros then $0 < w < w_0$ for all $r > r_{2L}$; hence $\lim_{r\to\infty} w(r) = 0$ and the theorem is proved for $\alpha_L = \gamma_L$. It remains to consider the possibility that h has exactly $2L+1$ zeros with $w > 0$ for every $r > 0$. We define the set $B_{2L+1} = \{\hat{\alpha} < \gamma_L |$ if $\hat{\alpha} < \alpha < \gamma_L$ then h has exactly $2L+1$ zeros. If $\gamma_L - \alpha > 0$ is sufficiently small then continuity and the definition of γ_L imply that h has exactly $2L+1$ zeros. Thus, by Lemma 3, B_{2L+1} is bounded below with $\gamma_{2L+1} \equiv \inf B_{2L+1} > \beta^\beta$. We consider the solution with $w(0) = \gamma_{2L+1}$, $w'(0) = 0$. There are several possibilities;

(iv) h has $2L+1$ zeros before $w = 0$, or

(v) h has at most $2L$ zeros followed by a finite value of r where $h = 0$, or

(vi) h has fewer than $2L$ zeros and $w > 0$ on $(0, \infty)$. These three possibilities are eliminated in the same way as in (i)-(iii) above. We conclude that h has exactly $2L$ zeros and $w > 0$ on $0 < r < \infty$. This implies that $0 < w < w_0$ for every $r > r_{2L}$ and $\lim_{r\to\infty} w(r) = 0$. Thus $\alpha_L \equiv \gamma_{2L+1}$ and the theorem is proved.

REFERENCES

[1] J. Bebernes and W. C. Troy, *Nonexistence for the Kassoy Problem*, this Journal, 18 (1987), to appear.

[2] D. Eberly, *Nonexistence for the Kassoy Problem for Dimensions* 1 *and* 2, preprint.

[3] D. Eberly and W. C. Troy, *Existence of logarithmic-type solutions to the Kassoy Problem in dimensions* $2 < N < 10$, J. Differential Equations, 1986, to appear.

[4] Y. Giga, *On elliptic equations related to self-seimilar solutions for nonlinear heat equations*, 1985, preprint.

[5] Y. Giga and R. V. Kohn, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297-319.

[6] A. Haraux and F. B. Weissler, *Non-uniqueness for a semilinear initial value problem*, Indiana Univ. Math. J., 31 (1982), pp. 167-189.

[7] F. Weissler, *Local existence and nonexistence for semilinear parabolic equations in L^p*, Indiana Univ. Math. J., 29 (1980), pp. 79-102.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

006

THE DRAWING AND WHIRLING OF STRINGS: SINGULAR GLOBAL MULTIPARAMETER BIFURCATION PROBLEMS*

STUART S. ANTMAN† AND MICHAEL REEKEN‡

Abstract. This paper treats the steady motion under gravity of both inextensible and elastic strings that are simultaneously whirled and drawn. The motion is governed by a quasilinear system of singular ordinary differential equations that depend on the two parameters ω and γ, which are the rates of whirling and drawing. The singular nonlinear problem is approximated by a sequence of regular problems to which global multiparameter bifurcation theory is applied. It is then shown that the sequence of solution sheets for the approximating problems converges to solution sheets of the actual problem. All these solution sheets, infinite in number, are shown to bifurcate from the ω- and γ-axes, which form the boundaries of the continuous spectra for the problem linearized about the trivial solution. (These linearized problems have no eigenvalues.) Nevertheless, each bifurcating sheet is characterized by a novel and distinctive nodal pattern. The nature of steady shocks in elastic strings is briefly discussed.

Key words. steady motions of strings, global multiparameter bifurcation theory, singular differential equations, nodal properties, steady shocks

AMS(MOS) subject classifications. 33A40, 34B15, 35L67, 47H12, 58E07, 73G99, 73H99, 73K03

1. Introduction. Global studies of the steady whirling of strings began with the work of Kolodner [17]. Such problems have stimulated many advances in global bifurcation theory. (A comprehensive bibliography is given by Alexander, Antman and Deng [3].) Of special relevance for our present investigation is the work of Stuart [28]. His model for a process by which fibers are spun is that of an inextensible string having a fixed configuration in a vertical plane rotating with constant angular speed about the vertical axis. The string is subjected to unusual boundary conditions. Stuart's beautiful analysis represents the first application of the global bifurcation theory of Crandall and Rabinowitz [10], [11] and Rabinowitz [23], [24] to a concrete problem from physics.

Although he described the manufacturing process as one in which fibers are both whirled and drawn, Stuart in fact ignored the longitudinal motion. As we shall show, the presence of both longitudinal and rotational motions results in a Coriolis acceleration, which prevents the fiber from remaining in a rotating plane, as Stuart required. Moreover, shocks can form in an elastic string that is being drawn. Thus the more realistic model we shall present introduces several qualitatively new and interesting phenomena.

The actual formulation of our problem, carried out in § 2, requires care because the material particles forming the fiber under study change with time. The problem has two basic parameters: the speed of drawing γ and the angular rotation speed ω.

In § 3 we work out the full bifurcation analysis for a special elastic string for which the boundary value problem admits a closed-form solution. This example shows that bifurcating sheets of solutions exhibit remarkable nodal properties characteristic

* Received by the editors August 5, 1985; accepted for publication (in revised form) January 7, 1986.

† Department of Mathematics, University of Maryland, College Park, Maryland 20742. This author's work was partially supported by a grant from the National Science Foundation.

‡ Mathematik, Bergische Universität, Gesamthochschule Wuppertal, 5600 Wuppertal 1, West Germany. This author conducted some of his work during visits to the University of Maryland, College Park, Maryland 20742, and Stanford University, Stanford, California 94305, in 1982–1983.

of solution sets for all kinds of strings. Other features of its solutions are not typical, however, as we show in the subsequent analysis.

In § 4 we carry out an analysis of the regularity of solutions, which enables us to reformulate the problem in a more tractable form. In § 5 we give a careful formulation of the equations for inextensible strings. This formulation supports the global bifurcation analysis of § 6, the heart of the paper. Here we employ a delicate regularization procedure to handle the troublesome singularities of the governing equations. We ultimately obtain the remarkable bifurcation diagram of Fig. 6.63. It shows that all bifurcating sheets of solutions bifurcate from the ω and γ axes (which are the boundaries of the continuous spectra for the linearization). Associated with each bifurcating sheet is a distinctive nodal pattern. This nodal pattern obviously cannot be inherited from that of the linearized problem, because the solutions of the linearized problem have no nodal pattern. In § 7 we examine some aspects of the behavior of elastic strings.

We denote ordinary derivatives by primes and partial derivatives by subscripts. Thus the partial derivative of $(\alpha, \beta, t) \mapsto f(\alpha, \beta, t)$ with respect to t at (α, β, t) is denoted $f_t(\alpha, \beta, t)$. The partial derivative of $(\alpha, t) \mapsto f(\alpha, b(t), t)$ with respect to t at (α, t) is denoted $(\partial/\partial t)f(\alpha, b(t), t) = f_\beta(\alpha, b(t), t)b'(t) + f_t(\alpha, b(t), t)$. We occasionally use the summation convention for twice repeated Latin indices ranging over 1, 2, 3.

2. Formulation of the governing equations. Let $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ represent a fixed right-handed orthonormal basis for Euclidean 3-space $\mathbb{E}^3$. Let $a \geqq 0$. We study the whirling motion under gravity of that part of a string lying between $\mathbf{0}$ and $a\mathbf{k}$ when it is being fed through an *inlet* at $\mathbf{0}$ and is being withdrawn through an *outlet* at $a\mathbf{k}$. Gravity is taken to act in either the $\mathbf{k}$ or $-\mathbf{k}$ direction.

Let ζ parametrize an unstressed reference configuration of the string. ζ identifies a typical material point of the string. Let $\xi(\tau)$ and $\eta(\tau)$ be the material points of the string that respectively pass through $\mathbf{0}$ and $a\mathbf{k}$ at time τ. Let $\mathbf{p}(\zeta, t)$ be the position of material point ζ at time t. Then by definition of ξ and η we have

$$(2.1) \qquad \mathbf{p}(\xi(t), t) = \mathbf{0}, \qquad \mathbf{p}(\eta(t), t) = a\mathbf{k}.$$

We assume that $\mathbf{p}(\cdot, t)$ is absolutely continuous so that it has a derivative almost everywhere and has a well defined length. The *stretch* at (ζ, t), which is the local ratio of deformed to reference length at (ζ, t), is

$$(2.2) \qquad \delta(\zeta, t) \equiv |\mathbf{p}_\zeta(\zeta, t)|.$$

The string is *inextensible* if δ is constrained to equal 1, no matter what system of forces acts on the string. The arc length of the curve $\mathbf{p}(\cdot, t)$ between $\mathbf{p}(\alpha, t)$ and $\mathbf{p}(\beta, t)$ is

$$(2.3) \qquad \sigma(\alpha, \beta, t) = \int_\alpha^\beta \delta(\zeta, t)\, d\zeta.$$

From (2.1) we find that the arc length of $\mathbf{p}(\cdot, t)$ between $\mathbf{0}$ and $\mathbf{p}(\xi(\tau), t)$ is $\sigma(\xi(t), \xi(\tau), t)$.

We now seek steady motions of the string in which it is fed in at $\mathbf{0}$ and withdrawn at $a\mathbf{k}$ at a constant rate and in which the configuration of the *active part* of the string, i.e., the part between $\mathbf{0}$ and $a\mathbf{k}$, occupies a rigid space curve that is rotating about the $\mathbf{k}$-axis with a constant angular velocity ω. We now translate these requirements into precise analytic forms. We assume that

$$(2.4) \qquad \xi'(\tau) = \eta'(\tau) = -\gamma < 0$$

so that

$$\xi(\tau)=\xi(0)-\gamma\tau, \qquad \eta(\tau)=\eta(0)-\gamma\tau. \tag{2.5}$$

Thus

$$l\equiv\eta(\tau)-\xi(\tau)=\eta(0)-\xi(0) \tag{2.6}$$

is independent of τ. (The prescription of l for uniform strings, which are defined below, is equivalent to the prescription of the amount of material in the active part of the string.) Let us set

$$s=\gamma(t-\tau)=\xi(\tau)-\xi(t). \tag{2.7}$$

Note that if $\xi(t)\leqq\xi(\tau)\leqq\eta(t)$, then $s\in[0, l]$. We require that the string lie on a rotating rigid space curve $\mathbf{r}(\cdot, t)$, which (without loss of generality) must have the form

$$\mathbf{r}(s, t)=x_m(s)\mathbf{e}_m(t) \quad \text{for } s\in[0, l], \tag{2.8}$$

$$\mathbf{e}_1(t)\equiv\cos\omega t\mathbf{i}+\sin\omega t\mathbf{j}, \quad \mathbf{e}_2(t)\equiv\mathbf{k}\times\mathbf{e}_1(t), \quad \mathbf{e}_3(t)\equiv\mathbf{k}, \tag{2.9}$$

by setting

$$\mathbf{p}(\xi(\tau), t)\equiv\mathbf{p}(\xi(t)+s, t)=\mathbf{r}(s, t)\equiv\mathbf{r}(\gamma(t-\tau), t). \tag{2.10}$$

The absolute continuity of $\mathbf{p}(\cdot, t)$ ensures that of $\mathbf{r}(\cdot, t)$. We are actually requiring that the motion of the string be a travelling wave with s representing the fixed phase.

Under these conditions we have

$$\mathbf{p}_\xi(\xi(\tau), t)=\mathbf{r}_s(s, t), \tag{2.11a}$$

so that (2.2) and (2.10) imply that

$$\delta(\xi(\tau), t)=|\mathbf{r}_s(s, t)|\equiv\nu(s) \tag{2.11b}$$

since $|\mathbf{r}_s(s, t)|$ does not depend explicitly on t. Moreover, the equality of the extremes of (2.10) implies that

$$\mathbf{p}_t(\xi(\tau), t)=\gamma\mathbf{r}_s(s, t)+\omega\mathbf{k}\times\mathbf{r}(s, t), \tag{2.12}$$

$$\mathbf{p}_{tt}(\xi(\tau), t)=\gamma^2\mathbf{r}_{ss}(s, t)+2\omega\gamma\mathbf{k}\times\mathbf{r}_s-\omega^2[x_1(s)\mathbf{e}_1(t)+x_2(s)\mathbf{e}_2(t)] \tag{2.13}$$

whenever $\mathbf{r}(\cdot, t)$ is twice continuously differentiable. Note that the speeds at $\mathbf{0}$ and $a\mathbf{k}$ of the material points occupying them are

$$|\mathbf{p}_t(\xi(t), t)|=\gamma\nu(0), \qquad |\mathbf{p}_t(\eta(t), t)|=\gamma\nu(l). \tag{2.14}$$

These are not necessarily equal. For a uniform string, (2.4) can be interpreted as a requirement that the mass fluxes at $\mathbf{0}$ and $a\mathbf{k}$ be equal. This requirement, rather than the equality of the entrance and exit speeds (given by (2.14)), is essential for steady motions.

We could have alternatively chosen to parametrize $\mathbf{r}$ by the arc length $\sigma(\xi(t), \xi(\tau), t)$ from $\mathbf{0}$ to $\mathbf{p}(\xi(\tau), t)$. We should then have to require that

$$\sigma(\xi(t+\lambda), \xi(\tau+\lambda), t+\lambda)=\sigma(\xi(t), \xi(\tau), t) \quad \forall\lambda. \tag{2.15a}$$

We would therefore obtain from (2.5) and (2.15a) that

$$\sigma_t(\xi(t), \xi(\tau), t)=\gamma[\delta(\xi(\tau), t)-\delta(\xi(t), t)]. \tag{2.15b}$$

In this case the resulting equations would be cluttered with δ's.

Having described the kinematics of the steady motion of our string, we now turn to the mechanics. To ensure that the string admits motion (2.10) solely under the action of gravity, we require that the string be *uniform*, i.e., that its mass density ρA per unit natural length be a positive constant and that its constitutive functions be independent of ζ. Otherwise (2.10) would represent a constraint, which would have to be maintained by artificial (time-varying) constraint forces. Since ρA is constant, equation (2.6) implies that the total mass of the active part of the string is independent of t.

Let $\mathbf{n}(\zeta, t)$ be the resultant contact force exerted by the material of $\{\chi: \chi > \zeta\}$ on that of $\{\chi: \chi \leqq \zeta\}$ at time t. The defining property of a string is that $\mathbf{n}(\zeta, t)$ be tangent to the curve $\mathbf{p}(\cdot, t)$ at $\mathbf{p}(\zeta, t)$, i.e., that $\mathbf{n}$ have the form

$$\mathbf{n}(\zeta, t) = \tilde{n}(\zeta, t)\mathbf{p}_\zeta(\zeta, t)/\delta(\zeta, t). \tag{2.16}$$

$\tilde{n}(\zeta, t)$ is the *tension* at (ζ, t). The string is *in tension* at (ζ, t) if $\tilde{n}(\zeta, t) > 0$ and *in compression* at (ζ, t) if $\tilde{n}(\zeta, t) < 0$.

The string is *elastic* (and *uniform*) if there is a function $(0, \infty) \ni \nu \mapsto N(\nu) \in \mathbb{R}$ such that

$$\tilde{n}(\zeta, t) = N(\delta(\zeta, t)). \tag{2.17}$$

We assume that N is continuously differentiable and that

$$N'(\nu) > 0, \quad N(1) = 0, \quad N(\nu) \to \infty \quad \text{as } \nu \to \infty, \quad N(\nu) \to -\infty \quad \text{as } \nu \to 0. \tag{2.18}$$

For the steady motions we study, we set

$$\tilde{n}(\xi(\tau), t) = \tilde{n}(s + \xi(0) - \gamma t, t) \equiv n(s, t) \tag{2.19}$$

(cf. (2.5), (2.7)). In this case (2.11) implies that (2.17) reduces to

$$n(s, t) = N(\nu(s)) \tag{2.20}$$

so that n is independent of t for elastic strings. We henceforth drop the argument t of n.

The string is *inextensible* if

$$\nu(s) = 1 \quad \forall s \in [0, l] \tag{2.21}$$

no matter what tension field is acting. For an inextensible string, the constraint force n is retained as a fundamental unknown of the problem. (It is the Lagrange multiplier maintaining the constraint of inextensibility.) We shall seek solutions for which n is locally integrable. As part of our definition of steady motion for such strings we require that n be independent of t.

Since we expect that the string might sustain shocks we formulate its equations of motion in the weakest possible form as impulse-momentum laws (cf. [5]). Let g denote the acceleration of gravity. If the only forces acting on the active part of the string executing the motion described above are its weight acting in the $\mathbf{k}$ or $-\mathbf{k}$ direction and the forces acting on its ends, then the impulse-momentum law for each material segment $(\xi(\tau_1), \xi(\tau_2))$ in $(\xi(t), \eta(t))$ and for each time interval (t_1, t_2) is

$$\begin{aligned} &\int_{t_1}^{t_2} \frac{\tilde{n}(\zeta, t)\mathbf{p}_\zeta(\zeta, t)}{\delta(\zeta, t)}\bigg|_{\zeta=\xi(\tau_1)}^{\zeta=\xi(\tau_2)} dt + \varepsilon\rho A g[\xi(\tau_2) - \xi(\tau_1)](t_2 - t_1)\mathbf{k} \\ &\quad = \rho A \int_{\xi(\tau_1)}^{\xi(\tau_2)} \mathbf{p}_t(\zeta, t)\bigg|_{t=t_1}^{t=t_2} d\zeta \end{aligned} \tag{2.22a}$$

where $\varepsilon = \pm 1$. We now use (2.5), (2.7), (2.11), (2.12) to convert (2.22a) to the following form

$$\int_{t_1}^{t_2} \frac{n(\gamma(t-\tau))\mathbf{r}_s(\gamma(t-\tau), t)}{\rho A\nu(\gamma(t-\tau))}\bigg|_{\tau=\tau_1}^{\tau=\tau_2} dt - \varepsilon g\gamma(\tau_2-\tau_1)(t_2-t_1)\mathbf{k}$$

$$\text{(2.22b)} \qquad = \int_{\xi(\tau_1)}^{\xi(\tau_2)} [\gamma\mathbf{r}_s(\zeta-\xi(0)+\gamma t, t)+\omega\mathbf{k}\times\mathbf{r}(\zeta-\xi(0)+\gamma t, t)]\bigg|_{t=t_1}^{t=t_2} d\zeta$$

$$= \int_{\gamma(t_2-\tau_1)}^{\gamma(t_2-\tau_2)} [\gamma\mathbf{r}_s(s, t_2)+\omega\mathbf{k}\times\mathbf{r}(s, t_2)]\, ds$$

$$- \int_{\gamma(t_1-\tau_1)}^{\gamma(t_1-\tau_2)} [\gamma\mathbf{r}_s(s, t_1)+\omega\mathbf{k}\times\mathbf{r}(s, t_1)]\, ds.$$

Since $\mathbf{r}(\cdot, t)$ is absolutely continuous and since $\mathbf{r}(s, \cdot)$ is analytic, we can differentiate (2.22) with respect to t_2 almost everywhere, obtaining

$$\text{(2.23)} \qquad \begin{aligned} &\left[\frac{n(s)}{\rho A\nu(s)}-\gamma^2\right]\mathbf{r}_s(s, t)\bigg|_{s_1}^{s_2} + \varepsilon g(s_2-s_1)k \\ &\qquad = 2\omega\gamma\mathbf{k}\times\mathbf{r}(s, t)\bigg|_{s_1}^{s_2} + \omega^2\int_{s_1}^{s_2} \mathbf{k}\times[\mathbf{k}\times\mathbf{r}(s, t)]\, ds \end{aligned}$$

where we have set $s_2 = \gamma(t_2-\tau_2)$, $s_1 = \gamma(t_2-\tau_1)$ and have replaced t_2 by t.

Let us now set

$$\text{(2.24a)} \qquad m(s, \gamma) = \frac{n(s)}{\rho A\nu(s)}-\gamma^2, \qquad M(\nu, \gamma) = \frac{N(\nu)}{\rho A\nu}-\gamma^2$$

(so that

$$\text{(2.24b)} \qquad m(s, \gamma) = M(\nu(s), \gamma)$$

for elastic strings),

$$\text{(2.25a)} \qquad u(s) = x_1(s)+ix_2(s), \qquad z(s) = x_3(s)$$

where i is the imaginary unit. Thus

$$\text{(2.25b)} \qquad \nu^2 = |u'|^2+(z')^2.$$

Then we can write (2.23) and the boundary conditions corresponding to (2.4) as

$$\text{(2.26)} \qquad m(y, \gamma)u'(y)\bigg|_c^s = 2i\omega\gamma u(y)\bigg|_c^s - \omega^2\int_c^s u(y)\, dy,$$

$$\text{(2.27)} \qquad u(0) = 0 = u(l),$$

$$\text{(2.28)} \qquad m(s, \gamma)z'(s) = \varepsilon g(b-s),$$

where b is a constant of integration,

$$\text{(2.29)} \qquad z(0) = 0, \qquad z(l) = a.$$

Equations (2.26) and (2.28) are to hold for all $[c, s]\subset[0, l]$.

If (2.26)-(2.29) has a solution with u absolutely continuous, then mu' is absolutely continuous and mz' is an affine function of s. We can therefore differentiate (2.26) and (2.28) with respect to s almost everywhere to obtain

$$\text{(2.30)} \qquad (mu')'-2i\omega\gamma u'+\omega^2 u = 0,$$

$$\text{(2.31)} \qquad (mz')'+\varepsilon g = 0.$$

If the configuration of the string is confined to a vertical plane through **k** rotating about **k** with angular velocity ω, then we can always take u to be real. In this case (2.26) and (2.27) imply that $\omega\gamma u=0$. If $\gamma>0$, then either the solution is trivial: $u=0$, or else $\omega=0$. In the latter case (2.26) implies that

$$mu'=\mathscr{C}\text{ (real const.).}\tag{2.32}$$

Now (2.28) and (2.32) imply that $m^2\nu^2=\mathscr{C}^2+g^2(b-s)^2$. We require that $\nu>0$. Thus if m should vanish on $[0,l]$, it could do so only at $s=b$ provided $b\in[0,l]$. In this case (2.32) would imply that $\mathscr{C}=0$ and therefore $u'=0$. Conditions (2.27) would then imply that $u=0$. If m vanishes nowhere, then (2.32) and (2.37) imply that $\mathscr{C}=0$. Thus *if* $\gamma>0$, *then the only planar solution of boundary value problem* (2.26)-(2.29) *is the trivial solution.*

The equations for an elastic string are obtained by supplementing (2.26)-(2.29) with the constitutive equation (2.24b). The equations for an inextensible string are obtained by retaining n or m as a fundamental variable in (2.23) and by setting $\nu=1$, so that

$$(z')^2=1-|u'|^2.\tag{2.33}$$

In principle, we can determine m from (2.28), (2.29), (2.33) in terms of $|u'|^2$. (Note that m does not now depend on the u' or z' through a composition of M with ν.) Solutions for m can be substituted into (2.26) to obtain a system for u alone. We shall discuss the details of this procedure in § 5.

3. Example. If $\nu_0>1$, then a conceivable form for the restriction of N to $[\nu_0,\infty)$ is

$$N(\nu)=\alpha^2\rho A\nu\quad\text{for }\nu\geqq\nu_0\tag{3.1}$$

where α^2 is a positive constant. The substitution of (3.1) into (2.30), (2.27)-(2.29) reduces them to

$$(\alpha^2-\gamma^2)u''-2i\omega\gamma u'+\omega^2u=0,\qquad u(0)=0=u(l),\tag{3.2}$$

$$(\alpha^2-\gamma^2)z'=\varepsilon g(b-s),\qquad z(0)=0,\quad z(l)=a.\tag{3.3}$$

γ^2 cannot equal α^2 if (3.3) is to have a solution. If $\gamma^2\neq\alpha^2$, then the solution of (3.3) is

$$z=\frac{\varepsilon gs(b-s/2)}{\alpha^2-\gamma^2}=\frac{as}{l}+\frac{\varepsilon gs(l-s)}{2(\alpha^2-\gamma^2)}\tag{3.4a}$$

with

$$b=\frac{l}{2}+\frac{\varepsilon(\alpha^2-\gamma^2)a}{gl}.\tag{3.4b}$$

Problem (3.2) has a nontrivial solution if and only if

$$\omega=\frac{k\pi}{\alpha l}(\alpha^2-\gamma^2),\qquad k=\pm1,\pm2,\cdots.\tag{3.5}$$

If l is fixed, (3.5) defines a countable family of parabolic eigencurves in the (ω,γ)-plane. If l is variable, then (3.5) defines a countable family of eigensurfaces in (ω,γ,l)-space. Corresponding to (3.5) are the nontrivial solutions

$$u^k=B\sin\frac{k\pi s}{l}\exp i\left(\frac{k\pi\gamma s}{\alpha l}\right).\tag{3.6}$$

Let us now check that $\nu \geqq \nu_0$. From (3.4) and (3.6) we obtain

$$(3.7) \qquad \nu^2 = \left(\frac{a}{l} + \frac{\varepsilon g(l-2s)}{2(\alpha^2-\gamma^2)}\right)^2 + \left(B\frac{k\pi}{l}\right)^2 \left(\cos^2\left(\frac{k\pi s}{l}\right) + \left(\frac{\gamma}{\alpha}\right)^2 \sin^2\left(\frac{k\pi s}{l}\right)\right).$$

We can ensure that $\nu \geqq \nu_0$ for all B by taking

$$(3.8) \qquad \frac{a}{l} + \frac{\varepsilon g(l-2s)}{2(\alpha^2-\gamma^2)} \geqq \nu_0,$$

i.e., by taking

$$(3.9) \qquad \frac{a}{l} - \frac{gl}{2|\alpha^2-\gamma^2|} \geqq \nu_0 \quad \text{or equivalently } |\gamma^2-\alpha^2| \geqq \frac{gl}{2(a-\nu_0 l)}.$$

Since $\nu_0 > 1$, the first form of (3.9) requires that $a/l > \nu_0 > 1$. This means that the distance a between inlet and outlet exceeds the natural length l of the string. When (3.9) holds, (3.4b) yields

$$(3.10) \qquad \left|b - \frac{l}{2}\right| \geqq \frac{l}{2}\left(\frac{a}{a-\nu_0 l}\right) > \frac{l}{2}.$$

Thus $b \notin [0, l]$ so that (3.3) implies that z is strictly increasing along the length of the curve.

In the space of (ω, γ, u) the solution pairs are represented by cylindrical surfaces above the eigencurves (3.5) with the points at which $\nu \leqq \nu_0$ excluded. (Thus the planes $\gamma = \pm\alpha$ are excluded.) We sketch the curves (3.5) in Fig. 3.11. On the kth sheet $|u^k|$ has exactly $(k-1)$ interior zeros (which are simple), but the number of zeros of x_1^k and x_2^k varies markedly with γ along the sheet. The shape of the string for $k=3$ and $\gamma/\alpha = 6$ is sketched in Fig. 3.12. (The virtues of (3.1) for other "nonlinear" problems for strings were recognized by Keller [16].)

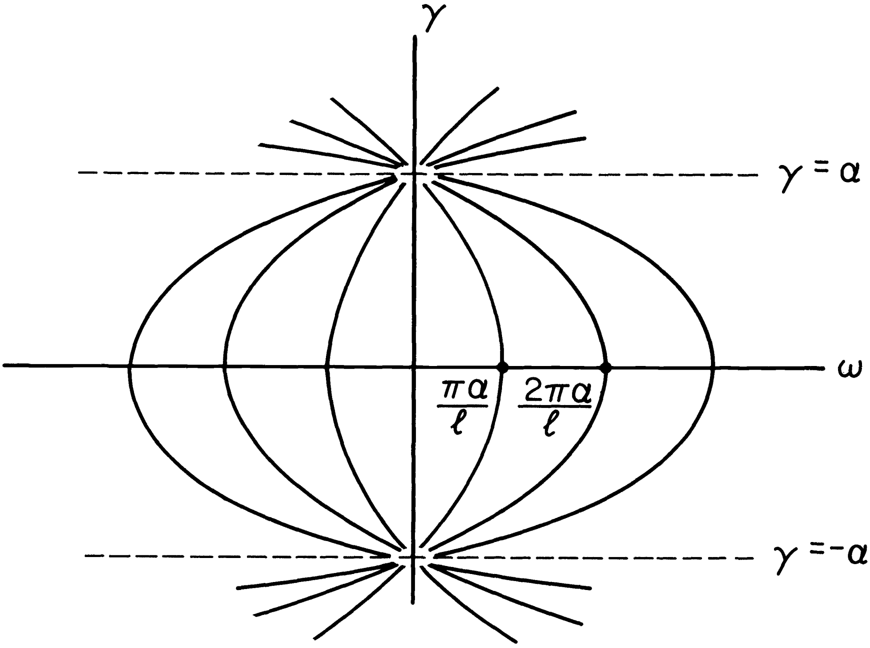


FIG. 3.11

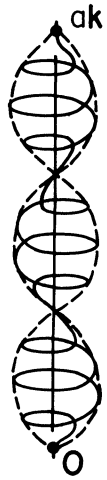


FIG. 3.12

Our goal in the remainder of this paper is to determine which properties of this example are typical of all materials.

4. Regularity of solutions. A solution of the boundary value problem (2.26)-(2.29) is *trivial* if $u=0$ and *nontrivial* otherwise. We seek nontrivial solutions of our boundary value problem for which u and z are absolutely continuous. Throughout this section we suppose that (2.26)-(2.29) has such solutions. Our results apply to them. Below we show that such solutions actually exist.

Equations (2.26) and (2.28) imply that mu' and mz' are absolutely continuous. Thus $|m\nu|$ is absolutely continuous. But this fact does not imply that $m\nu$ itself is continuous. It can be shown that if $m\nu$ is allowed to be discontinuous, then the governing equations admit an uncountable collection of pathological solutions. (Indeed, the methods of Reeken [25], [26] show that for an elastic string there are solutions of (2.26)-(2.29) with $m\nu$ nowhere negative on an arbitrarily prescribed measurable subset of $[0, l]$ and negative on its complement.) Such kinky solutions correspond to certain kinds of travelling shocks. These cannot exist when the string has some bending stiffness. But the presence of bending stiffness would also exclude other effects that are quite reasonable for our model. (For example, a stiff string supported at two nearby points on the same vertical line could not have a folded; pendant equilibrium configuration confined to the line beneath the supports. Such a configuration is a reasonable approximation to the actual one.) To avoid discussions on the physical admissibility of kinky solutions, we choose to preclude them from consideration by seeking solutions for which $m\nu$ is continuous.

Note that the continuity of $|m\nu|$ and the properties of N given in (2.18) ensure that ν has a positive lower bound for any solution of the boundary value problem for elastic strings. For inextensible strings ν is constrained to be 1.

If the string is inextensible, the requirement that $m\nu$ be continuous reduces to the requirement that m be continuous. If we compute m^2 from (2.26), (2.28) we find that m could only vanish on $[0, l]$ at b, provided $b \in [0, l]$. It then follows from (2.26) and (2.28) that u and z are continuously differentiable on $[0, l]\backslash\{b\}$.

The corresponding results for elastic strings are more delicate. Set

$$\gamma_0^2 = \inf_\nu \frac{N'(\nu)}{\rho A}. \tag{4.1}$$

Then (2.18) implies that if $\gamma^2 < \gamma_0^2$, then $\nu \mapsto M(\nu, \gamma)\nu$ is strictly increasing on $(0, \infty)$ and has a continuously differentiable inverse. The continuity of $s \mapsto M(\nu(s), \gamma)\nu(s)$ then implies that of ν when $\gamma^2 < \gamma_0^2$. For $\gamma^2 \geqq \gamma_0^2$ we cannot in general exclude the possibility of discontinuities in ν (density shocks). (γ_0 is the smallest longitudinal wave speed. For special N's, however, a priori estimates could restrict ν to an interval on which $s \mapsto M(\nu(s), \gamma)\nu(s)$ is invertible so that ν would be continuous.) The continuity of ν implies that u' and z' are bounded. It follows from (2.28) that m could have at most one zero on $[0, l]$, at b, provided $b \in [0, l]$. From (2.26) and (2.28) we then find that u and z are continuously differentiable on $[0, l] \backslash \{b\}$ when $\gamma^2 < \gamma_0^2$, or more generally, when the parameters are such that ν is continuous. Henceforth, without further comment, we shall restrict our attention to parameter ranges for elastic strings for which ν is continuous.

Equation (2.26) now implies that $s \mapsto m(s, \gamma)u'(s)$ is continuously differentiable on $[0, l] \backslash \{b\}$ so that (2.30) and (2.31) hold on this set in the classical sense. This fact does not, by itself, imply that u is twice continuously differentiable here. But if we compute $m^2\nu^2$ as before, we find that it is continuously differentiable on $[0, l] \backslash \{b\}$. Thus $m\nu$ is continuously differentiable because it cannot vanish here. For the inextensible string, m itself is continuously differentiable and does not vanish on $[0, l] \backslash \{b\}$. From (2.26) and (2.28) we find that u' and z' are continuously differentiable here. For elastic strings we proceed as before to show that ν is continuously differentiable (where it is continuous). Since ν does not vanish, m itself is continuously differentiable on $[0, l] \backslash \{b\}$ and we again find that u and $z \in C^2([0, l] \backslash \{b\})$. Now suppose that the complex-valued function u has a double zero at a point $s_0 \in [0, l] \backslash \{b\}$. Since m does not vanish on $[0, l] \backslash \{b\}$, the initial value problem for (2.30) on the connected component of $[0, l] \backslash \{b\}$ containing s_0 with initial conditions $u(s_0) = 0 = u'(s_0)$ has the unique solution $u = 0$ on this connected component. Thus we conclude that on each connected component of $[0, l] \backslash \{b\}$ either $u = 0$ or else all the zeros of u are simple. By the same token we find that if u is not the zero function on $[0, b]$ with $b \in (0, l]$, then the only place its zeros on $[0, b]$ could accumulate is at b. An analogous result holds for u nonzero on $[b, l]$.

Let u be a nontrivial solution and let

$$(4.2) \qquad E = \{s \in (0, l) \backslash \{b\}: u(s) \neq 0\}, \qquad F = \begin{cases} [0, b] & \text{if } u(s) = 0 \quad \text{for } s \in [0, b], \\ [b, l] & \text{if } u(s) = 0 \quad \text{for } s \in [b, l]. \end{cases}$$

Since u is continuous, E is open and can therefore be expressed as a countable union of disjoint open intervals. We now define real-valued functions v and ϕ on $E \cup F$ such that

$$(4.3) \qquad u(s) = v(s)e^{i\phi(s)}.$$

We do not require v to be positive. Clearly $v(s) = \pm|u(s)|$. On each component open interval of E, ϕ is determined by (4.3) to within an integral multiple of π. Representation (4.3) does not restrict ϕ on F.

Suppose that $b > 0$ and that u is not the zero function on $[0, b]$. Since the zeros of u could accumulate only at b, we can number the intervals of E starting with that having 0 as an end point. Let $0 \equiv s_0 < s_1 < s_2 < \cdots$ denote the zeros of u that are less than b. On (s_k, s_{k+1}) we set $v = (-1)^k|u|$, $k = 0, 1, \cdots$. Since u is twice continuously differentiable on $[0, b]$, it follows that v is Lipschitz continuous on $[0, b)$ and twice continuously differentiable on each interval (s_k, s_{k+1}). On $[0, s_1)$, we take

$$(4.4) \qquad \phi(s) = \arg u'(0) + \int_0^s \frac{[x_1(\sigma)x_2'(\sigma) - x_2(\sigma)x_1'(\sigma)]}{|u(\sigma)|^2} \, d\sigma$$

where arg $u'(0)$ is well defined by the requirement that it lie in $[0, 2\pi)$ (since $u'(0) \neq 0$). We now show that ϕ is defined at s_1. Since $u \in C^2([0, b))$, we have

$$(4.5) \qquad u(s) = (s - s_1)u'(s_1) + o(s - s_1), \qquad u'(s) = u'(s_1) + o(s - s_1).$$

The use of (4.5) in (4.4) shows not only that the integral converges when the upper limit is s_1, but also that $\phi'(s_1) = 0$. We can extend this argument to all the component open intervals of $[0, b)$. Thus (4.4) defines a continuously differentiable function on $[0, b)$ that is twice continuously differentiable on $E \cap [0, b)$. If $b \geqq l$, then v and ϕ are completely defined. If $b \in (0, l)$, then we use a similar process to define v and ϕ on $(b, l]$ when u is not identically zero here. If u is identically zero here, then we set $v = 0$ here. The remaining cases are treated similarly. We now substitute (4.3) into (2.30) to obtain

$$(4.6a) \qquad (mv')' - mv(\phi')^2 + 2\omega\gamma v\phi' + \omega^2 v = 0,$$

$$(4.6b) \qquad (m\phi')'v + 2m\phi'v' - 2\omega\gamma v' = 0,$$

which hold in the classical sense on each component open interval of E. We multiply (4.6b) by v and integrate the resulting expression to obtain

$$(4.7) \qquad m\phi'v^2 = \omega v^2 + c_n \quad \text{on } E_n$$

where E_n is a component open interval of E and c_n is a constant. But since $v(s) \to 0$ as s approaches an end point of E_n (which is not b), the constant c_n must be 0 (since $m\phi'$ is continuous on $\bar{E} \backslash \{b\}$). Since $v(s) \neq 0$ for $s \in E$, we obtain from (4.7) that

$$(4.8) \qquad m\phi' = \omega\gamma$$

on E. Since $m\phi'$ is continuous on $\bar{E} \backslash \{b\}$ it follows that (4.8) holds on $\bar{E} \backslash \{b\}$, the superposed bar denoting the closure. We now substitute (4.8) into (4.6a) and use (2.24a) to obtain

$$(4.9) \qquad (mv')' + \omega^2 m^{-1}(m + \gamma^2)v = 0$$

on $\bar{E} \backslash \{b\}$. (Note that $m + \gamma^2 = n/\rho A\nu$.)

There are still a few loose ends to tie up. We shall show that if u has a double zero on $[0, l]$, then $u = 0$. This result would imply that (4.9) holds on all of $[0, l]$. To prove these results we first obtain a proposition of some intrinsic interest.

PROPOSITION 4.10. *Let m and ν be continuous and let (u, z) be a nontrivial absolutely continuous solution of* (2.26)-(2.29) *for $\omega\gamma \neq 0$. If $b \in [0, l]$, then $m(b, \gamma) \neq 0$.*

Proof. We restrict our attention to independent variables s lying in $\bar{E} \backslash \{b\}$. From (2.25b) and (4.3) we obtain

$$(4.11) \qquad \nu^2 = (v')^2 + v^2(\phi')^2 + (z')^2.$$

Thus

$$(4.12) \qquad m^2\nu^2 \geqq m^2(v')^2 \geqq 0.$$

From (4.11) and (4.8) we get

$$(4.13) \qquad m^2\nu^2 \geqq \omega^2\gamma^2 v^2 \geqq 0.$$

From (4.11) and (2.28) we get

$$(4.14) \qquad |m(s, \gamma)|^{-1} \leqq g^{-1}|b - s|^{-1}\nu(s) \quad \text{for } s \neq b.$$

Let us assume for contradiction that $m(b, \gamma) = 0$. Since ν is continuous on $[0, l]$, (4.13) implies that

$$(4.15) \qquad v(b) = 0$$

and since mv' is continuous on $\bar{E}\backslash\{b\}$, (4.12) implies that

$$m(s,\gamma)v'(s)\to 0 \quad \text{as } s\to b. \tag{4.16}$$

The integration of (4.9) from b to s and the use of (4.16) yields

$$m(s,\gamma)v'(s)=-\omega^2\int_b^s [1+\gamma^2 m(\sigma,\gamma)^{-1}]v(\sigma)\,d\sigma. \tag{4.17}$$

We now multiply (4.17) by $m(s,\gamma)^{-1}v(s)$, integrate the resulting expression from b to s, and use (4.15) to get

$$\frac{v(s)^2}{2}=-\omega^2\int_b^s \frac{v(\sigma)}{m(\sigma,\gamma)}\int_b^\sigma\left[1+\frac{\gamma^2}{m(\tau,\gamma)}\right]v(\tau)\,d\tau\,d\sigma. \tag{4.18}$$

Next we divide (4.18) by $m(s)^2$ and integrate the resulting expression from b to s to get

$$\frac{1}{2}\int_b^s \frac{v(\sigma)^2}{m(\sigma,\gamma)^2}\,d\sigma=-\omega^2\int_b^s \frac{1}{m(\sigma,\gamma)^2}\int_b^\sigma \frac{v(\tau)}{m(\tau,\gamma)}\int_b^\tau\left[1+\frac{\gamma^2}{m(\chi,\gamma)}\right]v(\chi)\,d\chi\,d\tau\,d\sigma. \tag{4.19}$$

The integral on the left of (4.19) converges despite the singularity of $m(\sigma,\gamma)^{-1}$ at $\sigma=b$ because (4.13) implies that $|vm^{-1}|\leqq \nu|\omega\gamma|^{-1}$. This same inequality also implies that the double integral with respect to χ and τ on the right side of (4.19) behaves like $(\sigma-b)^2$. Thus (4.14) implies that the entire integral on the right side of (4.19) also converges.

Let us set $w=vm^{-1}$ and observe that

$$2\int_b^\sigma w(\tau)\int_b^\tau w(\chi)\,d\chi\,d\tau=\left[\int_b^\sigma w(\tau)\,d\tau\right]^2. \tag{4.20}$$

It then follows from (4.19) that

$$\begin{aligned}
\frac{1}{2}\int_b^s w(\sigma)^2\,d\sigma &\leqq -\omega^2\int_b^s m(\sigma,\gamma)^{-2}\int_b^\sigma w(\tau)\int_b^\tau m(\chi,\gamma)w(\chi)\,d\chi\,d\tau\,d\sigma\\
&\leqq \omega^2\left|\int_b^s m(\sigma,\gamma)^{-2}\int_b^\sigma w(\tau)\,d\tau\left[\int_b^\sigma m(\chi,\gamma)^2\,d\chi\right]^{1/2}\left[\int_b^s w(\chi)^2\,d\chi\right]^{1/2} d\sigma\right| \qquad (4.21)\\
&\leqq \omega^2\left[\int_b^s w(\sigma)^2\,d\sigma\right]M(s,b),
\end{aligned}$$

$$M(s,b)\equiv\left|\int_b^s m(\sigma,\gamma)^{-2}|\sigma-b|^{1/2}\left|\int_b^\sigma m(\chi,\gamma)^2\,d\chi\right|^{1/2} d\sigma\right|, \tag{4.22}$$

the last two inequalities of (4.21) coming from two applications of the Cauchy-Bunyakovskii-Schwarz inequality.

We now show that w must vanish near b by showing that $M(s,b)$ can be made arbitrarily small by taking $|s-b|$ sufficiently small. Thus we must show that the integrand of (4.22) is integrable.

Let us define the set

$$A\equiv\left\{\alpha>0:\left|\int_b^{s_0}|\sigma-b|^{-2\alpha}m(\sigma,\gamma)^2\,d\sigma\right|<\infty\right\} \tag{4.23}$$

where $s_0\neq b$ is some fixed number (in $\bar{E}\backslash\{b\}$). Since m is continuous, $(0,\frac{1}{2})\subset A$. Since ν has a positive lower bound, (4.14) implies that $[\frac{3}{2},\infty)\subset A^c$, the complement of A.

Let $\beta = \sup A$. Thus $\frac{1}{2} \leqq \beta \leqq \frac{3}{2}$ and

$$\left| \int_b^{s_0} |\sigma - b|^{-2(\beta-\varepsilon)} m(\sigma, \gamma)^2 \, d\sigma \right| < \infty, \tag{4.24}$$

$$\left| \int_b^{s_0} |\sigma - b|^{-2(\beta+\varepsilon)} m(\sigma, \gamma)^2 \, d\sigma \right| = \infty \tag{4.25}$$

for all $\varepsilon > 0$. Condition (4.25) implies that the reciprocal of its integrand is bounded:

$$m(s, \gamma)^{-2} \leqq C|s-b|^{-2(\beta+\varepsilon)} \quad \text{for } s \text{ between } b \text{ and } s_0. \tag{4.26}$$

Here and below C denotes a positive constant.

Combining (4.26) with (4.14) we have

$$m(s, \gamma)^{-2} \leqq C|s-b|^{-\mu} \quad \text{where } \mu \equiv \begin{cases} 2 & \text{if } \beta \geqq 1, \\ 2(\beta+\varepsilon) & \text{if } \beta < 1, \end{cases} \tag{4.27}$$

for s between b and s_0. Using (4.27) and the inequality

$$\left| \int_b^{\sigma} m(\chi, \gamma)^2 \, d\chi \right| \leqq |\sigma - b|^{2\alpha} \left| \int_b^{\sigma} \frac{m(\chi, \gamma)^2}{|\chi - b|^{2\alpha}} \, d\chi \right| \tag{4.28}$$

with $\alpha = b - \varepsilon$, we obtain

$$\begin{aligned} M(s, b) &\leqq \left| \int_b^s |\sigma - b|^{(1/2)-\mu+\beta-\varepsilon} \left| \int_b^{\sigma} \frac{m(\chi, \gamma)^2}{|\chi - b|^{2(\beta-\varepsilon)}} \, d\chi \right|^{1/2} d\sigma \right| \\ &\leqq C \int_b^s |\sigma - b|^{(1/2)-\mu+\beta-\varepsilon} \, d\sigma. \end{aligned} \tag{4.29}$$

The definition of μ in (4.27) shows that for sufficiently small ε, the exponent $\frac{1}{2} - \mu + \beta - \varepsilon$, appearing in (4.29), exceeds -1. Thus (4.21) implies that w and therefore v and u must vanish on a neighborhood of b. By the argument preceding (4.2) this fact ensures that $u = 0$ on $[0, l]$. □

By reproducing our earlier arguments we immediately obtain the following.

COROLLARY 4.30. *Let m and ν be continuous. Then absolutely continuous solutions* (u, z) *of* (2.26)-(2.29) *with $\omega\gamma \neq 0$ are continuously differentiable on* $[0, l]$ *and therefore twice continuously differentiable solutions of* (2.30), (2.31).

COROLLARY 4.31. *Let m and ν be continuous. If* (u, z) *is a nontrivial absolutely continuous solution of* (2.26)-(2.29) *with $\omega\gamma \neq 0$, then all the zeros of u are simple. Moreover, ϕ is continuously differentiable on* $[0, l]$ *and v is a twice continuously differentiable solution of* (4.9) *on* $[0, l]$.

5. Inextensible strings. We begin our study of inextensible strings, for which ν is constrained to equal 1, by examining the consequences of (2.28), (2.29), (2.33). Throughout this and the following section we tacitly consider only solutions for which (u, z) is absolutely continuous and m is continuous.

CASE 1. If z' has but one sign on $[0, l]$, which must be positive to accommodate (2.29), then

$$z' = \sqrt{1 - |u'|^2}, \quad \int_0^l \sqrt{1 - |u'|^2} \, ds = a, \quad m = \frac{\varepsilon g(b-s)}{\sqrt{1 - |u'|^2}}. \tag{5.1a, b, c}$$

CASE 2. If $b \in (0, l)$ and $z' > 0$ on $[0, b)$, $z' < 0$ on $(b, l]$, then

$$z'(s) = \operatorname{sign}(b-s)\sqrt{1-|u'|^2}, \tag{5.2a, b, c}$$
$$\int_0^l \operatorname{sign}(b-s)\sqrt{1-|u'|^2}\,ds = a, \qquad m = \frac{\varepsilon g|b-s|}{\sqrt{1-|u'|^2}}.$$

CASE 3. If $b \in (0, l)$ and $z' < 0$ on $[0, b)$, $z' > 0$ on $(b, l]$, then

$$z' = -\operatorname{sign}(b-s)\sqrt{1-|u'|^2}, \tag{5.3a, b, c}$$
$$-\int_0^l \operatorname{sign}(b-s)\sqrt{1-|u'|^2}\,ds = a, \qquad m = \frac{-\varepsilon g|b-s|}{\sqrt{1-|u'|^2}}.$$

For inextensible strings we necessarily restrict our attention to problems in which $(0 \leqq)\ a \leqq l$. If $a = l$, then (5.2b) and (5.3b) cannot be satisfied, while (5.1b) implies that the solution is trivial. Unless there is a statement to the contrary we assume throughout this section that $a < l$. For similar reasons we also assume that $\omega\gamma \neq 0$.

Using (4.3) and (4.8) we obtain from (5.1c), (5.2c), (5.3c) that

$$m^2 = \frac{g^2|b-s|^2 + \omega^2\gamma^2 v^2}{1-(v')^2} \tag{5.4}$$

from which we can get explicit forms for m in each case. Now we convert (4.9) into a system by setting

$$y = \omega\gamma v, \qquad \omega\gamma w = y'm \tag{5.5a, b}$$

so that (5.4) yields

$$\omega^2\gamma^2 w^2 = (y')^2 m^2 = (y')^2[g^2|b-s|^2 + y^2 + w^2]. \tag{5.5c}$$

In Cases 1, 2 and 3, we then have

$$m(s) = \varepsilon \operatorname{sign}(b-s)|m(s)|, \quad m = \varepsilon|m|, \quad m = -\varepsilon|m|, \tag{5.6a}$$

respectively. Here we take

$$|m(s)| = \sqrt{g^2|b-s|^2 + y(s)^2 + w(s)^2}. \tag{5.6b}$$

Then (4.9) reduces to each of the equivalent systems:

$$v' = \frac{w}{m}, \qquad w' = -\omega^2(1+\gamma^2/m)v, \tag{5.7a}$$

$$y' = \frac{\omega\gamma w}{m}, \qquad w' = \frac{-\omega}{\gamma}(1+\gamma^2/m)y. \tag{5.7b}$$

In (5.7b) m has the form (5.6) and, in (5.7a), m has the same form, but with y expressed in terms of v by (5.5a). In (5.7b) m does not depend upon the eigenvalue parameters ω and γ, but the value of $\gamma = 0$ is singular. This formulation is useful for comparison theorems, but the problem in which $\gamma = 0$ cannot be readily treated in this setting. In (5.7a), the right sides of the equations are regular in the parameters, but the dependence is complicated because m also depends on $\omega\gamma$. In view of (4.3), equation (5.7) has the boundary conditions

$$y(0) = 0 = y(l). \tag{5.8}$$

Note that Proposition 4.10 and (5.1c) imply that if there is a nontrivial solution corresponding to Case 1, then $b \notin [0, l]$. We can now consolidate the nontrivial cases of our problem. Let

$$\eta \equiv \begin{cases} 1 & \text{in Case 1 when } b \geqq l \quad \text{and} \quad \text{in Case 2,} \\ -1 & \text{in Case 1 when } b \leqq 0 \quad \text{and} \quad \text{in Case 3.} \end{cases} \tag{5.9}$$

Then (5.6) reduces to

$$m = \varepsilon\eta|m| \tag{5.10}$$

while (5.1a), (5.2a), (5.3a) yield

$$z'(s) = \eta \operatorname{sign}(b-s)\sqrt{1-|u'(s)|^2} = \eta g(b-s)/|m|. \tag{5.11}$$

Let us set

$$r = \sqrt{y^2+w^2}. \tag{5.12}$$

Thus (5.1b), (5.2b), (5.3b) have the form

$$A[b, r] \equiv g\int_0^l [(b-s)/|m|]\, ds = \eta a. \tag{5.13}$$

For $r \neq 0$, we readily find that

$$A[b, r] \to \pm l \quad \text{as } b \to \pm\infty, \qquad A_b[b, r] = g\int_0^l r^2|m|^{-3}\, ds. \tag{5.14}$$

This means that if $a < l$, we can solve (5.18) uniquely for b in terms of r and ηa whenever $r \neq 0$. We denote the solution as $b = \beta[r]$. The classical implicit function theorem in Banach space implies that β is continuously differentiable on $C^0\backslash\{0\}$. For trivial solutions, we find that $2b = 2\beta[0] = l + \eta a$ if $a < l$, but that b is not defined if $a = l$, in which case b can assume any value in $\mathbb{R}\backslash(0, l)$. In this case, however, $z' = 1$ and the configuration is well defined, but

$$m(s) = \varepsilon\eta g|b-s|. \tag{5.15}$$

Thus there is a whole family of compatible tensions when $a = l$. For $a < l$, m is uniquely determined from b by (5.15).

At this stage it is convenient to obtain some estimates for $A[b, r]$. Suppose that R is a positive number with $R \geqq r$. We have the following inequalities:

If $b \leqq 0$,

$$\begin{aligned} -l = A[b, 0] &\leqq A[b, r] \leqq A[b, R] \\ &= \sqrt{b^2+R^2/g^2} - \sqrt{(l-b)^2+R^2/g^2}, \end{aligned} \tag{5.16a}$$

if $b \in (0, l)$,

$$\begin{aligned} \sqrt{b^2+R^2/g^2} - R/g - (l-b) &= g\int_0^b \frac{(b-s)\, ds}{\sqrt{g^2(b-s)^2+R^2}} - g\int_b^l \frac{|b-s|\, ds}{\sqrt{g^2|b-s|^2}} \\ &\leqq g\int_0^l \frac{(b-s)\, ds}{\sqrt{g^2(b-s)^2+r^2}} \\ &\equiv A[b, r] \leqq b - g\int_b^l \frac{|b-s|\, ds}{\sqrt{g^2|b-s|^2+R^2}} \\ &= b + R/g - \sqrt{(l-b)^2+R^2/g^2}, \end{aligned} \tag{5.16b}$$

if $b \geqq l$,

$$\sqrt{b^2+R^2/g^2}-\sqrt{(b-l)^2+R^2/g^2}=A[b,R]\leqq A[b,r]\leqq A[b,0]=l. \tag{5.16c}$$

We consolidate these inequalities into

$$A^-(b,R)\leqq A[b,r]\leqq A^+(b,R) \tag{5.16d}$$

where $A^-(b, R)$ is defined by the leftmost terms of (5.16a, b, c) and $A^+(b, R)$ by the rightmost terms. We illustrate (5.16) in Fig. 5.17.

Equation (5.10) allows us to write (5.7a) as

$$v'=\frac{\varepsilon\eta w}{|m|}, \qquad w'=-\omega^2(1+\varepsilon\eta\gamma^2/|m|)v \tag{5.18}$$

where $|m|$ is given by (5.6b) with $b=\beta[r]$, $r^2=\omega^2\gamma^2v^2+w^2$. Equation (5.18) is equivalent to

$$(|m|v')'+\omega^2[\varepsilon\eta+\gamma^2/|m|]v=0. \tag{5.19}$$

The conversion of (5.7b) is analogous. Indeed, v satisfies (5.19) with the appropriate form of $|m|$.

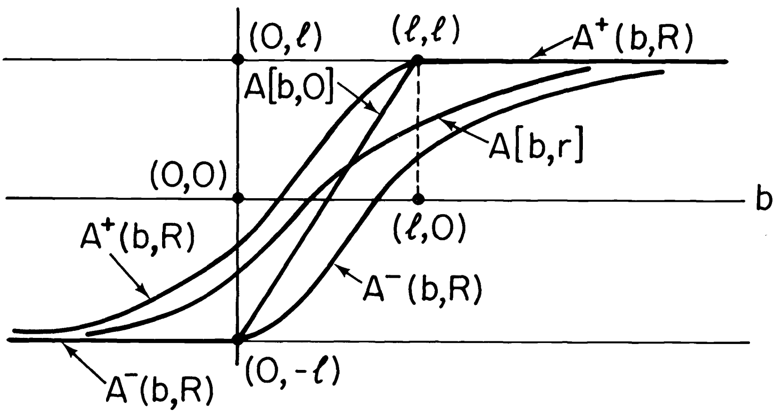


FIG. 5.17

The formal linearization of (5.19) is obtained by replacing $|m|$ in (5.10) with $g|s-(l+\eta a)/2|$. In this case (5.19) reduces to the singular Sturm-Liouville equation

$$g(|s-(l+\eta a)/2|v')'+\omega^2[\varepsilon y+\gamma^2g^{-1}|s-(l+\eta a)/2|^{-1}]v=0. \tag{5.20}$$

Independent solutions of (5.20) on either of the intervals $(0, (l+\eta a)/2)$ or $((l+\eta a)/2, l)$ are the Bessel functions

$$J_{\pm 2i\omega\gamma/g}(2[\varepsilon\eta\omega^2g^{-1}|s-(l+\eta a)/2|]^{1/2}). \tag{5.21}$$

The power series

$$J_\nu(z)=\sum_{m=0}^{\infty}\frac{(-1)^m(z/2)^{\nu+2m}}{m!\Gamma(\nu+m+1)} \tag{5.22}$$

shows that (5.21) is bounded near $s=(l+\eta a)/2$. Indeed, when $\varepsilon\eta=1$, the leading term of (5.21) about this point is

$$\begin{aligned}\Gamma(1\pm 2i\omega\gamma/g)^{-1}\Big\{&\cos\Big\{\frac{\omega\gamma}{g}\ln\Big[\frac{\omega^2}{g}\Big|s-\frac{(l+\eta a)}{2}\Big|\Big]\Big\}\\ &\pm i\sin\Big\{\frac{\omega\gamma}{g}\ln\Big[\frac{\omega^2}{g}\Big|s-\frac{(l+\eta a)}{2}\Big|\Big]\Big\}\Big\},\end{aligned} \tag{5.23}$$

which is discontinuous at $s=(l+\eta a)/2$. Analogous results hold for $\varepsilon\eta=-1$. Since the solutions of (5.20), (5.8) must be real combinations of the functions of (5.21), this problem can have no nontrivial *continuous* solutions. This disturbing observation is a manifestation of the failure of the integral operator corresponding to (5.18), (5.8), say, to be Fréchet differentiable on all of $C^0\times C^0$.

Remarks. (i) That solutions of our boundary value problems should be continuous is a consequence of the underlying physics. The problem (5.20), (5.8) can be given a very complete spectral analysis in weighted L_2-spaces by the Weyl theory and its refinements (cf. Coddington and Levinson [8], Dunford and Schwartz [13] and Naimark [22]). But the results of such an analysis are largely irrelevant for our problem. One inkling of the difficulty we face is that nontrivial solutions of (5.20), (5.8) in the appropriate Hilbert space have no discernible nodal structure.

(ii) Tabulations of the properties of Bessel functions of complex order are sparse (apparently because they were once wrongly deemed to be of little use in applications). The following references, kindly provided to us by F. W. J. Olver, may be consulted: Buckens [7], Luke and Weissman [20], Morgan [21].

(iii) For $\gamma=0$, independent solutions of (5.20) are the Bessel functions J_0 and N_0 of the argument shown in (5.21). Since N_0 is singular when its argument vanishes, its coefficient in the general solution must vanish. The remaining coefficient is generally insufficient to handle the two boundary conditions. A strategy for obtaining physically meaningful results in this case, which relies on the introduction of a or b as a second bifurcation parameter, is described in [3].

6. Global bifurcation theory for inextensible strings. To circumvent the difficulties portended by the singularities of (5.20), we let k denote a positive integer and replace (5.13) with the regularized problem

$$A_k[b,r]\equiv A[b,\sqrt{r^2+k^{-2}}]\equiv g\int_0^l \frac{(b-s)\,ds}{\sqrt{g^2|b-s|^2+r(s)^2+k^{-2}}}=\eta a. \tag{6.1}$$

For each k, considerations like those discussed in the treatment of (5.13) show that (6.1) has a unique solution for b, which we denote by $b=\beta_k[r]$. The functional β_k is continuously Fréchet differentiable on C^0. Next we set

$$m_k(s)=\varepsilon\eta\sqrt{g^2|\beta_k[r]-s|^2+r(s)^2+k^{-2}}. \tag{6.2}$$

We now study the regularized version of (5.18), (5.8)

$$v'=w/m_k,\quad w'=-\omega^2(1+\gamma^2/m_k)v,\quad v(0)=0=v(l) \tag{6.3}$$

and the associated integral equations

$$v(s)=\int_0^s m_k(t)^{-1}w(t)\,dt, \tag{6.4}$$

$$w(s)=\omega^2\Big[\int_0^l m_k(t)^{-1}\,dt\Big]^{-1}\int_0^l m_k(t)^{-1}\int_0^t[1+\gamma^2 m_k(\tau)^{-1}]v(\tau)\,d\tau\,dt \tag{6.5}$$

$$-\omega^2\int_0^s [1+\gamma^2 m_k(t)^{-1}]v(t)\,dt.$$

We abbreviate (6.4), (6.5) as

$$(v, w)=T_k(\omega, \gamma, v, w). \tag{6.6}$$

Since the right sides of (6.4) and (6.5) are innocuous, a simple application of the Ascoli-Arzelà theorem shows that T_k is a compact and continuous mapping from $\mathbb{R}\times[\mathbb{R}\backslash\{0\}]\times C^0\times C^0$ to $C^0\times C^0$. The linearization of (6.6) about the trivial solution, which we denote by

$$(v, w)=L_k(\omega, \gamma)(v, w), \tag{6.7}$$

corresponds to integral equations obtained by replacing m_k in (6.4), (6.5) with

$$m_k^0(s)=\varepsilon\eta\sqrt{g^2|\beta_k[0]-s|+k^{-2}} \tag{6.8}$$

where

$$2\beta_k[0]=2\beta[k^{-1}]=l+\eta a\sqrt{1+4[k^2g^2(l^2-a^2)]^{-1}}. \tag{6.9}$$

These integral equations in turn correspond to the regular Sturm-Liouville problem (5.19), (5.8) with $m(s)$ replaced by (6.8). In the formal limit as $k\to\infty$, these regularized problems reduce to the actual problems of § 5. Our goal is to justify this limit process.

Remark. Had we replaced the radical of (6.1) with $\sqrt{g^2|b-s|^2+r(s)^2+k^{-1}}$ and made analogous changes elsewhere, then we would not have obtained the resulting solution $b=\beta_k[0]$ of the equation $A_k[b, 0]=a$ in closed form. On the other hand, the boundary value problem corresponding to (6.7) could be solved explicitly in terms of Bessel functions of imaginary order.

To carry out the analysis we shall require the following results from the Sturmian theory for the boundary value problem

$$[p(s, \lambda)u']'+q(s, \lambda)u=0, \qquad u(0)=0=u(l). \tag{6.10}$$

The Prüfer transformation

$$u=\rho\sin\theta, \qquad pu'=\rho\cos\theta \tag{6.11}$$

takes (6.10) into the system

$$\rho'=\rho[p(s, \lambda)^{-1}-q(s, \lambda)]\sin\theta\cos\theta, \tag{6.12}$$

$$\theta'=p(s, \lambda)^{-1}\cos^2\theta+q(s, \lambda)\sin^2\theta, \tag{6.13}$$

$$\theta(0)=0, \qquad \theta(l)=(j+1)\pi \tag{6.14a, b}$$

where j is an integer. (Note that (6.13) can be solved in closed form when p and q are independent of s.)

THEOREM 6.15. *Let q be continuous and let p be continuous and positive on* $[0, l]\times[0, \infty)$. *Let the solution of* (6.13), (6.14a) (*known to exist on* $[0, l]$ *for each* $\lambda\in[0, \infty)$) *be denoted by* $\theta(\cdot, \lambda)$. *If*

$$\theta(l, 0)\leqq\pi \quad \text{and} \quad \theta(l, \lambda)\to\infty \quad \text{as } \lambda\to\infty, \tag{6.16}$$

then (6.10) *has a countable infinity of collections* e_j, $j=0, 1, \cdots$, *of eigenvalues with the following properties*: e_j *is a compact subset of* $(0, \infty)$,

$$0<\min e_0<\min e_1<\cdots,$$

$$0<\max e_0<\max e_1<\cdots,$$

$$\min e_j\to\infty \quad \text{as } j\to\infty,$$

the eigenspace corresponding to any eigenvalue is one-dimensional, an eigenfunction corresponding to any eigenvalue in e_j has exactly $j+2$ zeros on $[0, l]$ each of which is simple. If

$$\theta(l, \lambda) \text{ strictly increases to } \infty \quad \text{as } \lambda \to \infty, \tag{6.17}$$

then each e_j consists of but a single eigenvalue λ_j.

Let P and Q be continuous on $[0, l] \times [0, \infty)$ and satisfy

$$p(s, \lambda) \leqq P(s, \lambda), \qquad q(s, \lambda) \geqq Q(s, \lambda). \tag{6.18}$$

Let $\Theta(\cdot, \lambda)$ denote the solution of the modification of (6.13), (6.14a) *obtained by replacing p and q by P and Q. (Then $\Theta(s, \lambda) \leqq \theta(s, \lambda)$.) Let $\theta(\cdot, \lambda)$ and $\Theta(\cdot, \lambda)$ each satisfy* (6.16). *Let $\{E_j\}$ represent the collections of eigenvalues for the problem with P and Q. Then*

$$\min e_j \leqq \min E_j, \quad \max e_j \leqq \max E_j, \quad j = 0, 1, \cdots. \tag{6.19}$$

THEOREM 6.20. *Let $p(s, \lambda) = p(s)$, $q(s, \lambda) = \lambda h(s) - k(s)$ where p, h, k are continuous on $[0, l]$, p is positive on $[0, l]$, and h and k are positive on $(0, l)$. (In this case θ satisfies* (6.16) *and* (6.17).) *Then the eigenvalues λ_j of* (6.10) *are characterized by*

$$\lambda_j = \max_{E_j \in S_j} \min_{u \in E_j^\perp} \frac{\int_0^l [p(v')^2 + kv^2]\, ds}{\int_0^l hv^2\, ds} \tag{6.21}$$

where S_j is the class of all j-dimensional subspaces of the Sobolev space H_0^1 and $E_j^\perp$ is the orthogonal complement of E_j with respect to the inner product $(v_1, v_2) \mapsto \int_0^l hv_1v_2\, ds$. λ_j and its suitably normalized eigenfunction depend analytically on p, h, k in the topology of $C^0 \times C^0 \times C^0$.

Theorem 6.15 is a generalized version of standard results. Its proof relies on the simple observation that $\lambda \in e_j$ if and only if $\theta(b, \lambda) = (j+1)\pi$. (Cf. (6.14b).) Since it is a straightforward exercise (cf. Hille [14], e.g.) to verify that (6.16) and (6.17) hold for our problems, we shall not pause to give the details. Equation (6.21) is derived in [9, Chap. VI], e.g., it has various generalizations. When the hypotheses of Theorem 6.20 hold, the specialization of (6.19) is readily proved by means of (6.21). The last statement of Theorem 6.20 is based on Kato [15, Thm. II.5.16 and § IV.3.57].

We set

$$\chi = ((\omega^2, \gamma^2), (v, w)). \tag{6.22}$$

We define the norm $\|\cdot\|$ on $C^0 \times C^0$ by

$$\|(v, w)\| = \max_{s \in [0,l]} \sqrt{v(s)^2 + w(s)^2}. \tag{6.23}$$

We are now ready to use Theorems 6.15 and 6.20 to study (6.7). Our analysis is based on the observation that if χ satisfies (6.7), then (ω^2, γ^2, v) satisfies (5.19), (5.8) with $m(s)$ replaced by (6.8).

THEOREM 6.24. *Let $\varepsilon\eta = 1$. Then* (6.7) *has a countable infinity of analytic eigencurves $G_j^k \equiv \{(\omega^2, \gamma^2) \in [0, \infty] \times [0, \infty]: \omega^2 = \Omega_j^k(\gamma^2)\}$, $j = 0, 1, 2, \cdots$, such that*

$$0 < \Omega_0^k(\gamma^2) < \Omega_1^k(\gamma^2) < \cdots, \tag{6.25}$$

$$\Omega_j^k(\gamma^2) \to \infty \quad \text{as } j \to \infty, \tag{6.26}$$

$$\Omega_j^k(\gamma^2) = \max_{E_j \in S_j} \min_{v \in E_j^\perp} \frac{\int_0^l |m_k^0|(v')^2\, ds}{\int_0^l (\varepsilon\eta + \gamma^2/|m_k^0|)v^2\, ds}, \qquad \varepsilon\eta = 1, \tag{6.27}$$

$$\Omega_j^k(\gamma^2) \searrow 0 \quad \text{as } \gamma^2 \to \infty, \tag{6.28}$$

$$\Omega_j^k(\gamma^2) \searrow 0 \quad \text{as } k \to \infty. \tag{6.29}$$

(S_j is defined as in (6.21).) *The eigenfunction $V_j^k(\cdot, \gamma^2)$ corresponding to the eigenvalue $(\Omega_j^k(\gamma^2), \gamma^2)$ on G_j^k has exactly $j+2$ zeros on* $[0, l]$, *each of which is simple. V_j^k depends analytically on γ^2.*

Proof. All the statements of this theorem save (6.28) and (6.29) follow immediately from Theorems 6.15 and 6.20. Limits (6.28) and (6.29) are consequences of (6.27), the proof of the latter requiring the observation that if v does not vanish near $(l+\eta a)/2$ then the denominator of the Rayleigh quotient in (6.27) becomes infinite with k. (For k sufficiently large, (6.9) implies that $\beta_k[0]\in(0, l)$.) These results can also be established on the basis of (6.19) by means of a suitable comparison problem. □

In a similar way we obtain the following.

THEOREM 6.30. *Let $\varepsilon\eta=-1$. Then* (6.7) *has a countable infinity of analytic eigencurves $G_j^k\equiv\{(\omega^2, \gamma^2)\in[0,\infty)\times[0,\infty): \gamma^2=\Gamma_j^k(\omega^2)\}$, $j=0, 1, 2, \cdots$ such that*

$$(6.31)\qquad 0<\Gamma_0^k(\omega^2)<\Gamma_1^k(\omega^2)<\cdots,$$

$$(6.32)\qquad \Gamma_j^k(\omega^2)\to\infty \quad as\, j\to\infty,$$

$$(6.33)\qquad \Gamma_j^k(\omega^2)=\max_{E_j\in S_j}\min_{v\in E_j^\perp}\frac{\int_0^l[\omega^{-2}|m_k^0|(v')^2+v^2]\,ds}{\int_0^l|m_k^0|^{-1}v^2\,ds},$$

$$(6.34)\qquad \Gamma_j^k(\omega^2)\nearrow\infty \quad as\, \omega^2\searrow 0,$$

$$(6.35)\qquad \Gamma_j^k(\omega^2)\searrow 0 \quad as\, k\to\infty.$$

Suppose now that $\varepsilon\eta=-1$ and that γ^2 is a fixed positive number. The coefficient of $\omega^2 v$ in the second-order version of (6.7) is

$$(6.36)\qquad -1+\gamma^2[g^2|\beta_k[0]-s|^2+k^{-2}]^{-1/2}.$$

For k sufficiently large there is a neighborhood around $\beta_k[0]$, in fact, around $(l+\eta a)/2$, on which (6.36) is positive. In this case we can conclude from Theorem 6.15 that (6.7) has a countable infinity of eigenvalues $\Omega_j^k(\gamma^2)$ satisfying (6.25) and (6.26). In particular, if

$$(6.37)\qquad \gamma^4\geqq g^2l^2+1,$$

then (6.36) is nonnegative on $[0, l]$, whence it follows that (6.27)-(6.29) also hold. On the other hand, if

$$(6.38)\qquad \gamma^2<g(l+a)/2,$$

then (6.36) is negative on an open subset of $(0, l)$. In this case (6.7) also has a countable infinity of negative eigenvalues ω^2, which we ignore because ω^2 is confined to $[0,\infty)$ on physical grounds.

All these considerations show that the eigencurves of (6.7) have the forms shown in Fig. 6.39. When $\varepsilon\eta=-1$ we take Ω_j^k to be the inverse of Γ_j^k.

The simplicity of the eigenvalues, the nodal properties of the corresponding eigenfunctions, the Fréchet differentiability of T_k, and the compactness and continuity of T_k and L_k enable us to apply the global multiparameter bifurcation theory of Alexander and Antman [2] to conclude the following.

THEOREM 6.40. *Bifurcating from each eigencurve G_j^k of* (6.7) *is a connected family K_j^k of nontrivial solution pairs χ of* (6.6), *each point of which has Lebesgue dimension at least* 2. *The intersection of K_j^k with a plane of the form $\gamma^2=\gamma_0^2$ has at least one of the following properties*: (i) *K_j^k is unbounded in* $(0,\infty)\times(0,\infty)\times C^0\times C^0$, (ii) *there is an $i\neq j$ such that K_j^k connects* $((\Omega_j^k(\gamma^2), \gamma^2), (0, 0))$ *to* $((\Omega_i^k(\gamma^2), \gamma^2), (0, 0))$. *Moreover, if $\chi\in K_j^k$ and if $\|(v, w)\|$ is sufficiently small, then v has exactly $j+2$ zeros on* $[0, l]$, *each of which is simple.*

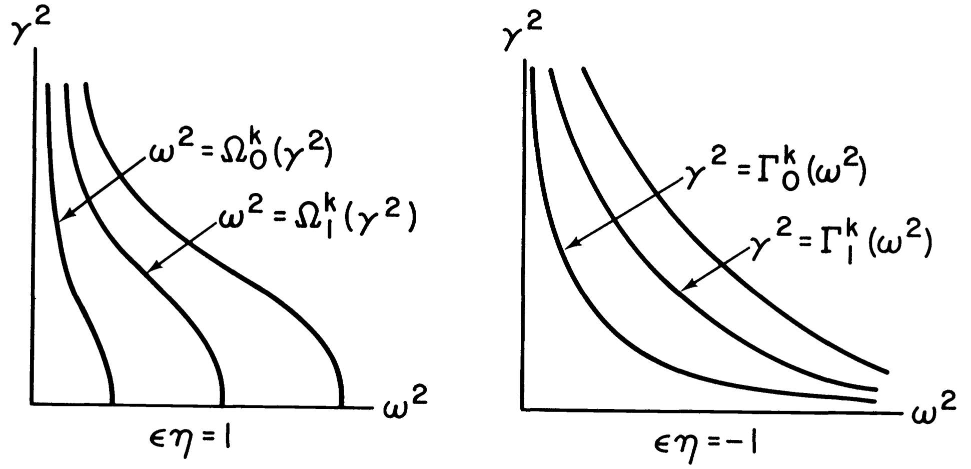


FIG. 6.39. *Eigencurves of* (6.7).

Corollaries 4.30 and 4.31 for our regularized problem and the last statement of Theorem 6.40 allow us to imitate a now standard argument of Crandall and Rabinowitz [10] (cf. e.g. [23], [3]) to deduce the following.

THEOREM 6.41. *Everywhere on K_j^k, v has exactly $j+2$ zeros, each of which is simple. For each γ_0^2, $K_j^k \cap \{\chi: \gamma^2 = \gamma_0^2\}$ is unbounded in* $(0, \infty) \times (0, \infty) \times C^0 \times C^0$, *does not satisfy property* (ii) *of Theorem* 6.40, *and does not meet K_i^k for $i \neq j$.*

We now study the behavior of K_j^k as $k \to \infty$. For this purpose we need estimates, such as those of Proposition 6.51, that are uniform in k. Unless there is a statement to the contrary, we assume that $\omega\gamma \neq 0$. Now (5.18) implies that $|v'| \leqq 1$. From this inequality and from (2.29), or more simply from the underlying geometry, we find that

$$4v^2 \leqq l^2 - a^2. \tag{6.42a}$$

If we use this inequality together with the inequality $|\omega\gamma v/m_k| \leqq 1$ in (6.5) we obtain

$$|w|^2 \leqq 2l^2\omega^2[\omega^2(l^2-a^2)/4+\gamma^2], \tag{6.42b}$$

whence

$$r^2 \leqq \omega^2\{[\gamma^2+2l^2\omega^2](l^2-a^2)/4+2l^2\gamma^2\}. \tag{6.42c}$$

Thus v and w are bounded when ω and γ are bounded.

We now construct related bounds, which are sharper for small $\|(v, w)\|$. Set

$$B(R) = \{(v, w): \|(v, w)\| \leqq R\}. \tag{6.43a}$$

If

$$\chi \in \{[0, \infty) \times [0, \infty) \times B(R)\} \cap K_j^k, \tag{6.43b}$$

then

$$r^2 \leqq (\omega^2\gamma^2+1)R^2 \equiv \delta^2 - 1 \tag{6.44}$$

and

$$A^-(b, \delta) \leqq A_k[b, r] \equiv A[b, \sqrt{r^2+k^{-2}}] \leqq A^+(b, \delta). \tag{6.45}$$

It follows from Fig. 5.17 that

$$\beta^-(\delta, \eta a) \leqq \beta_k[r] \leqq \beta^+(\delta, \eta a) \tag{6.46}$$

where $b = \beta^{\mp}(\delta, \eta a)$ is the solution of $A^{\pm}(b, \delta) = \eta a$. Indeed, we find from (5.16) that

$$\text{(6.47a)} \qquad \begin{aligned} &2\beta^{+}(\delta, \eta a) = (l+\eta a)\frac{[l+\eta a+2\delta/g]}{[l+\eta a+\delta/g]} \in (0, l) \\ &\text{when } 2\eta\delta < g(l^2-a^2)/a, \end{aligned}$$

$$\text{(6.47b)} \qquad \begin{aligned} &2\beta^{+}(\delta, \eta a) = l+\eta a\sqrt{1+4\delta^2 g^{-2}(l^2-a^2)^{-1}} \in [l, \infty) \\ &\text{when } \eta = 1, \qquad 2\delta \geqq g(l^2-a^2)/a. \end{aligned}$$

Now from Fig. 5.17 we find that

$$\text{(6.48a)} \qquad 0 < \beta^{-}(\delta, a) < \beta^{+}(\delta, a),$$

$$\text{(6.48b)} \qquad |\beta^{-}(\delta, a)| \leqq \beta^{+}(\delta, a).$$

Thus (6.46)-(6.48) imply that

$$\text{(6.49)} \qquad 2|\beta_k[r]| \leqq l + a\sqrt{1+\frac{4\delta^2}{g^2(l^2-a^2)}}.$$

Inequalities (6.44) and (6.49) imply that

$$\text{(6.50a)} \qquad |m_k|^2 \leqq 2g^2(|\beta_k|^2+l^2)+r^2 \leqq g^2(3l^2+a^2)+\frac{(3a^2-l^2)}{l^2-a^2}\delta^2-1,$$

$$\text{(6.50b)} \qquad |m_k| \leqq g\sqrt{3l^2+a^2}+\sqrt{\frac{3a^2+l^2}{l^2-a^2}}[\omega\gamma R+\sqrt{R^2+1}] \equiv \omega\gamma R\xi+\psi(R).$$

We now study the case in which $\varepsilon\eta = 1$. Since (6.43) is still in force, χ is a solution pair of the regularized version of (5.19), which we regard as a linear equation for v with coefficients known. We can compare the eigenvalues ω^2 of this nonlinear equation with those of the linear equation obtained by replacing the coefficient $|m_k|$ of v' with the right side of (6.50b) and replacing the coefficient $\omega^2[1+\gamma^2/|m_k|]$ with ω^2. Since the resulting comparison problem has constant coefficients, we find from Theorem 6.15 the following result.

PROPOSITION 6.51. *Let* $\varepsilon\eta = 1$ *and let* (6.44) *hold. Then*

$$\text{(6.52a)} \qquad 0 < \omega_j^2 \leqq \Omega_j^{+}(\gamma^2, R)$$

where

$$\text{(6.52b)} \qquad 2l^2\sqrt{\Omega_j^{+}(\gamma^2, R)} \equiv j^2\pi^2\xi\gamma R+[(j^2\pi^2\xi\gamma R)^2+4l^2j^2\pi^2\psi(R)]^{1/2}.$$

(The comparison problem has two eigenvalues ω for each j. We called the larger $\sqrt{(\Omega_j^{+}(\gamma^2, R))}^{1/2}$ because its square is also the larger, in consonance with Theorem 6.15.)

From Theorem 6.41, inequalities (6.42), and Proposition 6.51, we deduce the following.

THEOREM 6.53. *Let* $\varepsilon\eta = 1$ *and let* γ^2 *be fixed. Then* w *is unbounded and* ω^2 *is positive and unbounded on* K_j^k.

We now fix j, fix a number $\Gamma > 0$, and set

$$\text{(6.54)} \qquad \Delta_j(\Gamma) \equiv \{\chi \in [0, \infty) \times [0, \Gamma] \times C^0 \times C^0 \colon \omega^2 \leqq \Omega_j^{+}(\gamma^2, \|(v, w)\|)\} \cup \{\infty\}.$$

We define the topology of Δ_j (which is like a one-point compactification) by taking a neighborhood basis of ∞ to be

$$\text{(6.55)} \qquad \{\chi \colon \|(v, w)\| \geqq h, h = 1, 2, \cdots\}.$$

Let

$$H_j(\Gamma) \equiv \{\chi: v = 0 = w\} \cap \Delta_j(\Gamma), \tag{6.56}$$

$$\bar{K}_j^k(\Gamma) \equiv [K_j^k \cup H_j(\Gamma) \cup \{\infty\}] \cap \Delta_j(\Gamma). \tag{6.57}$$

Then $\bar{K}_j^k(\Gamma)$ is compact in $\Delta_j(\Gamma)$ and Theorem 6.53 says that $H_j(\Gamma)$ is not separated from $\{\infty\}$ in $\bar{K}_j^k(\Gamma)$. We define

$$K_j(\Gamma) \equiv \{\chi \in \Delta_j(\Gamma): \chi \text{ satisfies (6.6); } v \text{ has exactly } j+2 \text{ zeros on } [0, l], \text{ which are simple}\}, \tag{6.58}$$

$$\bar{K}_j(\Gamma) \equiv K_j(\Gamma) \cup H_j(\Gamma) \cup \{\infty\}. \tag{6.59}$$

It is conceivable that $K_j(\Gamma)$ is empty. Our basic result, which among other things asserts that it is not, is the following.

THEOREM 6.60. *Let $\varepsilon\eta = 1$. (i) $H_j(\Gamma)$ and $\{\infty\}$ are not separated in $\bar{K}_j(\Gamma)$. Thus $K_j(\Gamma)$ contains a connected subset $C_j(\Gamma)$ bifurcating from the jth eigencurve $\{((\omega^2, \gamma^2), (0, 0)): \omega^2 = \Omega_j(\gamma^2), \gamma^2 \in [0, \Gamma]\}$ and containing solution pairs with $\|w\|$ and ω^2 unbounded. (ii) $K_j(\Gamma) \cap K_i(\Gamma) = \varnothing$ for $i \neq j$. (iii) $C_j(\Gamma)$ has Lebesgue dimension at least 2 at each of its points.*

Proof. We hold j fixed. Property (ii) is a consequence of Corollary 4.31. To prove property (i) we appeal to Alexander's [1] generalization and unification of connectivity results of the sort treated by Kuratowski [18, Chap. V] and Whyburn [30, Chap. I]. The theory of Alexander shows that the following two properties suffice to establish property (i): (a) Let $\{k\}$ be any subsequence of the positive integers. If $\chi_j^k \in \bar{K}_j^k(\Gamma)$, then $\{\chi_j^k\}$ has a subsequence converging in the topology of $\Delta_j(\Gamma)$. (b) If $\chi_j^k \in \bar{K}_j^k(\Gamma)$ and if $\chi_j^k \to \chi_j$ in $\Delta_j(\Gamma)$ as $k \to \infty$, then $\chi_j \in \bar{K}_j(\Gamma)$. We first prove (a). If an infinite number of the points χ_j^k equal ∞, or if $\|w_j^k\|$ is unbounded, then $\{\chi_j^k\}$ has a subsequence converging to ∞ and (a) holds. If an infinite number of χ_j^k lie in $H_j(\Gamma)$, then the Bolzano-Weierstrass theorem ensures that (a) holds. Otherwise we may suppose that there is an $R > 0$ such that $(0, 0) \neq (v, w)_j^k \in B(R)$. Since elements of $K_j^k(\Gamma) \cap [\mathbb{R}^2 \times B(R)]$ are nontrivial solution pairs χ_j^k of (6.6) with $(\gamma^2)_j^k \in [0, \Gamma]$ and $(\omega^2)_j^k$ satisfying (6.52a), it suffices to show that $\bigcup_k T_k(\Delta_j(\Gamma) \cap \{\mathbb{R}^2 \times B(R)\}) \equiv D(\Gamma, R)$ is precompact. But the inequalities $|wm_k^{-1}|, |ym_k^{-1}| \leqq 1$ allow us to deduce from (6.4), (6.5) that $D(\Gamma, R)$ is uniformly bounded and equicontinuous, and, by virtue of the Ascoli-Arzelà theorem, is precompact.

We now prove (b), assuming that its hypotheses hold. As before, we can restrict our attention to the case that $R^{-1} \leqq \|w_j^k\| \leqq R$ with $R > 1$. Since χ_j^k satisfies (6.6) (i.e., (6.4) and (6.5)), we let $k \to \infty$. The Lebesgue Dominated Convergence theorem allows us to interchange the order of limit and integration to show that χ_j is a nontrivial solution pair of the limiting form of (6.6), which is equivalent to the boundary value problem posed in § 5. We need only show that v_j has exactly $j+2$ zeros on $[0, l]$, which are simple. If not, $(v, w)_j$ would be approximated in $C^0 \times C^0$ by functions $(v, w)_j^k$ with this nodal pattern. Thus $(v, w)_j$ would vanish at some point in $[0, l]$, whence $(v, w)_j = 0$ by Corollary 4.31, a contradiction. Thus (b) also holds, so that property (i) holds.

To prove property (iii) we observe that the statement that each point of $K_j^k(\Gamma)$ has Lebesgue dimension at least 2 is proved by computing Čech cohomology (cf. [2]). But Čech cohomology is continuous under the limit processes we have just carried out (cf. [1]). Thus the dimensionality properties of $K_j(\Gamma)$ are inherited from those of $K_j^k(\Gamma)$. □

We now study the behavior of K_j or C_j as $\|(v, w)\| \to 0$. We assume that (ω^2, γ^2) lies in a compact subset Σ of $[0, \infty) \times (0, \infty)$. Then (5.6) implies that

$$|m(s)| \leqq \sqrt{g^2|\beta[r] - s|^2 + \zeta^2\|r\|^2} \quad \text{where } \zeta^2 = \max\{\omega^2\gamma^2 + 1: (\omega^2, \gamma^2) \in \Sigma\}. \tag{6.61}$$

Now the analysis supporting Fig. 5.17 shows that $\beta[r] \in (0, l)$ if $\|r\|$ is sufficiently small. Thus we can imitate the proof leading to (6.29) to obtain

THEOREM 6.62 *Let $\varepsilon\eta = 1$. Let $\chi_j \in K_j \cap [\Sigma \times \{(v, w) \colon \|(v, w)\| \leqq \rho\}]$. Then $\omega_j^2 \to 0$ as $\rho \to 0$.*

We do not pause to refine our picture of C_j by the use of Theorems 6.15 and 6.20 buttressed by sharper estimates. We illustrate the appearance of C_j in Figs. 6.63 and 6.64. For fixed $\gamma^2 > 0$, a countable infinity of nontrivial solution branches bifurcate from $\omega^2 = 0 = (v, w)$. $\omega^2 = 0$ is the boundary of the continuous spectrum of the linearized problem. All the sheets of solutions we have treated bifurcate from the lines $\omega^2 = 0$ and $\gamma^2 = 0$.

The treatment of the problem for $\varepsilon\eta = -1$ is much easier than that for $\varepsilon\eta = 1$, paradoxically because the singular behavior as $\gamma^2 \to 0$ of the former is worse than that of the latter. One manifestation of this singular behavior is that the eigencurves of the regularized problem with $\varepsilon\eta = -1$, which are shown in Fig. 6.39, do not intersect the ω^2-axis. The presence of such intersections for $\varepsilon\eta = 1$ meant that we had to treat the behavior for small γ^2 with great care in the analysis beginning with (6.42). Since no such phenomena occur for $\varepsilon\eta = -1$, no such care is required. Consequently we can formulate the problem for $\varepsilon\eta = -1$ in terms of the variables y and w (cf. (5.5), (5.7)).

The integral equations for (y, w) corresponding to (6.4)-(6.6) are singular for $\gamma^2 = 0$, as is evident from (5.7). We set

$$\tilde{\chi} \equiv ((\omega^2, \gamma^2), (y, w)), \tag{6.65}$$

$$\tilde{K}_j^k = \{\tilde{\chi} \colon \chi \in K_j^k\} \tag{6.66}$$

where y is related to v by (5.5a). Then clearly Theorem 6.40 holds with χ and K_j^k

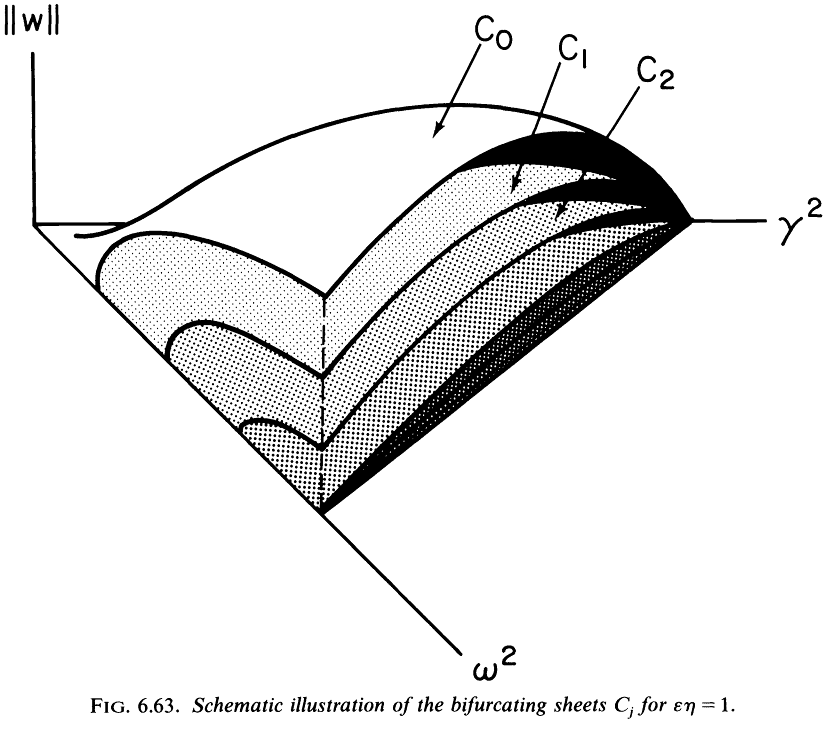


FIG. 6.63. *Schematic illustration of the bifurcating sheets C_j for $\varepsilon\eta = 1$.*

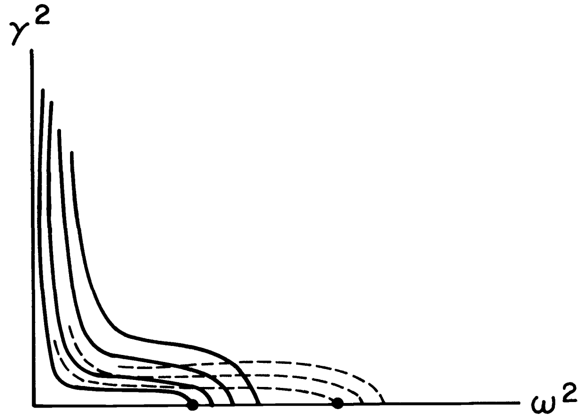


FIG. 6.64. *Schematic diagram of the level surfaces* $C_j \cap \{\chi: \|(v, w)\| = \rho\}$ *for different values of* ρ. *The solid lines correspond to* C_0 *and the dashed lines to* C_1.

replaced with $\tilde{\chi}$ and $\tilde{K}_j^k$ and with (6.6) interpreted as the equivalent equation for $\tilde{\chi}$. Theorem 6.41 holds with these substitutions together with the replacement of v by y.

Next we set

$$\tilde{B}(R) = \{(y, w): \|(y, w)\| \leqq R\}. \tag{6.67}$$

If

$$\tilde{\chi} \in \{[0, \infty) \times (0, \infty) \times \tilde{B}(R)\} \cap \tilde{K}_j^k, \tag{6.68}$$

then

$$r^2 \leqq R^2, \tag{6.69}$$

which is a great simplication over (6.44). It follows that (6.45)-(6.50) hold with

$$\delta^2 = R^2 + 1. \tag{6.70}$$

In analogy with Proposition 6.51 we have the following.

PROPOSITION 6.71. *Let* $\varepsilon\eta = -1$ *and let* (6.68) *hold. Then*

$$0 < \gamma^2 \leqq \Gamma_j^+(\omega^2, R) \equiv |\mu| + \frac{\mu^2}{\omega^2}\left(\frac{j\pi}{l}\right)^2 \tag{6.72}$$

where $\mu^2 \equiv 10g^2l^2 + \alpha^2(R^2+1)$.

(The estimates leading to (6.72) are painless.)

We now fix j, fix a number $\Omega > 1$ and set

$$\tilde{\Delta}_j(\Omega) \equiv \{\tilde{\chi}: \omega^2 \in [\Omega^{-1}, \Omega],\ \gamma^2 \leqq \Gamma_j^+(\omega^2, \|(y, w)\|)\} \cup \{\infty\}, \tag{6.73}$$

defining its topology just as we did for $\Delta_j(\Gamma)$. We set

$$\begin{aligned}\tilde{K}_j(\Omega) \equiv \{\tilde{\chi} \in \tilde{\Delta}_j(\Omega)&: \tilde{\chi} \text{ satisfies its equivalent version of (6.6)},\\ &y \text{ has exactly } j+2 \text{ zeros on } [0, l], \text{ which are simple.}\}\end{aligned} \tag{6.74}$$

Then we can duplicate the proof of Theorem 6.60 to obtain the following.

THEOREM 6.75. *Let* $\varepsilon\eta = -1$. *Then* $\tilde{K}_j(\Omega)$ *contains a connected subset* $\tilde{C}_j(\Omega)$ *bifurcating from the jth eigencurve* $\{((\omega^2, \gamma^2), (0, 0)): \gamma^2 = \Gamma_j(\omega^2), \omega^2 \in [\Omega^{-1}, \Omega]\}$ *and containing solution pairs with* $\|(y, w)\|$ *unbounded.* $\tilde{K}_j(\Omega) \cap \tilde{K}_i(\Omega) = \varnothing$ *for* $i \neq j$. $\tilde{C}_j(\Omega)$ *has Lebesgue dimension at least two at each of its points.*

Further analysis shows that the level curves on which $\|(y, w)\| = R$ resemble hyperbolas and that the bifurcation diagram corresponding to Fig. 6.63 looks like a quarter of a saddle surface. All the solution sheets bifurcate from the lines $\omega^2 = 0$, $\gamma^2 = 0$. (Incidentally, the use of (6.65) promotes the further analysis of the problem for which $\varepsilon\eta = 1$.)

From (4.8) we see that ϕ' depends on $\omega\gamma$. Hence the number of zeros of x_1 and x_2, although not the number of zeros of v, on any sheet changes with ω and γ. To study this dependence, we use (6.50a) with $\delta^2 - 1$ taken to be the right side of (6.42c). We thus obtain

$$|\phi'| \geqq \frac{\omega\gamma}{A\omega^2 + B\omega\gamma + C} \tag{6.76}$$

where A, B, C are positive constants that depend on l and a. Contrast this behavior with that of § 3. The Prüfer transformation $\omega = r \cos \theta$, $\gamma = r \sin \theta$ converts (5.7b) to a system for which

$$r(s) = r(s_0) \exp\left[-\frac{\omega}{\gamma} \int_{s_0}^{s} \sin \theta(t) \cos \theta(t)\, dt\right], \qquad s_0 \in [0, l]. \tag{6.77}$$

From (6.77), (4.8), (5.6b), (5.12) we then get

$$r(s) \geqq r(s_0) e^{-\omega l/\gamma}, \tag{6.78}$$

$$|\phi'| \leqq \frac{|\omega\gamma| e^{\omega l/\gamma}}{\sqrt{\omega^2 \gamma^2 v(s_0)^2 + w(s_0)^2}}. \tag{6.79}$$

We finally observe that when $r^2 > 0$, the differential equation (5.18) involves only analytic functions. For one-parameter bifurcation problems of the form $f(x, \lambda) = 0$ where x is in a Banach space, λ is real, and f is real-analytic, Dancer [12] has shown that where the set of solution pairs is compact, it consists of a locally finite union of finite-dimensional analytic manifolds. That λ is a scalar is irrelevant to Dancer's proof. Thus his theorem is applicable to our problem away from the bifurcation points. Moreover, the theorem of Alexander and Antman [2] shows that the set of nontrivial solution pairs contains a minimal set each point of which has dimension at least two. The combination of these two results shows that the set of nontrivial solution pairs for our problem contains a locally finite union of analytic manifolds of dimension at least two.

Several authors ([3], [4], [27], [29]) have developed methods to handle lack of compactness in one-parameter problems. See [29] for an extensive bibliography on this matter.

7. Elastic strings. Here we indicate the extent to which our results for inextensible strings are typical of those for elastic strings. We begin by observing that the bifurcation diagram (Fig. 6.63) bears absolutely no resemblance to that corresponding to Fig. 3.11, because the singularities arising in the problem for the inextensible string reflect the fact that the trivial state must be folded at $b \in (0, l)$ for the problem to have meaning. The conditions under which Fig. 3.11 is valid ensure that $b \notin [0, l]$ so that the trivial state is not folded and so that $z' > 0$ for any steady state. Solutions for elastic strings can be expected to exhibit the properties of those for inextensible strings when $l > a$. But there are other pathologies that can arise. We now examine the simplest manifestation of these.

We seek a trivial solution of the problem for elastic strings, i.e., we seek an absolutely continuous function z and a real number b satisfying (2.28), (2.29):

$$(7.1)\qquad N(|z'(s)|)-\gamma^2\rho A|z'(s)|=\varepsilon\rho Ag(b-s)\operatorname{sign} z'(s),\quad z(0)=0,\quad z(l)=a.$$

To be specific let us assume that $\varepsilon=1$, that γ is given, and that N is concave. Then $\nu\mapsto N(\nu)-\gamma^2\rho A\nu$ has the form shown in Fig. 7.2. $\mu(\gamma)$ denotes the maximum value of this function. (It decreases with increasing $|\gamma|$.)

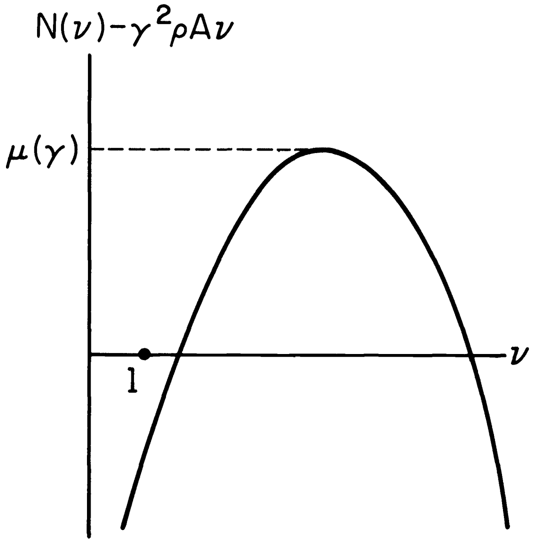


FIG. 7.2

We examine whether (7.1) admits a solution with z' continuous and positive. In this case (7.1) reduces to

$$(7.3\text{a, b, c})\qquad N(z'(s))-\gamma^2\rho Az'(s)=\rho Ag(b-s)\quad\text{for } s\in[0,l],\quad z(0)=0,\quad z(l)=a.$$

Equation (7.3a) has no solutions for z' if there is an s in $[0,l]$ such that $\rho Ag(b-s)>\mu(\gamma)$ and hence no solutions if $\rho Agb>\mu(\gamma)$. Equation (7.3a) has two continuous solutions for z' if $\rho Ag(b-s)\leqq\mu(\gamma)$ for all $s\in[0,l]$, i.e., if $\rho Agb\leqq\mu(\gamma)$. We denote the two continuous functions whose graphs form the inverse of that of Fig. 7.2 by $\nu^-(\cdot,\gamma)$ and $\nu^+(\cdot,\gamma)$ with $\nu^-(\xi,\gamma)<\nu^+(\xi,\gamma)$ for $\xi<\mu(\gamma)$. We set

$$(7.4)\qquad \Phi^\pm(g\rho Ab,\gamma)\equiv\int_0^l \nu^\pm(g\rho A(b-s),\gamma)\,ds\quad\text{when } g\rho Ab\leqq\mu(\gamma).$$

$\Phi^-(\cdot,\gamma)$ strictly increases from 0 at $b=-\infty$ to $\int_0^l \nu^-(\mu(\gamma)-g\rho As,\gamma)\,ds$ at $g\rho Ab=\mu(\gamma)$ while $\Phi^+(\cdot,\gamma)$ strictly decreases from ∞ at $b=-\infty$ to $\int_0^l \nu^+(\mu(\gamma)-g\rho As,\gamma)\,ds$ at $g\rho Ab=\mu(\gamma)$. Hence $\Phi^+(\mu(\gamma),\gamma)>\Phi^-(\mu(\gamma),\gamma)$. The graphs of $\Phi^\pm$ are shown in Fig. 7.5.

From Fig. 7.5 we immediately see that (7.3) has a continuous solution with $z'=\nu^\pm(\rho Ag(b-s))$ when $a\notin(\Phi^-(\mu(\gamma),\gamma),\Phi^+(\mu(\gamma),\gamma))$, for then one of the equations $\Phi^\pm(g\rho Ab,\gamma)=a$ has a unique solution for b in terms of a, these equations corresponding to the condition that $z(l)-z(0)=a$.

We do not get a solution with z' continuous if $a\in(\Phi^-(\mu(\gamma),\gamma),\Phi^+(\mu(\gamma),\gamma))$. In this case we can construct a discontinuous z' (nonuniquely) to satisfy (7.3). Now,

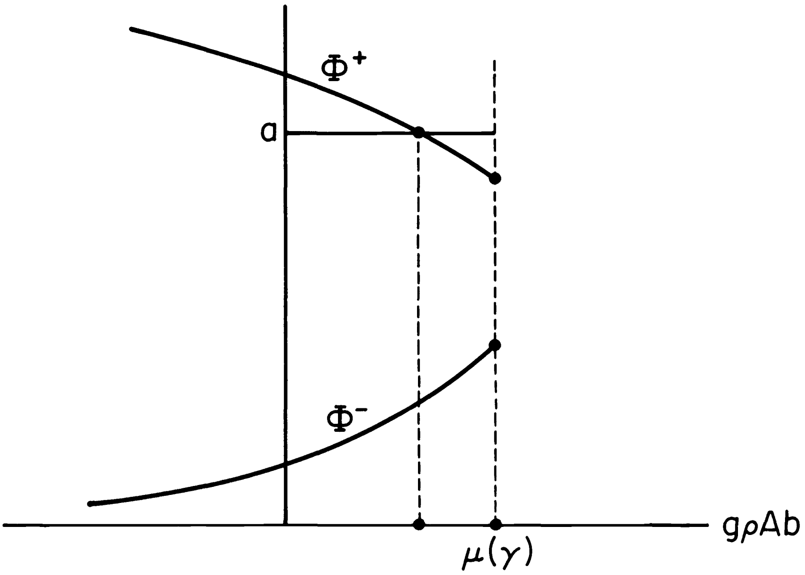


FIG. 7.5

since $N(\nu)\to\infty$ as $\nu\to\infty$, we can make $\mu(\gamma)$ as large as we like by taking $|\gamma|$ small enough. Suppose that $\mu(\gamma)\geqq g\rho Al$. Then $\Phi^-(\mu(\gamma),\gamma)>l\nu^-(\mu-g\rho Al)\geqq l$. Thus we fail to get a continuous solution for a range of a's exceeding l. In this range the average stretch a/l of the string exceeds one, yet the only possible steady motion is a standing shock.

We can similarly treat other cases under the requirement, introduced at the end of § 4, that $m\nu$ be continuous. Since the left side of (7.1) is just $m\nu$ for the class of trivial solutions under consideration, the continuity of $m\nu$ requires that z' could only change sign in $(0, l)$ at b, when $b\in(0, l)$.

The same difficulties arise for the full problem. Figure 7.2 represents a typical graph for $m\nu$. Other such graphs may have several local maxima and minima with $m\nu$ approaching ∞ as $\nu\to\infty$ if N is asymptotically superlinear and approaching $-\infty$ as $\nu\to\infty$ if N is asymptotically sublinear. The requirement that $m\nu$ be continuous allows many possible jumps in ν. An entropy condition is needed to eliminate solutions that are not physically realistic. Unfortunately the concept of being physically realistic is not mathematically precise. Admissibility conditions such as those of Liu [19] are based on the interpretation of the original hyperbolic system as a singular limit of a parabolic system with a mechanism for viscous dissipation. Such mechanisms are appropriate for treating shocks in fluids, but their suitability for solids is not so well established (cf. [6]). For our problem the jump conditions they provide are incompatible with those based on criteria such as Maxwell's equal area rule, which has a more purely thermodynamical motivation. Neither kind of jump condition can be relied upon to pick out a unique system of shocks for the steady motions we study. Our preliminary studies of shock conditions also show that they are very sensitive to an interpretation of the equations of motion for strings as singular limits of those for rods, which have flexural and torsional stiffness.

Even if we could confidently adopt an entropy condition as realistic, we have no assurance that the resulting operator equations admit a global bifurcation theory. (If they do, then a proof of that fact would likely require an analysis based on a regularization argument like that of § 6.) Nevertheless, we can still get useful global information about nontrivial solutions for elastic strings by exploiting the techniques of § 6 when $b \in (0, l)$ and by less drastic techniques otherwise:

THEOREM 7.6. *The number of (necessarily simple) zeros of v on $[0, l]$ is constant on any connected set of nontrivial shock-free solution pairs. (The topology of solution-parameter space is that of § 6.) Let D be a region of (ω^2, γ^2)-space for which the trivial solutions are shock-free. If the b for the trivial state lies outside $[0, l]$ for all $(\omega^2, \gamma^2) \in D$, then there is a neighborhood of D in solution-parameter space in which the bifurcation diagram resembles that corresponding to Fig. 3.11 with the bifurcating sheets inheriting their nodal properties from those of the eigenfunctions of the linearized problem. Let $D \subset \{(\omega^2, \gamma^2): \omega^2 \neq 0, \gamma^2 \neq 0\}$ and let $D \neq \emptyset$. If the b for the trivial state lies inside $(0, l)$ for all $(\omega^2, \gamma^2) \in D$, then there is a neighborhood of D in solution-parameter space in which the bifurcation diagram resembles those of § 6.*

Note that all these conclusions hold for $\gamma^2 < \gamma_0^2$, where γ_0^2 is defined in (4.1).

Finally we may comment briefly on the limit as an elastic string becomes inextensible. Let ν^* be the inverse of N and let $n = N(\nu)$. Then (2.24) yields

$$n = \rho A(m + \gamma^2)\nu^*(n). \tag{7.7}$$

If ν^* is sufficiently "flat," e.g., if ν^* is concave on $(0, \infty)$, then (7.7) has a unique solution $n = \hat{n}(m, \gamma)$ for n in terms of m. In the limit that the string becomes inextensible, i.e., as $\nu^*(n) \to 1$ for all n, this solution is obviously $\hat{n}(m, \gamma) = \rho A(m + \gamma^2)$. We now replace (2.33) with

$$(z')^2 = \nu^*(\hat{n}(m, \gamma))^2 - |u'|^2. \tag{7.8}$$

We may then follow the approach of §§ 5 and 6, observing than an equation such as (5.1c) becomes an implicit equation for m. It appears from our preliminary analysis of this question that the resulting bifurcation diagrams for slightly extensible strings converge nonuniformly to those for inextensible strings, the nonuniformity occurring for large values of γ^2. Note that the special material used in § 3 does not admit a natural limit process allowing the extensibility to go to zero.

Acknowledgments. We are indebted to J. C. Alexander, E. N. Dancer, and F. W. J. Olver for their very helpful comments on parts of this paper.

REFERENCES

[1] J. C. ALEXANDER, *A primer on connectivity*, in Proc. Conf. on Fixed Point Theory, 1980, E. Fadell and G. Fournier, eds., Springer Lecture Notes in Maths, Berlin, 886, 1981, pp. 455–483.

[2] J. C. ALEXANDER AND S. S. ANTMAN, *Global behavior of bifurcating multidimensional continua of solutions for multiparameter nonlinear eigenvalue problems*, Arch. Rational Mech. Anal., 76 (1981), pp. 339–354.

[3] J. C. ALEXANDER, S. S. ANTMAN AND S.-T. DENG, *Nonlinear eigenvalue problems for the whirling of heavy elastic strings* II: *New methods of global bifurcation theory*, Proc. Roy. Soc. Edinburgh, Sect. A, 93 (1983), pp. 197–227.

[4] C. J. AMICK AND J. F. TOLAND, *On solitary water-waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–95.

[5] S. S. ANTMAN, *The equations for large vibrations of strings*, Amer. Math. Monthly, 87 (1980), pp. 359–370.

[6] S. S. ANTMAN AND R. MALEK-MADANI, *Travelling waves in nonlinearly viscoelastic media and shock structure in elastic media*, to appear.

[7] F. BUCKENS, *Tables of Bessel functions of imaginary order*, Technical report, Dept. of App. Mech. and Math., Univ. of Louvain, Belgium, 1963.

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

[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol.* I, Wiley-Interscience, New York, 1953.

[10] M. G. Crandall and P. H. Rabinowitz, *Nonlinear Sturm-Liouville problems and topological degree*, J. Math. Mech., 19 (1970), pp. 1083-1102.

[11] ———, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321-340.

[12] E. N. Dancer, *Global structure of the solutions of nonlinear real analytic eigenvalue problems*, Proc. London Math. Soc. (3), 27 (1973), pp. 747-765.

[13] N. Dunford and J. T. Schwartz, *Linear Operators, Part* II, Wiley-Interscience, New York, 1963.

[14] E. Hille, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1969.

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[16] J. B. Keller, *Large amplitude motion of a string*, Amer. J. Phys., 27 (1959), pp. 584-586.

[17] I. I. Kolodner, *Heavy rotating string—a nonlinear eigenvalue problem*, Comm. Pure Appl. Math., 8 (1955), pp. 394-408.

[18] K. Kuratowski, *Topology*, 4th ed., Academic Press, New York, 1966.

[19] T.-P. Liu, *Shock waves in the nonisentropic gas flow*, J. Differential Equations, 22 (1975), pp. 442-452.

[20] S.-K. Luke and S. Weissman, *Bessel functions of imaginary argument*, Inst. for Molecular Physics, Univ. of Maryland, College Park, MD, Report DA-ARO(D)-31-124-G466 No. 1, 1964.

[21] S. P. Morgan, *Tables of Bessel functions of imaginary order and imaginary argument*, California Inst. of Technology, Pasadena, CA, 1947.

[22] M. A. Naimark, *Linear Differential Operators*, English transl., Unger, New York, 1967.

[23] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487-513.

[24] ———, *Some aspects of nonlinear eigenvalue problems*, Rocky Mountain J. Math., 3 (1973), pp. 161-202.

[25] M. Reeken, *The rotating string*, Math. Ann., 268 (1984), pp. 59-84.

[26] ———, *Exotic equilibrium states of the elastic string*, Proc. Roy. Soc. Edinburgh, Sect. A, 96 (1984), pp. 289-302.

[27] C. A. Stuart, *Some bifurcation theory for k-set contractions*, Proc. London Math. Soc. (3), 27 (1973), pp. 531-550.

[28] ———, *Spectral theory of rotating chains*, Proc. Roy. Soc. Edinburgh, Sect. A, 73 (1975), pp. 199-214.

[29] ———, *Bifurcation from the essential spectrum*, Proceedings of Equadiff 82, Springer Lecture Notes in Mathematics 1017, Berlin, 1983, pp. 575-596.

[30] G. T. Whyburn, *Topological Analysis*, rev. ed., Princeton Univ. Press, Princeton, NJ, 1964.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

007

CONVOLUTION OPERATORS IN A FADING MEMORY SPACE: THE CRITICAL CASE*

G. S. JORDAN†, OLOF J. STAFFANS‡ AND ROBERT L. WHEELER§

Abstract. We study the linear system of convolution equations

$$\mathscr{L}x(t) \equiv x'(t) + \mu * x(t) = f(t), \qquad t \in (-\infty, \infty),$$

where x, f are n-dimensional column vectors and μ is an $n \times n$ matrix-valued measure which is finite with respect to a suitable weight function. We describe the null space and the range of the operator $\mathscr{L}$ in a fading memory space. Our results include the previously untreated critical case when there may be a finite number of eigenvalues of the Laplace transform $\hat{L}(z) = zI + \hat{\mu}(z)$ of $\mathscr{L}$ on the boundary of the strip of convergence of $\hat{\mu}$. Our description is given in terms of the Jordan chains at the eigenvalues of the locally analytic matrix-valued function $\hat{L}(z)$. We prove a new Smith factorization theorem for locally analytic matrix functions. At the eigenvalues on the boundary of the strip of convergence, sufficient conditions for the existence of such a factorization are given in terms of the Banach algebra concept of the order of smoothness of a locally analytic matrix function and the structure of the Smith factorization. The authors have previously developed such Banach algebra methods to analyze scalar locally analytic functions.

Key words. convolution equation, locally analytic matrix function, Smith factorization, null space, range, fading memory

AMS(MOS) subject classifications. 45F05, 45Mxx, 34K20

1. Introduction.

We study the linear system of convolution equations

$$\mathscr{L}x(t) \equiv x'(t) + \mu * x(t) = f(t), \qquad t \in R \equiv (-\infty, \infty), \tag{1.1}$$

where x, f are n-dimensional column vectors and μ is an $n \times n$ matrix-valued measure which is finite with respect to a suitable weight function. As usual $\mu * x$ denotes the convolution

$$\mu * x(t) \equiv \int_{-\infty}^{\infty} d\mu(s) x(t-s).$$

We describe the null space and the range of the operator $\mathscr{L}$ in a fading memory space in the critical case when there may be a finite number of eigenvalues of the Laplace transform $\hat{L}(z) = zI + \hat{\mu}(z)$ of $\mathscr{L}$ (that is, zeros of the determinant of $\hat{L}(z)$) on the boundary of the strip of convergence of $\hat{\mu}$. Our descriptions are in terms of Jordan chains of vectors at the eigenvalues of the Laplace transform $\hat{L}$ of $\mathscr{L}$, and they depend on a Smith factorization theorem which we prove for locally analytic matrix functions.

Conditions sufficient to guarantee the existence of local and global Smith factorizations of locally analytic matrix functions are given in §3. These results are central to the subsequent analysis and are the most difficult theorems of the paper. The global Smith factorization theorem, Theorem 3.2, is the matrix analogue of the L^1-quotient theorem for scalar locally analytic functions [7, Thm. 3.4]. Since we are primarily concerned with permitting eigenvalues on the boundary of the strip of convergence of

* Received by the editors February 16, 1984, and in revised form November 1, 1985.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300.

‡ Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo, Finland.

§ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-4097. The work of this author was partially supported by the National Science Foundation under grants MCS 83-00559 and DMS 8500947.

locally analytic matrix functions, the development of these theorems is, of necessity, rather technical. For eigenvalues on the boundary of this strip, these theorems require that the components of the locally analytic matrix function satisfy certain smoothness hypotheses with the required order of smoothness being determined by the maximal partial multiplicity of the factored matrix. In the interior of the strip of convergence, the locally analytic matrix function is analytic and these smoothness assumptions hold automatically.

In § 2 we describe the weighted measure spaces in which we work, and review the concepts of local analyticity and smoothness of scalar locally analytic functions as developed in [7]. We remark that in [7, Lemma 4.3] we gave conditions on the components of the measure $\mu(t)$ that are necessary and sufficient to guarantee that the components of $\hat{\mu}(z)$, and hence of $\hat{L}(z)$, are smooth of a prescribed order. We conclude § 2 with a discussion of the fading memory spaces on which the operator $\mathscr{L}$ in (1.1) acts.

We describe the null space and the range of the operator $\mathscr{L}$ in (1.1) in §§ 5 and 6, respectively. The results established here are in the same spirit as those of [12]. However, they are more general because the eigenvalues of $\hat{L}(z)$ are not restricted to the interior of the strip of convergence of $\hat{\mu}$ as was the case in [12]. The results here also sharpen those in [12] since, roughly, our descriptions take account of polynomial rates of growth (or decay), whereas the descriptions in [12] only distinguish growth (or decay) rates that are exponentially separated. The methods of proof in [12] do not apply in the present situation; instead we must use the factorization theorems that we develop in § 3. Our description relates the null space and range of $\mathscr{L}$ to the Jordan chains of $\hat{L}(z)$. The notion of Jordan chains of vectors at the eigenvalues of a locally analytic matrix function and some of their elementary properties are described in § 4.

The principal application which motivates the results given here is the same as in [12] and [11], i.e., a study of the asymptotic behavior of the linear, infinite delay, autonomous system of functional differential equations

$$\begin{aligned} &x'(t)+\mu * x(t)=f(t), \qquad t\in R^+\equiv[0,\infty),\\ &x(t)=\phi(t), \qquad t\in R^-\equiv(-\infty,0], \end{aligned} \tag{1.2}$$

where $\mu(t)$ is supported on R^+. We analyze (1.2) in the sequel paper [8]. Again the results in [8] improve those in [12] and [11] since in [8] we study the critical case where the components of the solution of (1.2) are not exponentially separated as well as the noncritical case studied in [12] and [11].

In addition to the description of the null and range spaces of the Fredholm convolution operator (1.1) given in §§ 5 and 6, the analysis of the functional differential equation (1.2) in [8] requires us to study the singular part expansion of the inverse of a locally analytic matrix function at its eigenvalues. Conditions necessary to ensure the existence of such a singular part expansion follow from the Smith factorization theorems proved here. Since this expansion is not used in our study of (1.1), we postpone its development to [8]. Kappel and Wimmer [9] have used similar factorization and singular part expansions to decompose the null space of a system of linear functional differential equations with finite delay. However, in the case of finite delay equations, the characteristic equation $zI+\hat{\mu}(z)$ is entire, so the analytic difficulties that we encounter at the boundary of the strip of convergence do not arise.

Finally, we remark that the scalar L^1-quotient theorem as well as less precise results concerning singular part expansions have been used to study the asymptotic behavior of both scalar and system versions of Volterra integrodifferential equations of convolution type whose eigenvalues lie on the boundary of the half-plane of convergence of the Laplace transform of the kernel [7, Props. 5.1, 5.2, 5.3], [4], [5], [6], [10] (both in weighted and nonweighted L^1-spaces). The Smith factorization theorem for locally analytic matrix functions obtained here as well as the related results concerning singular part expansions given in [8] can be applied to sharpen and extend these earlier results for systems of integrodifferential equations.

2. Weighted measures and fading memory spaces. Let us begin with a short presentation of weighted L^1-spaces and locally analytic functions. For more details see, e.g., [7], [11] and [12].

We call the function ρ a *weight function*, if ρ is a Borel measurable, strictly positive function on R, ρ and ρ^{-1} are locally bounded, and ρ is submultiplicative, i.e.,

$$\rho(s+t) \leqq \rho(s)\rho(t), \qquad s, t \in R.$$

Without loss of generality one may always take $\rho(0)=1$. A typical weight function to keep in mind is

$$\rho(t)=\begin{cases} e^{-at}(1+|t|)^{\delta}, & t \in R^-, \\ e^{-bt}(1+t)^{\gamma}, & t \in R^+, \end{cases}$$

where $\delta, \gamma \geqq 0$ and $-\infty < b \leqq a < \infty$.

The space $L^1(C; \rho)$ consists of all complex measurable functions x on R for which

$$\|x\| \equiv \int_R |x(t)|\rho(t)\, dt < \infty.$$

For $x, y \in L^1(C; \rho)$, define the convolution of x and y by

$$x * y(t) = \int_R x(t-s)y(s)\, ds.$$

Then $L^1(C; \rho)$ becomes a commutative normed ring with convolution multiplication. We let $V(C; \rho)$ denote the ring one obtains from $L^1(C; \rho)$ by adjoining a unit.

Both $L^1(C; \rho)$ and $V(C; \rho)$ are closed subrings of $M(C; \rho)$, the space of locally finite complex Borel measures μ on R satisfying

$$\|\mu\| \equiv \int_R \rho(t)\, d|\mu|(t) < \infty,$$

where $|\mu|$ is the total variation measure of μ. The convolution of $a \in L^1(C; \rho)$ and $\mu \in M(C; \rho)$ is defined by

$$\mu * a(t) = a * \mu(t) = \int_R a(t-s)\, d\mu(s).$$

Define

$$\alpha = -\sup_{t<0} \frac{\log \rho(t)}{t} = -\lim_{t\to-\infty} \frac{\log \rho(t)}{t}, \tag{2.1}$$
$$\omega = -\inf_{t>0} \frac{\log \rho(t)}{t} = -\lim_{t\to\infty} \frac{\log \rho(t)}{t}$$

(in [7] these numbers were denoted by ρ^* and ρ_*, respectively). Then $-\infty < \omega \leqq \alpha < \infty$, and the maximal ideal space of $V(C; \rho)$ can be identified with $\bar{\Pi} = \Pi \cup \{\infty\}$, where

$$\Pi \equiv \{z \in C \mid \omega \leqq \operatorname{Re} z \leqq \alpha\}.$$

For the example of a weight function given earlier, $\omega = b$ and $\alpha = a$. We note that for the special case $\rho(t) \equiv 1$, $\omega = \alpha = 0$ and Π is the imaginary axis.

The bilateral Laplace transforms defined by

$$\hat{a}(z) = \int_R e^{-zt} a(t)\, dt, \qquad a \in L^1(C; \rho),$$

$$\hat{\mu}(z) = \int_R e^{-zt}\, d\mu(t), \qquad \mu \in M(C; \rho)$$

converge absolutely for $z \in \Pi$.

We let $C^{n \times n}$ stand for the space of $n \times n$ complex-valued matrices, and define $L^1(C^{n\times n}; \rho)$, $V(C^{n\times n}; \rho)$ and $M(C^{n\times n}; \rho)$ as above, but with complex-valued functions and measures replaced by $C^{n\times n}$-valued functions and measures. Also these are normed rings, if one defines the convolutions componentwise in the obvious way, but they are no longer commutative due to the fact that matrix multiplication is not commutative. One can define the bilateral Laplace transforms $\hat{a}(z)$ and $\hat{\mu}(z)$ of a function $a \in L^1(C^{n\times n}; \rho)$ and a measure $\mu \in M(C^{n\times n}; \rho)$, e.g., componentwise, and these too converge for all $z \in \Pi$.

We use the same concept of "local analyticity" as in [7], and call a complex function ϕ *locally analytic at a point* $z_0 \in \Pi$, if there exist measures $\mu_0, \cdots, \mu_k$ in $M(C; \rho)$ and a function $\psi(z, \xi_1, \cdots, \xi_k)$ analytic at $(z_0, \hat{\mu}_1(z_0), \cdots, \hat{\mu}_k(z_0))$ such that

$$\phi(z) = \psi(z, \hat{\mu}_1(z), \cdots, \hat{\mu}_k(z))$$

in a neighborhood of z_0. We say that ϕ is *locally analytic at infinity* if there exist functions $a_1, \cdots, a_m$ in $L^1(C; \rho)$, measures $\mu_1, \cdots, \mu_k$ in $M(C; \rho)$, and a function $\psi(z, \eta_1, \cdots, \eta_m, \xi_1, \cdots, \xi_k)$ analytic at $(0, 0, \cdots, 0)$ such that

$$\phi(z) = \psi(z^{-1}, \hat{a}_1(z), \cdots, \hat{a}_m(z), \hat{\mu}_1(z)/z, \cdots, \hat{\mu}_k(z)/z)$$

in a neighborhood of infinity. Throughout, the word "neighborhood" means an open subset of $\bar{\Pi}$ rather than an open subset of the compactified plane. Finally, we call ϕ *locally analytic* if it is locally analytic at each point of $\bar{\Pi}$.

The "smoothness" concept which we use is also the same as in [7]. We say that a complex function ϕ has a *zero of order at least m* at $z_0 \in \Pi$ if

$$\limsup_{z \to z_0,\, z \in \Pi} |(z - z_0)^{-m} \phi(z)| < \infty.$$

The point $z_0 \in \Pi$ is a *zero of order m* of ϕ if $\lim_{z \to z_0} (z - z_0)^{-m} \phi(z)$ exists and is nonzero. A very important subclass of zeros consists of those which are locally analytic; specifically, z_0 is a *locally analytic zero of order* (*at least*) *m* of ϕ if it is a zero of order (at least) m and $(z - z_0)^{-m} \phi(z)$ is locally analytic at z_0. Finally, we define a complex locally analytic function to be *smooth of order m* at z_0 [7, Def. 3.5] if $\phi(z) = \psi(z) + \zeta(z)$ near z_0 where ψ is analytic at z_0 and ζ is locally analytic at z_0 with a locally analytic zero of order at least m at z_0. Observe that a complex locally analytic function ϕ has a locally analytic zero of order (at least) m at z_0 if and only if it has a zero of order (at least) m at z_0 and is smooth of order m at z_0.

The scalar concepts listed above carry over to vector and matrix functions in an obvious way. We say that a vector or matrix function is locally analytic (at a point), has a zero of order at least m, has a locally analytic zero of order at least m, or is smooth of order m, if each component has the same scalar property.

In the following lemmas we list some elementary properties of the preceding definitions. They apply to scalar, vector and matrix functions, except when stated otherwise.

LEMMA 2.1. *A function ϕ is smooth of order m at z_0 if and only if it is of the form $\phi = p + \zeta$, where p is a polynomial of degree at most $m-1$, and ζ has a locally analytic zero of order at least m at z_0. The functions p and ζ in this decomposition are uniquely determined.*

Proof. Clearly, if $\phi = p + \zeta$, with p and ζ as above, then ϕ is smooth of order m at z_0. Also, if we have two decompositions $\phi = p_1 + \zeta_1 = p_2 + \zeta_2$, then $p_1 - p_2$ has a zero of order at least m at z_0; hence it vanishes identically. This proves the uniqueness of the decomposition.

To prove the existence of the decomposition, one first writes ϕ in the form $\phi = \psi + \zeta_1$, where ψ is analytic at z_0 and ζ_1 has a zero of order at least m at z_0, and then uses the Taylor series for ψ at z_0 to get

$$\phi = p + \zeta_1 + \zeta_2,$$

where p is a polynomial of degree at most $m-1$, ζ_2 is the remainder of order m in the Taylor series, and $\zeta_1 + \zeta_2$ has a zero of order at least m at z_0. □

If ϕ is smooth of order m at z_0, then we can define the *generalized derivatives* $\phi'(z_0), \cdots, \phi^{(m-1)}(z_0)$ of ϕ at the point z_0 to be

$$\phi^{(i)}(z_0) = p^{(i)}(z_0) = i!\,p_i, \qquad 1 \leqq i \leqq m-1$$

where

$$p(z) = \sum_{i=0}^{m-1} p_i (z - z_0)^i,$$

is the polynomial in the decomposition in Lemma 2.1. The generalized derivative $\phi^{(m)}(z_0)$ can be defined analogously by

$$\phi^{(m)}(z_0) = \lim_{z \to z_0,\, z \in \Pi} (z - z_0)^{-m} \eta(z),$$

where η is the remainder in the decomposition in Lemma 2.1. If ζ is m times continuously differentiable, then the generalized derivatives defined above coincide with the ordinary derivatives of ϕ at z_0. If both ϕ and ψ are smooth of order m at z_0, then so is $\phi\psi$ (cf. Lemma 2.2 below), and the generalized derivatives $(\phi\psi)^{(j)}(z_0)$, $0 \leqq j \leqq m$, of $\phi\psi$ at z_0 satisfy the same conditions as ordinary derivatives do, i.e.,

$$(\phi\psi)^{(j)}(z_0) = \sum_{i=0}^{j} \binom{j}{i} \phi^{(i)}(z_0) \psi^{(j-i)}(z_0). \tag{2.2}$$

DEFINITION 2.1. A scalar, vector or matrix quasipolynomial is an expression of the form $\sum_{k=0}^{m} A_k (z-c)^{-k}$, where the A_k are scalars, vectors or $n \times n$ matrices. Here $c < \omega$ is a fixed real number.

Note that a quasipolynomial may be expressed in the alternative form $\sum_{k=0}^{m} B_k((z - z_0)/(z-c))^k$ where $z_0 \neq c$ can be chosen arbitrarily, and each B_k is a scalar, vector or $n \times n$ matrix.

LEMMA 2.2. *Let ϕ and ψ be functions which are locally analytic at z_0.*

(i) *If both ϕ and ψ are smooth of order l at z_0, then each of $\phi \pm \psi$ is smooth of order l at z_0.*

(ii) *If ϕ has a zero of order at least m and is smooth of order $l \geqq m$ at z_0, and if ψ is smooth of order $l-m$ and has a zero of order at least $r \leqq l-m$ at z_0, then $\phi\psi$ and $\psi\phi$ are smooth of order l and have a zero of order at least $m+r$ at z_0.*

(iii) *If ϕ has a zero of order at least r at z_0, ψ is scalar-valued and has a zero of order m at z_0 and ϕ and ψ are smooth of order l at z_0, where $m \leqq r \leqq l$, then ϕ/ψ is* *smooth of order $l-m$ and has a zero of order at least $r-m$ at z_0. Moreover, there is a quasipolynomial p such that $\phi - p\psi$ has a zero of order at least l at z_0.*

Proof. The proofs of (i) and (ii) are obvious and will not be given. To prove (iii), first express ϕ and ψ in the form $\phi(z)=(z-z_0)^r\tilde{\phi}(z)$ and $\psi(z)=(z-z_0)^m\tilde{\psi}(z)$, where $\tilde{\phi}$ and $\tilde{\psi}$ are smooth of order $l-r$ and $l-m$, respectively. Since $\tilde{\psi}(z_0)\neq 0, 1/\tilde{\psi}$ is smooth of order $l-m$ at z_0 by [7, Lemma 4.2]. Now (ii) implies $\tilde{\phi}/\tilde{\psi}$ is smooth of order $l-r$, and thus $\phi(z)/\psi(z)=(z-z_0)^{r-m}\tilde{\phi}(z)/\tilde{\psi}(z)$ is smooth of order $l-m$, and has a zero of order at least $r-m$ at z_0.

As ϕ/ψ is smooth of order $l-m$, we can use Lemma 2.1 to write $\phi/\psi = p_1+\zeta_1$ where p_1 is a polynomial of degree at most $l-m-1$ and ζ_1 has a zero of order at least $l-m$ at z_0. If p is a quasipolynomial whose Taylor series expansion at z_0 up to order $l-m-1$ equals p_1, then also $\zeta=\phi/\psi-p=p_1-p+\zeta_1$ has a zero of order at least $l-m$ at z_0. By Lemma 2.2(ii), $\phi-\psi p=\zeta\psi$ has a zero of order at least l at z_0. □

Every measure $\mu\in M(C^{n\times n};\rho)$ induces a continuous operator μ^* on certain spaces of (fading) memory type. These spaces are defined as follows (for more details, see [11] or [12]).

We call η an *influence function dominated* by ρ if η is Borel measurable, strictly positive, $\eta(0)=1$, and

$$\eta(s+t)\leqq \eta(s)\rho(t), \qquad s,t\in R$$

(in [12] both ρ and η were supposed to be continuous, but that assumption was not important). In particular, ρ is an influence function dominated by itself. For each influence function η we define the *adjoint influence function* $\tilde{\eta}$ by

$$\tilde{\eta}(t)=[\eta(-t)]^{-1}, \qquad t\in R. \tag{2.3}$$

If η is dominated by ρ, then so is $\tilde{\eta}$. Moreover, every influence function dominated by ρ must satisfy

$$\tilde{\rho}(t)\leqq \eta(t)\leqq \rho(t), \qquad t\in R, \tag{2.4}$$

so $\tilde{\rho}$ and ρ, respectively, are the smallest and largest influence functions dominated by ρ. We point out, however, that not every η satisfying (2.4) is an influence function dominated by ρ. For example, if $\rho(t)=(1+|t|)^p$ and

$$\eta(t)=\begin{cases}(1+|t|)^{-q_-}, & t\in R^-,\\ (1+t)^{q_+}, & t\in R^+,\end{cases}$$

with p, q_- and q_+ nonnegative, then it is easy to check that η is dominated by ρ if and only if $q_-+q_+\leqq p$.

Every influence η induces a number of (fading) memory spaces. We let $L^p(C^n;\eta)$, $1\leqq p\leqq\infty$, be the space of measurable functions x: $R\to C^n$, with norm

$$\|x\|=\begin{cases}\left\{\int_R [\eta(t)\|x(t)\|]^p\,dt\right\}^{1/p}, & 1\leqq p<\infty,\\ \operatorname*{ess\,sup}_{t\in R}\eta(t)\|x(t)\|, & p=\infty.\end{cases}$$

We let the space $BUC(C^n;\eta)$ consist of those continuous functions $x\in L^\infty(C^n;\eta)$ which satisfy $\|\tau_t x-x\|\to 0$ as $t\to 0$, where τ_t is the translation operator $\tau_t x(s)=x(t+s)$, $s,t\in R$. An important subclass of $BUC(C^n;\eta)$ is $BC_0(C^n;\eta)$, which is made up of those functions x in $BUC(C^n;\eta)$ which satisfy

$$\lim_{t\to\infty}\operatorname*{ess\,sup}_{|s|\geqq t}\eta(t)\|x(t)\|=0.$$

We use the notation $\mathscr{B}(C^n; \eta)$ to represent any one of the preceding spaces, i.e., all results which are formulated in terms of $\mathscr{B}(C^n; \eta)$ remain true when $\mathscr{B}(C^n; \eta)$ is replaced by $L^p(C^n; \eta)$, $1 \leqq p \leqq \infty$, $BUC(C^n; \eta)$, and $BC_0(C^n; \eta)$. Finally, we define $\mathscr{B}^m(C^n; \eta)$ to be the space of all functions $x \in \mathscr{B}(C^n; \eta)$ whose distribution derivatives up to order m also belong to $\mathscr{B}(C^n; \eta)$. In particular, $\mathscr{B}^0(C^n; \eta)$ is the same as $\mathscr{B}(C^n; \eta)$.

The convolution of a measure $\mu \in M(C^{n\times n}; \rho)$ and a function $x \in \mathscr{B}^m(C^n; \eta)$ is defined by

$$\mu * x(t) = \int_R d\mu(s)x(t-s).$$

In the case when $m=0$ and $\mathscr{B}^m(C^n; \eta) = L^p(C^n; \eta)$ this convolution is well defined only a.e., but in the other cases it is defined for all $t \in R$. The convolution operator $\mu*$ maps each $\mathscr{B}^m(C^n; \eta)$, $m \geqq 0$, continuously back into itself [11, Lemma 2.1].

3. Factorization of locally analytic matrix functions. In this section we obtain local and global Smith factorizations for locally analytic matrix functions. Theorem 3.2, giving the global Smith factorization, is the matrix analogue of [7, Thm. 3.4]. Our development relies on the theory for matrix polynomials as developed in [2] (see also [3] for the Smith form for analytic matrix functions).

DEFINITION 3.1. Let the matrix function M be locally analytic at $z_0 \in \Pi$. We say that M has a local Smith factorization at z_0 if it has a right local Smith factorization

$$M(z) = R_1(z)D_1(z)P_1(z) \tag{3.1}$$

and a left local Smith factorization

$$M(z) = P_2(z)D_2(z)R_2(z) \tag{3.2}$$

in a neighborhood of z_0. Here P_i are unimodular (determinant identically one) quasipolynomials, R_1 and R_2 are locally analytic at z_0 with $\det R_1(z_0) \neq 0 \neq \det R_2(z_0)$, and D_1 and D_2 are diagonal quasipolynomials with diagonal entries of the form $((z-z_0)/(z-c))^k$, where the exponents are nonnegative integers which are nondecreasing as one moves down the diagonal of each of D_1 and D_2.

Trivially, if $M(z_0)$ is invertible, then M has a local Smith factorization at z_0 (take $D_1(z) \equiv P_1(z) \equiv D_2(z) \equiv P_2(z) \equiv I$). The interesting case is when $M(z_0)$ is not invertible. In this case we follow [2] and [3] and call z_0 an eigenvalue of M at z_0. (In our terminology, the constant function $M(z) \equiv A$, where A is invertible, has no eigenvalues. To get the "ordinary" eigenvalues of A, one has to study the analytic function $zI - A$ instead of the function $M(z) \equiv A$.)

The following lemma implies that the two diagonal matrices D_1 and D_2 in Definition 3.1 must be the same:

LEMMA 3.1. *Let* $z_0 \in \Pi$, *and let the matrix functions A and B be locally analytic at* z_0 *with* $\det A(z_0) \neq 0 \neq \det B(z_0)$. *If* D_1 *and* D_2 *are diagonal quasipolynomials having the structure of the diagonal quasipolynomials of Definition* 3.1, *and if*

$$D_2(z) = A(z)D_1(z)B(z) \tag{3.3}$$

in a neighborhood of z_0, *then* $D_1 = D_2$.

It follows immediately from Lemma 3.1 that if M has a local Smith factorization at z_0, then the diagonal quasipolynomials D_1 and D_2 occurring in the left and right local Smith factorizations of M are uniquely determined and equal. We denote this diagonal quasipolynomial by D. Also, the nonnegative integral powers $k_1 \leqq k_2 \leqq \cdots \leqq k_n$ which occur in D are called the *partial multiplicities* of M at z_0. The partial multiplicity k_n is called the *maximal partial multiplicity* of M at z_0, and the sum $k=\sum_{i=1}^n k_i$ is called the *algebraic order* of M at z_0 (z_0 is an eigenvalue of M if and only if the algebraic order of M at z_0 is positive). Alternative characterizations of k_n and k will be given in Theorem 3.1.

Proof of Lemma 3.1. For $i=1,2$ we have

$$D_i(z)=\operatorname{diag}\left(\left(\frac{z-z_0}{z-c}\right)^{k_{1i}},\cdots,\left(\frac{z-z_0}{z-c}\right)^{k_{ni}}\right),$$

where the nonnegative integers k_{ji} satisfy $k_{1i}\leqq k_{2i}\leqq\cdots\leqq k_{ni}$, $i=1,2$. Since the entries a_{ij} and b_{ij} of A and B are continuous at z_0 (in the relative topology of Π), a standard argument (see [1, Vol. I, § VI 3]), which uses the Binet-Cauchy formula to expand the right-hand side of (3.3), yields that

$$k_{12}+\cdots+k_{j2}\geqq k_{11}+\cdots+k_{j1},\qquad 1\leqq j\leqq n.$$

The reverse inequality is obtained by applying the same argument to $D_1=A^{-1}D_2B^{-1}$, and Lemma 3.1 follows immediately. □

We continue with another preliminary lemma:

LEMMA 3.2. *Suppose M is a locally analytic matrix function which is smooth of order l at $z_0\in\Pi$,* det *M has a zero of integral order $k\geqq 0$ at z_0, and all minors Δ of M of order $n-1$ satisfy*

$$\Delta(z)=O((z-z_0)^{k-l}),\quad z\to z_0,\quad z\in\Pi. \tag{3.4}$$

Then in each row and each column of M there is at least one element which has a locally analytic zero of integral order at most l at z_0.

Proof. We prove the result for rows only, since the proof for columns is completely analogous.

Develop det M along an arbitrary row i to obtain

$$\det M=\sum_{j=1}^{n}(-1)^{i+j}m_{ij}\Delta_{ij}, \tag{3.5}$$

where m_{ij} is the (i,j)-element of M and Δ_{ij} is the corresponding minor of M of order $n-1$. If each m_{ij}, $1\leqq j\leqq n$, satisfies $m_{ij}(z)=o((z-z_0)^l)$, $z\to z_0$, $z\in\Pi$, then by (3.4), (3.5) $\det M(z)=o((z-z_0)^k)$, $z\to z_0$, $z\in\Pi$. This estimate contradicts the assumption that det M has a zero of order exactly k at z_0. Thus, some element m_{ij} does not satisfy $m_{ij}(z)=o((z-z_0)^l)$, $z\to z_0$, $z\in\Pi$, and, being smooth of order l, it must have a locally analytic zero of order $q\leqq l$, q being an integer, possibly zero. □

We now can establish a sufficient condition for a matrix to have a local Smith factorization at z_0.

THEOREM 3.1. *Let the $n\times n$ matrix function $M=(m_{ij})$ be locally analytic at $z_0\in\Pi$ and assume that* det *M has a zero of integral order $k\geqq 0$ at z_0. If $n>1$, let $\sigma=\sigma(M)$ be the smallest nonnegative integer such that every minor Δ of M of order $n-1$ has a zero of order at least $k-\sigma$ at z_0; in the scalar case $n=1$ set $\sigma=k$. If M is smooth of order σ at z_0, then M has a local Smith factorization at z_0. In addition, the maximal partial multiplicity of M at z_0 equals σ, and the algebraic order of M at z_0 equals k.*

Note that z_0 is not assumed to be a *locally analytic* zero of order k of det M. That this is, in fact, the case follows from the conclusion of Theorem 3.1 since M satisfies (3.1) and (3.2) with

$$D_1(z)=D_2(z)=\operatorname{diag}\left(\left(\frac{z-z_0}{z-c}\right)^{k_1},\cdots,\left(\frac{z-z_0}{z-c}\right)^{k_n}\right)$$

and $k_1+k_2+\cdots+k_n=k$.

The smoothness assumption on M in Theorem 3.1 is far from necessary for M to have a local Smith factorization at z_0. For example, if ϕ_1 and ϕ_2 are scalar locally analytic functions with $\phi_1(z_0)\neq 0\neq \phi_2(z_0)$, then the matrix function $\operatorname{diag}[\phi_1(z),(z-z_0)\phi_2(z)]$ has a Smith factorization at z_0 with maximal partial multiplicity one, but it is not smooth of order one unless ϕ_1 is smooth of order one at z_0. It is also evident from the proof given below that not all the elements of M need be smooth of order σ for the construction to go through.

Proof of Theorem 3.1. We construct only a right local Smith factorization for M since the construction of a left local Smith factorization is completely analogous.

By multiplying M from the left and from the right by (unimodular) permutation matrices (i.e., $n\times n$ matrices $Q=(q_{st})$ which for some distinct i and j, $1\leqq i,j\leqq n$, satisfy $q_{st}=1$ if $s=t\neq i,j$, $q_{ij}=q_{ji}=1$, $q_{st}=0$, otherwise), we may assume that the order $k_1\geqq 0$ of the zero at z_0 of the element in the position (1, 1) is the minimum of the orders of the zeros at z_0 of all elements of M. Let us denote this rearranged matrix by $\tilde{M}=(\tilde{m}_{ij})$. It follows from Lemma 3.2 that $k_1\leqq\sigma$. By Lemma 2.2(iii), the quotients $\tilde{m}_{i1}/\tilde{m}_{11}$, $1\leqq i\leqq n$, are locally analytic and smooth of order $\sigma-k_1$. This means that we may successively left-multiply $\tilde{M}$ by unimodular locally analytic matrices to add the product of $-\tilde{m}_{i1}/\tilde{m}_{11}$ and the first row of $\tilde{M}$ to the ith row of $\tilde{M}$, where $2\leqq i\leqq n$; for fixed i, $2\leqq i\leqq n$, the required matrix has the form $R=(r_{st})$, where $r_{ss}=1$, $1\leqq s\leqq n$, $r_{i1}=-\tilde{m}_{i1}/\tilde{m}_{11}$ and $r_{st}=0$, otherwise. The result of these multiplications is a matrix $M_1=(m_{ij}^{(1)})$, where $m_{i1}^{(1)}\equiv 0$ for $2\leqq i\leqq n$. As our elimination matrices are smooth of order $\sigma-k_1$, and each of the elements of M has a zero of order at least k_1, it follows from Lemma 2.2 that each element of M_1 is smooth of order σ, and has a zero of order at least k_1. Also, the Binet–Cauchy formula shows that $\sigma(M_1)$, defined analogously to $\sigma(M)$, satisfies $\sigma(M_1)=\sigma(M)=\sigma$.

The first stage of the construction of a right local Smith factorization for M is now complete. In the next stage a similar sequence of steps will be used to place in the (2, 2) position an element with a zero at z_0 satisfying a certain minimality condition, and to assure that all off-diagonal elements in the second column vanish. Attaining the latter property requires an additional step to treat the elements above the diagonal; we begin with this step.

By Lemma 2.2(iii), there is a quasipolynomial p_{1j} for each j, $2\leqq j\leqq n$, such that $m_{1j}^{(1)}-p_{1j}m_{11}^{(1)}$ has a zero of order at least σ at z_0. Thus, we may successively right-multiply M_1 by unimodular quasipolynomial matrices to add the product of $-p_{1j}$ and the first column of M_1 to the jth column of M_1, where $2\leqq j\leqq n$; for fixed j, $2\leqq j\leqq n$, the required matrix has the form $U=(u_{st})$ where $u_{ss}=1$, $1\leqq s\leqq n$, $u_{1j}=-p_{1j}$ and $u_{st}=0$, otherwise. The effect of these multiplications is to replace each element $m_{1j}^{(1)}$, $j\geqq 2$, of M_1 with a locally analytic function which has a zero of order at least σ at z_0. Let us denote the new matrix by $\bar{M}_1=(\bar{m}_{ij}^{(1)})$.

Now left- and right-multiply $\bar{M}_1$ by permutation matrices to rearrange the $(n-1)\times(n-1)$ submatrix $(\bar{m}_{ij}^{(1)})$, $2\leqq i,j\leqq n$, so that its element which minimizes the order of the zero at z_0 is in its upper left-hand corner; by Lemma 3.2, if k_2 is the order of the zero at z_0 of the minimizing element, then $k_1\leqq k_2\leqq\sigma$. Observe that even after the permutations, all the elements in the first column except for the first vanish, and all the elements in the top row except for the first have a zero of order at least σ. As before, we may left-multiply the resulting matrix by unimodular locally analytic matrices to replace all the off-diagonal elements in the second column with identically zero functions. We get a matrix $M_2=(m_{ij}^{(2)})$ with the following properties: Each element of M_2 is smooth of order σ, $m_{i1}^{(2)}\equiv 0$, $i\neq 1$, $m_{i2}^{(2)}\equiv 0$, $i\neq 2$, $m_{11}^{(2)}=m_{11}^{(1)}$ has a zero at z_0 of order $k_1\geqq 0$, $m_{22}^{(2)}$ has a zero at z_0 of order k_2 with $k_1\leqq k_2\leqq\sigma$, each element $m_{1k}^{(2)}$,

$3 \leqq k \leqq n$, has a zero at z_0 of order at least σ, and each element $m_{ij}^{(2)}$, $2 \leqq i, j \leqq n$, has a zero at z_0 of order at least k_2. Moreover, $\sigma(M_2) = \sigma(M_1) = \sigma(M) = \sigma$.

Continue this process of right-multiplying by unimodular quasipolynomial matrices, multiplying by permutation matrices and then left-multiplying by unimodular locally analytic matrices to construct, successively, matrices $M_3, M_4, \cdots, M_n$ such that the elements $m_{ij}^{(l)}$ of M_l are smooth of order σ, $m_{ij}^{(l)} \equiv 0$ for $1 \leqq i, j \leqq l$, $i \neq j$, $m_{ii}^{(l)}$ has a zero at z_0 of order k_i, $1 \leqq i \leqq l$, with $0 \leqq k_1 \leqq k_2 \leqq \cdots \leqq k_l \leqq \sigma$, and, if $l < n$, each element $m_{ij}^{(l)}$ for $1 \leqq i \leqq l-1$, $l+1 \leqq j \leqq n$, has a zero at z_0 of order at least σ and each element $m_{ij}^{(l)}$ for $l \leqq i \leqq n$, $l+1 \leqq j \leqq n$ has a zero at z_0 of order at least k_l. In addition, $\sigma(M_l) = \sigma$, $1 \leqq l \leqq n$.

Clearly, each element $m_{ii}^{(n)}$, $1 \leqq i \leqq n$, of the diagonal matrix M_n may be written in the form

$$m_{ii}^{(n)}(z) = \left(\frac{z - z_0}{z - c}\right)^{k_i} \tilde{m}_{ii}^{(n)},$$

where $\tilde{m}_{ii}^{(n)}$ is locally analytic at z_0 and $\tilde{m}_{ii}^{(n)}(z_0) \neq 0$. Thus, we may left-multiply M_n by the invertible locally analytic matrix diag $(1/\tilde{m}_{11}^{(1)}, \cdots, 1/\tilde{m}_{nn}^{(n)})$ to obtain the diagonal matrix

$$D(z) = \operatorname{diag}\left(\left(\frac{z - z_0}{z - c}\right)^{k_1}, \cdots, \left(\frac{z - z_0}{z - c}\right)^{k_n}\right).$$

To complete the right local Smith factorization we now simply invert all the different left-factors which we have obtained and combine them into one locally analytic factor R_1 with det $R_1(z_0) \neq 0$, and likewise we invert all the different right-factors and combine them into one unimodular quasipolynomial factor P_1.

Finally, it was observed earlier that $\sum_{i=1}^{n} k_i = k$, i.e., that the algebraic order of M at z_0 equals k. We also note that because of the nature of the diagonal elements of M_n and because $\sigma(M_n) = \sigma$, we have $k - \sigma = \sum_{i=1}^{n-1} k_i$. Thus, the maximal partial multiplicity k_n of M at z_0 is given by $k_n = \sigma$. This completes the proof of Theorem 3.1. □

As an immediate consequence of Theorem 3.1 we have the following.

COROLLARY 3.1. *Let M have a local Smith factorization at $z_0 \in \Pi$ with diagonal quasipolynomial D and with maximal partial multiplicity σ. Suppose that the locally analytic matrix function Q is smooth of order σ at z_0, and that* det $Q(z_0) \neq 0$. *Then MQ and QM have local Smith factorizations at z_0 with the same diagonal quasipolynomial D.*

Proof. We prove the result for MQ; the proof for QM is completely analogous. The left factorization of MQ clearly holds with R_2 from (3.2) replaced by R_2Q. To obtain the right factorization from (3.1) we must rewrite DP_1Q. By Theorem 3.1 we may rewrite this expression as $DP_1Q = \tilde{R}DP$, where $\tilde{R}$ is locally analytic at z_0 with det $\tilde{R}(z_0) \neq 0$ and P is a unimodular quasipolynomial matrix. By Lemma 3.1 we still have the same diagonal quasipolynomial D on the right-hand side; thus, substituting this new expression for DP_1Q into (3.1) one gets the desired result for MQ. □

We conclude this section by giving conditions which ensure the existence of a "global Smith factorization" on $\bar{\Pi}$ of a locally analytic matrix function M that has only a finite number of eigenvalues, each of which is in Π and admits a local Smith factorization. This result is the matrix analogue of the scalar L^1-quotient theorem [7, Thm. 3.4].

THEOREM 3.2. *Let the n-by-n matrix function M be locally analytic on $\bar{\Pi}$, assume that* det $M(z) \neq 0$ *except on a finite set $Z = \{z_1, \cdots, z_N\} \subset \Pi$, and let M have a local Smith factorization at each $z_l \in Z$. Then M has a global Smith factorization on $\bar{\Pi}$, i.e., a* *right global Smith factorization*

$$M(z)=R_1(z)D(z)P_1(z), \qquad z\in\bar{\Pi}, \tag{3.6}$$

and a left global Smith factorization

$$M(z)=P_2(z)D(z)R_2(z), \qquad z\in\bar{\Pi}. \tag{3.7}$$

Here P_1 *and* P_2 *are unimodular quasipolynomials,* R_1 *and* R_2 *are locally analytic on* $\bar{\Pi}$ *with* $\det R_i(z)\neq 0$, $i=1,2$, *for all* $z\in\Pi$, *and* D *is the diagonal quasipolynomial*

$$D=\prod_{l=1}^{N} D_l, \tag{3.8}$$

where, for $1\leqq l\leqq N$, D_l *is the diagonal quasipolynomial occurring in the local Smith factorization of* M *at* z_l.

Proof. Once again we construct only the right factorization since the construction of the left one is completely analogous.

The matrix M has the local Smith factorization

$$M(z)=R_{11}(z)D_1(z)P_{11}(z) \tag{3.9}$$

in a neighborhood of z_1. Since D_1 and P_{11} have analytic inverses in $\bar{\Pi}\backslash\{z_1\}$, we may solve (3.9) for R_{11} in a neighborhood of z_1 and extend R_{11} to a locally analytic matrix defined on all of $\bar{\Pi}$ in such a way that (3.9) holds on $\bar{\Pi}$. By Corollary 3.1, this extended R_{11} has a right local Smith factorization $R_{11}(z)=R_{21}(z)D(z)P_{21}(z)$ at z_2 with D_2 being the diagonal quasipolynomial in the factorization of M at z_2. Substituting this expression for R_{11} in (3.9), one gets a factorization for M in a neighborhood of z_2.

Clearly, we may solve for R_{21} in a neighborhood of z_2 and then extend its domain to all of $\bar{\Pi}$ in the same way that the domain of R_{11} was extended. Continuing the argument above and then repeating it until all the eigenvalues $z_1,\cdots,z_N$ are included, one arrives at a decomposition of M on $\bar{\Pi}$ of the form

$$M=R_{N1}D_NP_{N1}D_{N-1}P_{N-1,1}\cdots D_1P_{11}. \tag{3.10}$$

Here $R_{l1}=R_{l+1,1}D_{l+1}P_{l+1,1}$, $1\leqq l\leqq N-1$, and each D_l is the diagonal quasipolynomial from the local Smith factorization of M at z_l.

The product $Q\equiv D_NP_{N1}D_{N-1}P_{N-1,1}\cdots D_1P_{11}$ is a quasipolynomial, i.e., a polynomial in $w\equiv(z-c)^{-1}$; hence, by [2, Thm. S1.1], Q has a global Smith factorization $Q=RDP_1$ where P_1 is a unimodular quasipolynomial, R is a quasipolynomial with constant nonzero determinant, $D=\operatorname{diag}(d_1,\cdots,d_n)$ and each d_i is a monic scalar quasipolynomial. Because of the form of Q it is easy to express the minors of Q in terms of the minors of the factors of Q (cf. [1, p. 12]) and then to use [2, Thms. S1.2 and S1.4] to determine the d_i and to establish (3.8). Thus, substituting the global Smith factorization of Q in (3.10) and letting $R_1=R_{N1}R$, we obtain the right global Smith factorization (3.6). □

4. Jordan chains. In order to describe the null space and the range of the convolution operator $\mathscr{L}$ defined in the introduction we have to introduce the concept of Jordan chains of a locally analytic matrix function with a Smith factorization. The easiest way to do this seems to be via the concept of root functions.

DEFINITION 4.1. Let q be a positive integer, and let the matrix function M be locally analytic at $z_0\in\Pi$. We say that a (column) vector-valued function r is a right root function of M of order at least q at z_0 if r is locally analytic and smooth of order q at z_0, $r(z_0)\neq 0$, and Mr has a zero of order at least q at z_0. Define

$$r_i=\frac{1}{i!}r^{(i)}(z_0), \qquad 0\leqq i\leqq q-1, \tag{4.1}$$

where $r^{(i)}(z_0)$ is the generalized derivative of order i of r at z_0. For any $p, 0 \leqq p \leqq q-1$, we call the vector sequence $r_0, \cdots, r_p$ a right Jordan chain of length $p+1$ of M at z_0; the vector $r(z_0)=r_0$ is said to be a right eigenvector corresponding to the eigenvalue z_0.

If M is smooth of order p at z_0, then so is Mr, and it follows from (2.2) and (4.1) that

$$\sum_{i=0}^{j} \frac{1}{i!} M^{(i)}(z_0) r_{j-i} = 0, \qquad 0 \leqq j \leqq p, \tag{4.2}$$

where $M^{(i)}(z_0)$ is the generalized derivative of order i of M at z_0. This means that our definition of a right Jordan chain of length $p+1$ agrees with, e.g., the definition given in [3, p. 91]. Observe that $r_0, r_1, \cdots, r_p$ is a right Jordan chain of M at z_0 if and only if the polynomial $\sum_{i=0}^{p} r_i(z-z_0)^i$ is a right root function of order at least $p+1$ of M at z_0.

LEMMA 4.1. *Let M and T be $n \times n$ matrix functions that are locally analytic at $z_0 \in \Pi$, and assume that* det $T(z_0) \neq 0$. *Then r is a right root function of M of order at least q at z_0 if and only if r is a right root function of TM of order at least q at z_0. In particular, if M has a local Smith factorization $R_1 D P_1$ at z_0, then the right root functions and right Jordan chains of M are identical with those of the right factor DP_1 at z_0.*

Proof. If r is a right root function of M of order at least q at z_0, then

$$\limsup_{z \to z_0,\, z \in \Pi} |(z-z_0)^{-q} T(z) M(z) r(z)| \leqq |T(z_0)| \limsup_{z \to z_0,\, z \in \Pi} |(z-z_0)^{-q} M(z) r(z)| < \infty.$$

Conversely, if r is a right root function of TM of order at least q at z_0, it follows that

$$\limsup_{z \to z_0,\, z \in \Pi} |(z-z_0)^{-q} M(z) r(z)| \leqq |T(z_0)^{-1}| \limsup_{z \to z_0,\, z \in \Pi} |(z-z_0)^{-q} T(z) M(z) r(z)| < \infty,$$

and Lemma 4.1 is proved. □

We emphasize that the right root functions and right Jordan chains of a locally analytic matrix function M having a local Smith factorization at z_0 do not depend on the particular factorization chosen; moreover, since they are determined by the quasipolynomial right factor DP_1 at z_0, the theory of right root functions and right Jordan chains for analytic matrix functions as developed in [3] (see also [2]) can be directly applied in this setting. In particular, the length of a right Jordan chain at a point cannot exceed the maximal partial multiplicity of M at that point.

If we multiply a matrix function M from the right by a sufficiently smooth matrix with nonvanishing determinant, then the right Jordan chains change in the same way as in the analytic case:

LEMMA 4.2. *Let the matrix function M have a local Smith factorization at $z_0 \in \bar{\Pi}$ with maximal partial multiplicity σ, and let the matrix function T be smooth of order σ at z_0, and satisfy* det $T(z_0) \neq 0$. *Then $r_0, \cdots, r_p$ is a right Jordan chain of length $p+1$ of MT at z_0 if and only if*

$$s_j = \sum_{i=0}^{j} \frac{1}{i!} T^{(i)}(z_0) r_{j-i}, \qquad 0 \leqq j \leqq p \tag{4.3}$$

is a right Jordan chain of M at z_0.

The proof is essentially the same as the proof of, e.g., [3, p. 93] and [2, Prop. 1.11]. It is a consequence of the fact that the polynomial $r(z)=\sum_{i=0}^{p} r_i(z-z_0)^i$ is a right root function of MT of order at least $p+1$ at z_0 if and only if Tr is a right root function of M of order at least $p+1$ at z_0, and that the polynomial part of Tr in the decomposition in Lemma 2.1 equals $\sum_{i=0}^{p} s_i(z-z_0)^i$.

Lemma 4.1 allows us to immediately transfer the concept of a canonical set of right Jordan chains as developed for matrix analytic functions [3] (see also [2, p. 32]) to the present setting.

DEFINITION 4.2. Let the $n \times n$ matrix function M be locally analytic and have a local Smith factorization at $z_0 \in \Pi$, and assume that $\det M$ has a zero of positive integer order k at z_0. Let $k_1 \leqq \cdots \leqq k_s$, $s \leqq n$, be positive integers, and let

$$r_{i0}, \cdots, r_{i,k_i-1}, \qquad 1 \leqq i \leqq s, \tag{4.4}$$

be a set of right Jordan chains of M at z_0. The set of sequences (4.4) is said to be a canonical set of right Jordan chains of M at z_0 if the vectors $r_{10}, \cdots, r_{s0}$, are linearly independent, and $k_1 + \cdots + k_s = k$.

We note that since the right Jordan chains of $M = R_1 D P_1$ are determined by the quasipolynomial right factor DP_1, the existence of a canonical set of right Jordan chains of M at z_0 is guaranteed by [3, pp. 92–93], and the properties of canonical sets of right Jordan chains are exactly those for analytic matrices as developed in [3]. In particular, the positive integers k_j, $1 \leqq j \leqq s$, are exactly the nonzero partial multiplicities $k_{n-s+1} \leqq \cdots \leqq k_n$ of M at z_0 [3]. The number s in Definition 4.2 equals the dimension of the null space (eigenspace) of $M(z_0)$, and it is often called the *geometric order* of M at z_0.

The importance of a canonical set of right Jordan chains of M at z_0 derives from the fact that its elements can be used to generate a basis for the set of all right Jordan chains of M at z_0. More precisely, let σ be the maximal partial multiplicity of the zero of M at z_0, and let $\mathcal{N}$ be the subspace of $C^{n \times \sigma}$ consisting of all sequences of the form $(0, 0, \cdots, 0, r_0, \cdots, r_p)$, where $r_0, \cdots, r_p$ is a right Jordan chain of M at z_0, and the total number of vectors is σ. Then it follows [2, Prop. 1.15] that the set of sequences (4.4) is a canonical set of right Jordan chains of M at z_0 if and only if the set of sequences

$$\gamma_{jp} = (0, \cdots, 0, r_{j0}, \cdots, r_{jp}), \qquad 0 \leqq p \leqq k_j - 1, \quad 1 \leqq j \leqq s,$$

where the number of zero vectors preceding r_{j0} in γ_{jp} is $\sigma - (p+1)$, forms a basis for $\mathcal{N}$. Observe that the total number of sequences γ_{jp} above equals the algebraic order of the zero of M at z_0. In other words, the dimension of $\mathcal{N}$ is the same as the algebraic order of M at z_0.

The notions of left root function and left Jordan chain of a locally analytic matrix M are defined analogously to the corresponding notions of right root function and right Jordan chain except that column vectors are replaced by row vectors and left multiplications are replaced by right multiplications throughout. The results concerning left root functions and left Jordan chains are completely analogous to the corresponding results for right root functions and right Jordan chains. In particular, if M has a left global Smith factorization $P_2 D R_2$, then the left root functions and left Jordan chains of M at z_0 are identical with those of the left factor $P_2 D$. A canonical set of left Jordan chains is defined analogously to a canonical set of right Jordan chains.

5. The null space of $\mathcal{L}$. In this section we describe the null space of the convolution operator

$$\mathcal{L}x(t) \equiv x'(t) + \mu * x(t) \tag{5.1}$$

in a space of fading memory type. Here $\mu \in M(C^{n \times n}; \rho)$. A similar description has been given in [12], but that one is not as general as the one given here (in [12] it is assumed that all the critical points belong to the interior of Π), and also is less precise (no connection between the null space of $\mathscr{L}$ and the right Jordan chains of $\hat{L}$ is made in [12]).

Formally, the Laplace transform of $\mathscr{L}$ is the function

$$\hat{L}(z) \equiv z + \hat{\mu}(z), \qquad z \in \Pi, \tag{5.2}$$

which is locally analytic in Π. We assume that $\hat{L}$ has only a finite set $Z = \{z_1, \cdots, z_N\}$ of eigenvalues. Moreover, we assume that $\hat{L}$ has a local Smith factorization at each point z_l of Z. Of course, this assumption is automatically satisfied at points of Z in the interior of Π. The function $\hat{L}$ is unbounded at infinity, and cannot therefore be locally analytic there. However, if we define

$$M(z) \equiv (z-c)^{-1}\hat{L}(z) = (z-c)^{-1}(z + \hat{\mu}(z)), \qquad z \in \bar{\Pi}, \tag{5.3}$$

where as before $c < \omega$, then M is locally analytic on all of $\bar{\Pi}$. It also has a local Smith factorization at each point z_l of Z. By Theorem 3.2, M has a right global Smith factorization

$$M(z) = R_1(z)D(z)P_1(z), \qquad z \in \bar{\Pi}. \tag{5.4}$$

Substituting this factorization into (5.2), we get the factorization

$$\hat{L}(z) = (z-c)R_1(z)D(z)P_1(z), \qquad z \in \Pi, \tag{5.5}$$

for $\hat{L}$ itself. This factorization is the main tool which we use to determine the nullspace of $\mathscr{L}$ in $\mathscr{B}^{m+1}(C^n; \eta)$.

To see how the factorization (5.5) can be used, note that since $\det R_1(z) \neq 0$, $z \in \bar{\Pi}$, the locally analytic matrix function R_1 is the transform of an invertible element $\xi_1 \in V(C^{n\times n}; \rho)$. Also the quasipolynomial $Q_1(z) \equiv D(z)P_1(z) = \sum_{i=0}^{q} A_i(z-c)^{-i}$, A_0 invertible, is the transform of the element $\nu_1 = \sum_{i=0}^{q} A_i e^{i*}$ in $V(C^{n\times n}; \rho)$. Here

$$e(t) = \exp(ct)I, \quad t \geqq 0, \qquad e(t) = 0, \quad t < 0,$$

$e^{0*} = \delta I$ with I the unit point mass at zero, and e^{i*}, $i = 1, 2, \cdots$, denotes the i-fold convolution $e * e * \cdots * e$.

LEMMA 5.1. *The operator* $\mathscr{L}$: $\mathscr{B}^{m+1}(C^n; \eta) \to \mathscr{B}^m(C^n; \eta)$ *has the same null space as the operator* $\mathscr{Q}_1 : \mathscr{B}^{m+1}(C^n; \eta) \to \mathscr{B}^{m+1}(C^n; \eta)$ *defined by*

$$\mathscr{Q}_1 x \equiv \nu_1 * x. \tag{5.6}$$

Proof. Since ξ_1 is invertible in $V(C^{n\times n}; \rho)$, Lemma 5.1 of [12] gives that the operator $(d/dt - c)\xi_1 *$maps $\mathscr{B}^{m+1}(C^n; \eta)$ one-to-one and onto $\mathscr{B}^m(C^n; \eta)$. Thus, since $\mathscr{Q}_1$ maps $\mathscr{B}^{m+1}(C^n; \eta)$ into $\mathscr{B}^{m+1}(C^n; \eta)$ and $\mathscr{L}x = (d/dt - c)\xi_1 * (\mathscr{Q}_1 x)$, the proof of Lemma 5.1 is complete. □

The null space of $\mathscr{Q}_1$ in $\mathscr{B}^{m+1}(C^n; \eta)$ is easy to compute. To do this, we follow the procedure in [11] and for $a \geqq b$ let $\rho_{a,b}$ denote the weight function

$$\rho_{a,b}(t) = \begin{cases} e^{-at}, & t \leqq 0, \\ e^{-bt}, & t > 0. \end{cases}$$

Choose α_1 and ω_1 so that $c < \omega_1 < \omega \leqq \alpha < \alpha_1 < \infty$. Then $\nu_1 \in V(C^{n\times n}; \rho_{\alpha_1,\omega_1})$. Also, by (2.1), (2.3) and (2.4), there exists a $T > 0$ such that

$$\tilde{\rho}_{\alpha_1,\omega_1}(t) \leqq \tilde{\rho}(t) \leqq \eta(t), \qquad |t| \geqq T;$$

hence, $\mathscr{B}^{m+1}(C^n; \eta) \subset \mathscr{B}^{m+1}(C^n; \tilde{\rho}_{\alpha_1,\omega_1})$.

Now when $\operatorname{Re} z=\alpha_1$ or $\operatorname{Re} z=\omega_1$, $\det Q_1(z)\neq 0$, so Lemmas 4.1 and 4.3 of [12] tell us that all solutions of

$$\left(\frac{d}{dt}-c\right)\mathscr{Q}_1 x(t)=0$$

in $\mathscr{B}^{m+1}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$ are of the form

$$x(t)=\sum_{l=1}^{N} p_l(t)\,e^{z_l t}, \tag{5.7}$$

where the p_l are (column) vector polynomials in t of degree at most one less than the order of the zero of $\det Q_1$ at z_l. Since $(d/dt-c)$ maps $\mathscr{B}^{m+1}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$ one-to-one onto $\mathscr{B}^{m}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$ (use [12, Lemma 3.1]), formula (5.7) characterizes the form of all functions x in the null space of $\mathscr{Q}_1$ in $\mathscr{B}^{m+1}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$. As $\mathscr{B}^{m+1}(C^n;\eta)\subset\mathscr{B}^{m+1}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$, it also characterizes the form of all functions x in the null space of $\mathscr{Q}_1$ in $\mathscr{B}^{m+1}(C^n;\eta)$, and by Lemma 5.1, the form of all functions x in the null space of $\mathscr{L}$ in $\mathscr{B}^{m+1}(C^n;\eta)$.

Formula (5.7) gives the form of all functions in the null space of $\mathscr{Q}_1$, but not all functions of the form (5.7) actually do belong to the null space. To identify exactly the null space of $\mathscr{Q}_1$ in $\mathscr{B}^{m+1}(C^n;\tilde{\rho}_{\alpha_1,\omega_1})$, we need the notion of right Jordan chains, developed in § 4.

LEMMA 5.2. *Let $\nu\in M(C^{n\times n};\rho_{\alpha_1,\omega_1})$, and let x be of the form*

$$x(t)=\sum_{i=0}^{p}\frac{t^i}{i!}r_{p-i}\,e^{z_0 t}, \tag{5.8}$$

with $r_0\neq 0$ and $\alpha_1<\operatorname{Re} z_0<\omega_1$. Then $\nu*x=0$ if and only if $r_0, r_1,\cdots, r_p$ is a right Jordan chain of $\hat{\nu}$ at z_0.

Proof. First use the binomial theorem, then interchange the order of the summation, and finally make a change of summation variables to get

$$\begin{aligned}\nu*x(t)&=\sum_{i=0}^{p}\int_R d\nu(s)\frac{(t-s)^i}{i!}r_{p-i}\,e^{z_0(t-s)}\\&=e^{z_0 t}\sum_{j=0}^{p}\frac{t^{p-j}}{(p-j)!}\sum_{i=0}^{j}\frac{1}{i!}\left\{\int_R(-s)^i e^{-z_0 s}\,d\nu(s)\right\}r_{j-i}.\end{aligned}$$

By elementary Laplace transform theory,

$$\int_R(-s)^i e^{-z_0 s}\,d\nu(s)=\hat{\nu}^{(i)}(z_0),$$

so we have

$$\nu*x(t)=e^{z_0 t}\sum_{j=0}^{p}\frac{t^{p-j}}{(p-j)!}\sum_{i=0}^{j}\frac{1}{i!}\hat{\nu}^{(i)}(z_0)r_{j-i}.$$

Clearly, this implies that $\nu*x=0$ if and only if

$$\sum_{i=0}^{j}\frac{1}{i!}\hat{\nu}^{(i)}(z_0)r_{j-i}=0,\qquad 0\leqq j\leqq p,$$

and since $\hat{\nu}$ is analytic at z_0 this condition is equivalent to $r_0, r_1, \cdots, r_p$ being a right Jordan chain of $\hat{\nu}$ at z_0. □

Combining Lemmas 5.1 and 5.2 with Lemma 4.1, we get the following.

COROLLARY 5.1. *A function $x \in \mathscr{B}^{m+1}(C^n; \eta)$ of the form* (5.8) *with $r_0 \neq 0$ and $z_0 \in \Pi$ belongs to the null space of $\mathscr{L}$ if and only if $r_0, r_1, \cdots, r_p$ is a right Jordan chain of $\hat{L}$ at z_0.*

Observe in particular the requirement that $x \in \mathscr{B}^{m+1}(C^n; \eta)$. This may or may not restrict the growth of x in (5.8) at plus or minus infinity in such a way that only those right Jordan chains whose length does not exceed a fixed number generate functions in the null space of $\mathscr{L}$. In the two settings discussed in [12, §§ 5 and 6], either all right Jordan chains at z_0 generate functions in the null space of $\mathscr{L}$, or none of them does.

By combining Corollary 5.1 with the discussion preceding Lemma 5.2, we can give a complete description of the null space $\eta(\mathscr{L})$ of $\mathscr{L}$ in $\mathscr{B}^{m+1}(C^n; \eta)$.

THEOREM 5.1. *Let $\mathscr{L}$ be the operator in* (5.1) *acting on $\mathscr{B}^{m+1}(C^n; \eta)$. Assume that $\hat{L}$ has only a finite set $Z = \{z_1, \cdots, z_N\}$ of eigenvalues, and that $\hat{L}$ has a local Smith factorization at each point of Z. Then the null space $\mathscr{N}(\mathscr{L})$ is the direct sum*

$$\mathscr{N}(\mathscr{L}) = \mathscr{N}_1 \oplus \cdots \oplus \mathscr{N}_N, \tag{5.9}$$

where each $\mathscr{N}_l$ is the set of functions $x \in \mathscr{N}(\mathscr{L})$ in Corollary 5.1 *with z_0 replaced by $z_l \in Z$, plus the zero function.*

6. The range of $\mathscr{L}$. Here we describe the range of the convolution operator $\mathscr{L}$ defined in (5.1) as a mapping from $\mathscr{B}^{m+1}(C^n; \eta)$ into $\mathscr{B}^m(C^n; \eta)$. The setting is exactly the same as in § 5. Namely, $\hat{L}$ defined in (5.2) has only a finite set $Z = \{z_1, \cdots, z_N\}$ of eigenvalues, and $\hat{L}$ has a local Smith factorization at each point z_l of Z. By Theorem 3.2, the locally analytic matrix M defined in (5.3) has a left global Smith factorization; substituting this factorization into (5.2), we get the factorization

$$\hat{L}(z) = (z - c)P_2(z)D(z)R_2(z), \qquad z \in \Pi. \tag{6.1}$$

This factorization is the tool which we use to determine the image of $\mathscr{L}$ in $\mathscr{B}^m(C^n; \eta)$.

In analogy with § 5, we use the fact that $\det R_2(z) \neq 0$, $z \in \bar{\Pi}$, to see that R_2 is the transform of an invertible element $\xi_2 \in V(C^{n \times n}; \rho)$. Also, the quasipolynomial $Q_2 = P_2 D$ is the transform of the element

$$\nu_2 \equiv Q_2(e) = \sum_{i=0}^{q} B_i e^{i*}$$

in $V(C^{n \times n}; \rho)$.

LEMMA 6.1. *The operator $\mathscr{L}$: $\mathscr{B}^{m+1}(C^n; \eta) \to \mathscr{B}^m(C^n; \eta)$ has the same range as the operator $\mathscr{Q}_2 : \mathscr{B}^m(C^n; \eta) \to \mathscr{B}^m(C^n; \eta)$ defined by*

$$\mathscr{Q}_2 x \equiv \nu_2 * x. \tag{6.2}$$

Proof. Using Lemma 3.5 of [11], we get

$$\mathscr{L}x = \left(\frac{d}{dt} - c\right) \mathscr{Q}_2(\xi_2 * x) = \mathscr{Q}_2\left[\left(\frac{d}{dt} - c\right) \xi_2 * x\right]$$

when $x \in \mathscr{B}^{m+1}(C^n; \eta)$. Now as in the proof of Lemma 5.1, we use the fact that $(d/dt - c)\xi_2 *$ maps $\mathscr{B}^{m+1}(C^n; \eta)$ one-to-one onto $\mathscr{B}^m(C^n; \eta)$ to complete the proof of Lemma 6.1. □

Our description of the range of $\mathscr{L}$ is given in terms of the left Jordan chains of $\hat{L}$ (which, by the analogue of Lemma 4.1 for left Jordan chains, are the same as the left Jordan chains of P_2D), and the concept of (left) Jordan chains of functions in $\mathscr{B}^m(C^n; \eta)$, which can be defined in the following way:

DEFINITION 6.1. The row vectors $v_0, \cdots, v_p$ with $v_0 \neq 0$ form a Jordan chain of length $p+1$ of $f \in \mathscr{B}^m(C^n; \eta)$ at $z_0 \in \Pi$ if there exist scalar functions $F_1, \cdots, F_{p+1} \in \mathscr{B}^{m+1}(C; \eta)$ satisfying

$$\left(\frac{d}{dt} - z_0\right) F_1 = v_0 f, \quad \left(\frac{d}{dt} - z_0\right) F_{j+1} = F_j + v_j f, \quad 1 \leqq j \leqq p. \tag{6.3}$$

In the case when $\eta = \rho$ and $\mathscr{B} = L^1$ one could also use an obvious modification of Definition 4.1 and define the notion of a left Jordan chain of $\hat{f}$ at a point $z_0 \in \Pi$. The connection between these two types of Jordan chains is very simple; the vectors $v_0, \cdots, v_p$ form a left Jordan chain of $\hat{f}$ at z_0 if and only if they form a Jordan chain of f at z_0 in the sense of Definition 6.1 (a proof of one direction of this claim is contained in the proof of Lemma 6.2 below; see, in particular, (6.4)). In other words, one should regard the notion of a Jordan chain of a function $f \in \mathscr{B}^m(C^n; \eta)$ defined above as a generalization of the notion of a left Jordan chain of $\hat{f}$ (Definition 6.1 makes sense even when $\hat{f}$ does not exist).

THEOREM 6.1. *Let $\mathscr{L}$ be the operator* (5.1) *acting on $\mathscr{B}^{m+1}(C^n; \eta)$. Assume that $\hat{L}$ has only a finite set $Z = \{z_1, \cdots, z_N\}$ of eigenvalues, and that $\hat{L}$ has a local Smith factorization at each point of Z. Then f in $\mathscr{B}^m(C^n; \eta)$ belongs to the range of $\mathscr{L}$ if and only if, for every $z_l \in Z$, every left Jordan chain of $\hat{L}$ at z_l is also a Jordan chain of f at z_l.*

In a way this result is very natural. Thinking of the interpretation of a Jordan chain of f as a left Jordan chain of $\hat{f}$, one could interpret Theorem 6.1 as a factorization result. It generalizes the statement that $\hat{L}$ is a left divisor of $\hat{f}$ if and only if every left Jordan chain of $\hat{L}$ is also a left Jordan chain of $\hat{f}$. In the setting of analytic matrix-valued functions this is a well-known result (see e.g., [3, Thm. 1.4]).

In [12] the range of $\mathscr{L}$ was characterized in two different cases. In [12, § 5] the influence function η is "small" at infinity, and $\mathscr{L}$ is a surjection, i.e., the range of $\mathscr{L}$ is all of $\mathscr{B}^m(C^n; \eta)$. The same result can be deduced from Theorem 6.1 because it is not difficult to show that, under the assumption in [12, § 5], for every $f \in \mathscr{B}^m(C^n; \eta)$, every sequence $v_0, \cdots, v_p$ of vectors with $v_0 \neq 0$ is a Jordan chain of f at those points z_l allowed in [12]. On the other hand, in [12, § 6], the influence function is "large" at infinity, and that forces the Laplace transform of f to converge and be analytic in a neighborhood of every eigenvalue z_l allowed in [12]. In that situation one could equally well work with the notion of Jordan chains of $\hat{f}$ mentioned above. Much more can be said about the significance of Theorem 6.1, and we shall discuss this question in a forthcoming paper.

We begin the proof of Theorem 6.1 by first proving the necessity of the Jordan chain condition in Theorem 6.1:

LEMMA 6.2. *Let $\nu \in M(C^{n \times n}; \rho)$, let $z_0 \in \Pi$ be an eigenvalue of $\hat{\nu}$, and assume that $\hat{\nu}$ is smooth of order q at z_0. Suppose that $f = \nu * x$, where $x \in \mathscr{B}^m(C^n; \eta)$. If $0 \leqq p \leqq q-1$, then every left Jordan chain of length $p+1$ of $\hat{\nu}$ at z_0 is also a Jordan chain of f at z_0.*

We remark that Lemma 6.2 gives a necessary condition for f to belong to the range of $\mathscr{L}$ even in the case when $\hat{L}$ has an infinite number of eigenvalues. This is true because the factorization (6.1) is not used to prove the necessity of the condition in Theorem 6.1; all that is required is that $\hat{L}$ be suitably smooth at each of its eigenvalues. We also remark that we can apply Lemma 6.2 to the measure $\nu_2 = Q_2(e)$ obtained from the factorization (6.1) since $\hat{\nu}_2$ is analytic on all of Π.

Proof of Lemma 6.2. Let $v_0, \cdots, v_p$, with $v_0 \neq 0$, be a left Jordan chain of $\hat{\nu}$ at z_0, where $0 \leqq p \leqq q-1$. Define

$$v(z) = \sum_{i=0}^{p} v_i (z-z_0)^i,$$

and note that

$$v(z)\hat{\nu}(z) = \sum_{i=0}^{p} v_i \hat{\nu}(z)(z-z_0)^i$$

has a locally analytic zero of order at least $p+1$ at z_0. In particular, the partial sum

$$\sum_{i=0}^{j} v_i \hat{\nu}(z)(z-z_0)^i$$

has a locally analytic zero of order at least $j+1$ at z_0 for each j, $0 \leqq j \leqq p$. Hence, if we define $\hat{a}_j$, $1 \leqq j \leqq p+1$, inductively by

$$(z-z_0)\hat{a}_1(z) = v_0\hat{\nu}(z), \qquad (z-z_0)\hat{a}_{j+1}(z) = \hat{a}_j(z) + v_j\hat{\nu}(z), \qquad 1 \leqq j \leqq p,$$

then each $\hat{a}_j$ is the Laplace transform of a function $a_j \in L^1(C^n; \rho)$ (see [7, Prop. 2.3]). Using Lemma 3.1 of [11], we find that the functions a_j satisfy $a_j' \in M(C^n; \rho)$ (where prime denotes a distribution derivative), and

$$a_1' - z_0 a_1 = v_0 \nu, \quad a_{j+1}' - z_0 a_{j+1} = a_j + v_j \nu, \quad 1 \leqq j \leqq p. \tag{6.4}$$

Define $F_j = a_j * x$ for $1 \leqq j \leqq p+1$. Then by [11, Lemma 3.6], $F_j' = a_j' = a_j' * x$, so $F_j \in \mathscr{B}^{m+1}(C; \eta)$. Moreover, by (6.4),

$$\left(\frac{d}{dt} - z_0\right) F_1 = v_0 \nu * x = v_0 f,$$

and

$$\left(\frac{d}{dt} - z_0\right) F_{j+1} = a_j * x + v_j \nu * x = F_j + v_j f$$

for $1 \leqq j \leqq p$. Hence $v_0, \cdots, v_p$ is a Jordan chain of f at z_0. □

By Lemma 6.1, f is in the range of $\mathscr{L}$ if and only if it is in the range of the operator $\mathscr{Q}_2$ corresponding to the quasipolynomial left factor $P_2 D$. The next lemma describes the action of the invertible operator corresponding to the unitary quasipolynomial P_2, and thereby further reduces the problem to an examination of the range of the operator corresponding to the diagonal quasipolynomial D.

LEMMA 6.3. *Let ν and ν^{-1} both belong to $M(C^{n \times n}; \rho_{\alpha_1, \omega_1})$ and let z_0 satisfy $\alpha_1 < \operatorname{Re} z_0 < \omega_1$. If $f \in \mathscr{B}^m(C^n; \eta)$, then $v_0, \cdots, v_p$ is a Jordan chain of $\nu * f$ at z_0 if and only if*

$$w_j = \sum_{i=0}^{j} \frac{1}{i!} v_{j-i} \hat{\nu}^{(i)}(z_0), \qquad 0 \leqq j \leqq p, \tag{6.5}$$

is a Jordan chain of f at z_0.

The proof given below is very similar to the proof of Lemma 6.2.

Proof. First, observe that it suffices to prove Lemma 6.3 in one direction only since if $g = \nu * f$, then $f = \nu^{-1} * g$, and, by a Taylor series argument (cf. Lemma 4.2), (6.5) is equivalent to

$$v_j = \sum_{i=0}^{j} \frac{1}{i!} w_{j-i} (\nu^{-1})^{\wedge(i)}(z_0), \qquad 0 \leqq j \leqq p.$$

Also observe that $w_0 \neq 0$ if and only if $v_0 \neq 0$.

Suppose that $w_0, \cdots, w_p$ is a Jordan chain of f at z_0, and let $F_1, \cdots, F_{p+1}$ in $\mathcal{B}^{m+1}(C;\eta)$ satisfy relation (6.3) with all the v_j's replaced by w_j's. By (6.5) and the fact that $\hat{\nu}$ is analytic at z_0, we find that, for $0 \leqq j < p$, the partial sum

$$\psi_{j+1}(z) \equiv \sum_{i=0}^{j} (v_i\hat{\nu}(z) - w_i)(z-z_0)^i$$

has an analytic zero of order at least $j+1$ at z_0. Hence, for $1 \leqq j \leqq p+1$, there exist functions $a_j \in L^1(C^n; \rho_{\alpha_1,\omega_1})$ so that

$$\hat{a}_j(z) = (z-z_0)^{-j}\psi_j(z), \qquad z \in \Pi.$$

From this formula and Lemma 3.1 of [11] we see that the distribution derivatives a_j' satisfy

$$\text{(6.6)} \qquad a_1' - z_0 a_1(z) = v_0\nu - w_0\delta, \quad a_{j+1}' - z_0 a_{j+1} = a_j + v_j\nu - w_j\delta, \quad 1 \leqq j \leqq p,$$

where δ is the unit point mass at zero.

Define $G_j \in \mathcal{B}^{m+1}(C;\eta)$ by

$$G_j = F_j + a_j * f, \qquad 1 \leqq j \leqq p+1.$$

Then by (6.6), [11, Lemma 3.6] and (6.3), we get

$$\left(\frac{d}{dt} - z_0\right) G_1 = w_0 f + v_0\nu * f - w_0 f = v_0 \nu * f,$$

$$\left(\frac{d}{dt} - z_0\right) G_{j+1} = F_j + w_j f + a_j * f + v_j \nu * f - w_j f = G_j + v_j \nu * f, \qquad 1 \leqq j \leqq p.$$

Thus, $v_0, \cdots, v_p$ is a Jordan chain of $\nu * f$ at z_0. □

Finally, to complete our proof that the Jordan chain condition in Theorem 6.1 is sufficient for f to belong to the range of $\mathcal{L}$, we have the following lemma for scalar quasipolynomials:

LEMMA 6.4. *Let $z_1, \cdots, z_l$ be distinct points satisfying* $\operatorname{Re} z_l > c$, *and let $p_1, \cdots, p_N$ be positive integers. Define the operator d on $\mathcal{B}^m(C;\eta)$ by*

$$\text{(6.7)} \qquad dx \equiv d * x \equiv \prod_{l=1}^{N} * (\delta - (z_l - c)e)^{p_l *} * x, \qquad x \in \mathcal{B}^m(C;\eta).$$

Let $f \in \mathcal{B}^m(C;\eta)$, and assume that for each l, $1 \leqq l \leqq N$, the sequence $1, 0, \cdots, 0$ *is a Jordan chain of length p_l of f at z_l. Then f belongs to the range of d.*

Here the notation $\prod_{l=1}^{N} * d_l$ denotes the convolution product $d_1 * d_2 * \cdots * d_N$ when $N \geqq 2$, and d_1 when $N = 1$.

Proof. The Laplace transform of the measure d in (6.7) is given by

$$\hat{d}(z) = \prod_{l=1}^{N} \left(\frac{z-z_l}{z-c}\right)^{p_l}.$$

Write

$$\text{(6.8)} \qquad \prod_{l=1}^{N} \left(\frac{z-z_l}{z-c}\right)^{p_l} = \sum_{l=1}^{N} \sum_{j=0}^{p_l} a_{lj}(z-z_l)^{-j}$$

for appropriate constants a_{lj}. Define d_{lj} by

$$d_{lj}(t) = \left(\frac{d}{dt} - z_l\right)^{p_l - j} e^{p_l *}(t)$$

for $1 \leqq l \leqq N$, $0 \leqq j \leqq p_l$. It is easy to check that, for $1 \leqq l \leqq N$, $d_{lj} \in L^1(C; \rho)$ for $0 \leqq j \leqq p_l - 1$ and $d_{lp_l} \in V(C; \rho)$. Observe that

$$\hat{d}_{lj}(z) = (z + z_l)^{p_l - j}(z - c)^{-p_l},$$

so d is given by

$$d = \prod_{l=1}^{N} * \, d_{l0}.$$

Moreover, it follows from (6.8) that

$$\sum_{l=1}^{N} \sum_{j=0}^{p_l} a_{lj} \prod_{\substack{i=1 \\ i \neq l}}^{N} * \, d_{i0} * d_{lj} = \delta. \tag{6.9}$$

Set $F_{l0} = f$ for $1 \leqq l \leqq N$, and let $F_{l1}, \cdots, F_{lp_l}$ be the functions in Definition 6.1 corresponding to the Jordan chains $v_{l0} = 1, v_{l1} = \cdots = v_{l,p_l - 1} = 0$ of f at z_l. Note in particular that, due to the special structure of these Jordan chains, $F_{lj} \in \mathscr{B}^{m+j}(C; \eta)$, and that the repeated application of Lemma 3.5 of [11] yields

$$d_{l0} * F_{lj} = d_{lj} * f \tag{6.10}$$

for $1 \leqq l \leqq N, 0 \leqq j \leqq p_l$.

Define $x \in \mathscr{B}^m(C; \eta)$ by

$$x(t) = \sum_{l=1}^{N} \sum_{j=0}^{p_l} a_{lj} F_{lj}.$$

It follows from the commutativity of convolution, (6.9) and (6.10) that

$$\begin{aligned} dx &= \prod_{i=1}^{N} * \, d_{i0} * \sum_{l=1}^{N} \sum_{j=0}^{p_l} a_{lj} F_{lj} \\ &= \sum_{l=1}^{N} \sum_{j=0}^{p_l} a_{lj} \prod_{\substack{i=1 \\ i \neq l}}^{N} * \, d_{i0} * d_{lj} * f = f, \end{aligned}$$

so f belongs to the range of d. □

Proof of Theorem 6.1. As stated above, the necessity of the condition in Theorem 6.1 for f in $\mathscr{B}^m(C^n; \eta)$ to belong to the range of $\mathscr{L}$ is an immediate consequence of Lemmas 6.1 and 6.2, and the fact that the left Jordan chains of $\hat{L}$ are identical with those of the left factor $\hat{\nu}_2 = P_2 D$ in (6.1).

Conversely, suppose that $f \in \mathscr{B}^m(C^n; \eta)$ and, for every $z_l \in Z$, every left Jordan chain of $\hat{L}$ at z_l is also a Jordan chain of f at z_l. By Lemma 6.1 and Lemma 4.1 for left Jordan chains it suffices to prove that f belongs to the range of the operator $\mathscr{Q}_2$. Let $\nu_3 = P_2(e)$ and $\nu_4 = D(e)$ be the measures whose transforms are the factors P_2 and D in (6.1), respectively. If α_1 and ω_1 satisfy $c < \omega_1 < \omega \leqq \alpha < \alpha_1$, then ν_3, ν_3^{-1} and ν_4 all belong to $V(C^{n \times n}; \rho_{\alpha_1, \omega_1})$. Since $\nu_2 = \nu_3 * \nu_4$, the function f belongs to the range of $\mathscr{Q}_2$ if and only if $g \equiv \nu_3^{-1} * f$ belongs to the range of the diagonal operator $\mathscr{D}x \equiv \nu_4 * x$. Thus, by Lemma 6.3 and the analogue of Lemma 4.2 for left Jordan chains, it suffices to show that g in $\mathscr{B}^m(C^n; \eta)$ belongs to the range of $\mathscr{D}$ whenever, for each eigenvalue $z_l \in Z$, every left Jordan chain of $\hat{D}$ is also a Jordan chain of g. But $D = \operatorname{diag}[d_1, \cdots, d_n]$, where each d_i is the monic scalar quasiopolynmial

$$d_i(z) = \prod_{l=1}^{N} \left(\frac{z - z_l}{z - c} \right)^{k_{il}}.$$

Here $k_{il} \geqq 0$ is the ith partial multiplicity of D at z_l. For each l with $k_{il} > 0$, the vectors $v_{i0} = (0, 0, \cdots, 1, 0, \cdots, 0)$ with the one in the ith coordinate, and $v_{i1} = \cdots = v_{i, k_{il}-1} = 0$ form a left Jordan chain of length k_{il} of D at z_l; hence, by assumption, $v_{i0}, \cdots, v_{i,k_{il}-1}$ is a Jordan chain of g at z_l. Lemma 6.4 shows that the ith component of g belongs to the range of the scalar operator d_i corresponding to d_i. Thus, g belongs to the range of $\mathscr{D}$. □

Acknowledgment. The authors wish to thank Professor J. A. Ball for several helpful conversations.

REFERENCES

[1] F. R. GANTMACHER, *The Theory of Matrices*, Vols. I and II, Chelsea, New York, 1959.

[2] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[3] I. GOHBERG AND L. RODMAN, *Analytic matrix functions with prescribed local data*, J. d'Analyse Math., 40 (1981), pp. 90-128.

[4] K. B. HANNSGEN, *A Wiener-Lévy theorem for quotients, with applications to Volterra equations*, Indiana Univ. Math. J., 29 (1980), pp. 103-120.

[5] G. S. JORDAN AND R. L. WHEELER, *Asymptotic behavior of unbounded solutions of linear Volterra integral equations*, J. Math. Anal. Appl., 55 (1976), pp. 596-615.

[6] ———, *Weighted L^1-remainder theorems for resolvents of Volterra equations*, this Journal, 11 (1980), pp. 885-900.

[7] G. S. JORDAN, O. J. STAFFANS AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra Equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749-782.

[8] ———, *Subspaces of stable and unstable solutions of a functional differential equation in a fading memory space: The critical case*, Report HTKK-MAT-A229, Helsinki University of Technology, Helsinki, 1985.

[9] F. KAPPEL AND H. K. WIMMER, *An elementary divisor theory for autonomous linear functional differential equations*, J. Differential Equations, 21 (1976), pp. 134-147.

[10] R. K. MILLER, *Structure of solutions of unstable linear Volterra integrodifferential equations*, J. Differential Equations, 15 (1974), pp. 129-157.

[11] O. J. STAFFANS, *On a neutral functional differential equation in a fading memory space*, J. Differential Equations, 50 (1983), pp. 183-217.

[12] ———, *The null space and range of a convolution operator in a fading memory space*, Trans. Amer. Math. Soc., 281 (1984), pp. 361-388.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

008

NONLINEAR EVOLUTION EQUATIONS WITH ALMOST PERIODIC TIME DEPENDENCE*

GEORGE SEIFERT†

Abstract. We give conditions under which solutions of the equation $u'(t) = Au(t)+f(t)$ are asymptotically almost periodic; here $u(t)$ and $f(t)$ are functions on the reals with values in a real Banach space X, $f(t)$ is almost periodic in the sense of Bohr and A is a function on a subset $D(A)$ of X, to X. We apply the result to a nonlinear one-dimensional heat equation with an almost periodic time-dependent heat supply.

Key words. nonlinear evolution equation, almost periodic function, asymptotically almost periodic function

AMS(MOS) subject classification. 34G20

1. Introduction. Let $\{X, |\ |\}$ denote a real Banach space. Let A denote a function on $D(A) \subset X$ to X, and f a function on R to X; R the set of reals. If we assume that f is almost periodic (a.p.) in the usual sense of Bohr, it is the purpose of this paper to give conditions on A so that strong solutions of

$$u' = Au + f(t) \tag{1}$$

on $[s, \infty)$ for some fixed $s \in R$ are asymptotically almost periodic (a.a.p.) and approach an a.p. generalized solution of (1) as $t \to \infty$.

As an application, we obtain a result for the nonlinear one-dimensional heat equation with a.p. time-dependent heat supply:

$$\begin{aligned} &u_t(t, \xi) = (\sigma(u_\xi))_\xi + r(t, \xi), \quad t \in R, \quad 0 < \xi < 1, \\ &u(t, 0) = u(t, 1) = 0 \quad \text{for } t \in R, \end{aligned} \tag{2}$$

where $r(t, \xi)$ is a.p. in t. Our result for (2) is in terms of almost periodicity in t for $L^2[0, 1]$-valued functions in terms of the $L^2[0, 1]$ norm; it is the purpose of a future investigation to find additional conditions under which almost periodicity with respect to the usual real valued norm is obtained.

Our methods are based on some results in [1]. The case where f is periodic is treated in [1] where a monotonicity condition on A, similar to the one we use, is imposed.

We assume without loss of generality that $A0 = 0$, the zero in X, since the conditions we impose on f also hold for $f(t) + x_0$, x_0 any element in X.

2. Definitions and preliminary results. We say that the function $u : [s, \infty) \to X$ is asymptotically almost periodic (a.a.p. for short) if it is continuous and if there exists an a.p. function $v : R \to X$ such that $u(t) - v(t) \to 0$ as $t \to \infty$.

We will use the following results, which can easily be established by following the proofs for the corresponding finite-dimensional results given in, for example, the book by Fink [2]; it is important to observe that the range of any a.p. function is contained in a compact subset of X (cf. [3]).

PROPOSITION 1. *The continuous function $u : [s, \infty) \to X$ is a.a.p. if and only if given a sequence $\{t'_n\}$, $t'_n \geqq s$, $t'_n \to \infty$ as $n \to \infty$, there exists a subsequence $\{t_n\}$ such that $\{u(t + t_n)\}$ converges uniformly for $t \geqq s$.*

* Received by the editors March 25, 1985; accepted for publication April 7, 1986.

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

PROPOSITION 2. *If $v:R\to X$ is a.p. there exists a sequence $\{t_n\}$, $t_n\to\infty$ as $n\to\infty$, such that $v(t+t_n)-v(t)\to 0$ as $n\to\infty$ uniformly for $t\in R$.*

The next results are basically in [1] and are concerned with solutions of (1). We define $u:[s,\infty)\to X$ to be a strong solution of (1) if $u(t)$ is continuous, is absolutely continuous on each compact interval of $[s,\infty)$, is differentiable almost everywhere on (s,∞) and satisfies (2) almost everywhere on (s,∞). If $u(s)=x$, x is called the initial value of $u(t)$. We shall henceforth refer to a strong solution of (1) as simply a solution of (1).

For each (t,s) in a subset of $R\times R$ of the form $a\leqq s\leqq t\leqq b$, a function $U(t,s):X\to X$ is called an evolution operator on X if

(i) $U(s,s)x=x$ for each $x\in X$;

(ii) $U(t,s)U(s,r)=U(t,r)$ for $a\leqq r\leqq s\leqq t\leqq b$;

(iii) $U(t,s)x$ is continuous in (t,s) for each $x\in X$.

Here and henceforth we always suppose $a\leqq s\leqq t\leqq b$ with $a=-\infty$ and $b=\infty$ possible.

We say that $U(t,s)$ is an evolution operator for (1) if for each solution $u:[s,b)\to X$ of (1), $u(t)=U(t,s)u(s)$.

If $U(t,s)$ is an evolution operator for (1) and $x\in X$ is arbitrary, we call $U(t,s)x$ a generalized solution of (1) on $[s,b)$. If $a=-\infty$ and $b=\infty$, we say that $U(t,s)x$ is a generalized solution on R.

We note that if there exists an evolution operator for (1) any solution $u:[s,b)\to X$, $a<s<b$, is unique, and similarly, if there exists such a solution u of (1), any evolution operator for (1) is unique. If $A:D(A)\to X$, we say that $A\in\mathscr{A}(\omega)$, $\omega\in R$, if

$$|x-y|-|x-y-\lambda(Ax-Ay)|\leqq\lambda\omega|x-y| \tag{3}$$

for $x,y\in D(A)$ and $\lambda>0$ with $\lambda\omega<1$ (cf. [1, § 1]).

It is not difficult to show that fixed x and y,

$$\lambda^{-1}(|x-y|-|x-y-\lambda(Ax-Ay)|)$$

is a nonincreasing function of λ for $\lambda>0$ (cf. [4, Lemma 5.1, p. 37]). Thus if (3) holds for some $\lambda_0>0$, it holds for all $\lambda>\lambda_0$.

PROPOSITION 3. *Let $A:D(A)\to X$ and $A\in\mathscr{A}(0)$. Let $\overline{D(A)}=X$ where $\overline{D(A)}$ is the closure of $D(A)$. Let $f:R\to X$ be continuous. Let $(I-\lambda_0A)D(A)=X$ for some $\lambda_0>0$; here and henceforth I denotes the identity operator on X. Then there exists an evolution operator $U(t,s)$ for (1) for $t\geqq s$.*

For a proof we may use Theorem 5.1 in [1]; we also use the well-known fact that if $A\in\mathscr{A}(0)$ and $(I-\lambda_0A)D(A)=X$, then $(I-\lambda A)D(A)=X$ for any $\lambda>0$.

PROPOSITION 4. *Let $A\in\mathscr{A}(\omega)$ and $\overline{D(A)}=X$ and f and g be continuous on R to X. Let $U_f(t,s)$ and $U_g(t,s)$ be evolution operators for (1) and an equation like (1) with f replaced by g. Then*

$$|U_f(t,s)x-U_g(t,s)x|\leqq\int_s^t|f(\tau)-g(\tau)|\exp\omega(t-\tau)\,d\tau\quad \textit{for}\ x\in X,\quad t\geqq s. \tag{4}$$

This is essentially Lemma 5.2 in [1].

PROPOSITION 5. *Let A satisfy the hypotheses of Proposition 4 and f be continuous on R to X. Let $U(t,s)$ be an evolution operator for (1). Then*

$$|U(t,s)x-U(t,s)y|\leqq|x-y|\exp\omega(t-s),\qquad t\geqq s\quad \textit{and}\quad x,y\in X. \tag{5}$$

For a proof, cf. Theorem 2.1 in [1] where actually weaker conditions on A are used.

3. Main results.

THEOREM 1. *Let A and f be functions as given in the introduction with $A \in \mathscr{A}(\omega)$ for some $\omega < 0$, and f a.p. Let $\overline{D(A)} = X$, and suppose there exists a $\lambda_0 > 0$ such that $(I - \lambda_0 A)D(A) = X$. Then each solution of* (1) *on* $[0, \infty)$ *is a.a.p. and there exists a unique generalized a.p. solution $\bar{u}(t)$ of* (1) *on R such that if $u(t)$ is any solution of* (1), *$u(t) - \bar{u}(t) \to 0$ as $t \to \infty$.*

Proof. Since $\omega < 0$, we have $A \in \mathscr{A}(0)$, and so the hypotheses of Proposition 3 hold. Let $\{t'_n\}$ be a sequence with $t'_n > 0$, $t'_n \to \infty$ as $n \to \infty$. Since f is a.p. there exists a subsequence $\{t_n\}$ of $\{t'_n\}$ and an a.p. function $g: R \to X$ such that

$$f(t + t_n) - g(t) \to 0 \quad \text{as } n \to \infty \tag{6}$$

uniformly for $t \in R$. We may clearly suppose $t_{n+1} > t_n$, $n = 1, 2, 3, \cdots$.

Let $U_f(t, s)$ be the evolution operator for (1) and $u(t)$ be a solution of (1) on $[0, \infty)$; let $u(0) = x$. Then

$$U_f(t + t_n, 0)x = u_n(t) \text{ is a solution of}$$

$$u' = Au + f_n(t) \tag{2n}$$

for $t \geqq -t_n$; here $f_n(t) = f(t + t_n)$. Let $U_{f_n}(t, s)$ be the evolution operator for (2n); then

$$U_f(t + t_n, 0)x = U_{f_n}(t, -t_n)x, \qquad t \geqq -t_n.$$

Since g is continuous on R, there exists an evolution operator $U_g(t, s)$ for $u' = Au + g(t)$ and a generalized solution

$$v_n(t) = U_g(t, -t_n)x, \qquad t \geqq -t_n$$

for this equation. Fix $t \geqq 0$ and m and n such that $m > n$. Then

$$|u_n(t) - u_m(t)| \leqq |u_n(t) - v_n(t)| + |v_n(t) - v_m(t)| + |v_m(t) - u_m(t)|. \tag{7}$$

By Proposition 4 with (6) and the condition $\omega < 0$, we have given $\varepsilon > 0$, there exists an N_ε so that $n > N_\varepsilon$ implies

$$|u_n(t) - v_n(t)| \leqq \int_{-t_n}^{t} |f_n(\tau) - g(\tau)| \exp \omega(t - \tau)\, d\tau \leqq \varepsilon, \tag{8}$$

and a similar result with n replaced by m.

Using Proposition 5, we have

$$\begin{aligned} |v_n(t) - v_m(t)| &\leqq |v_n(-t_n) - v_m(-t_n)| \exp \omega(t + t_n) \\ &\leqq |x - U_g(-t_n, -t_m)x| \exp \omega t_n. \end{aligned} \tag{9}$$

Using Proposition 4 again with $U_0(t, s)$ the evolution operator for $u' = Au$, we have

$$|U_g(-t_n, -t_m)x - U_0(-t_n, -t_m)x| \leqq B_1 \tag{10}$$

for some constant B_1; this follows since g is a.p. and hence bounded on R. By Proposition 5 and the fact that $u = 0$ is a solution of $u' = Au$ on R, we get

$$|U_0(-t_n, -t_m)x| \leqq |x|. \tag{11}$$

Using (10) and (11) in (9) yields

$$|v_n(t) - v_m(t)| \leqq (2|x| + B_1) \exp \omega t_n$$

and using this with (8) in (7) shows that for N_ε sufficiently large $n > N_\varepsilon$ and $m > n$ imply $|u_n(t) - u_m(t)| \leqq 3\varepsilon$. Thus $\{u_n(t)\}$ is Cauchy uniformly on $[0, \infty)$ and so converges uniformly on that interval. By Proposition 1 we conclude that $u(t)$ is a.a.p. and so by definition $u(t) - w(t) \to 0$ as $t \to \infty$ for some $w: R \to X$, w a.p.

Since f is a.p. there exists by Proposition 2 a sequence $\{\tau_n\}$, $\tau_n > 0$, $\tau_n \to \infty$ as $n \to \infty$, such that $f(t+\tau_n) - f(t) \to 0$ as $n \to \infty$; since w is also a.p., we may suppose $w(t+\tau_n) \to \bar{u}(t)$ as $n \to \infty$ uniformly for $t \in R$, where $\bar{u}(t)$ is a.p.

We show that $\bar{u}(t)$ is a generalized solution of (1). Let $t \in R$ be given and fix $t_0 \leqq t$. We have for n sufficiently large

$$|U_f(t, t_0)\bar{u}(t_0) - \bar{u}(t)| \leqq |U_f(t, t_0)\bar{u}(t_0) - u(t+\tau_n)| + |u(t+\tau_n) - \bar{u}(t)|. \tag{12}$$

Since $u(t+\tau_n) = U_{f_n}(t, -\tau_n)u(0)$ where $f_n(t) = f(t+\tau_n)$ and $U_{f_n}(t, -\tau_n) = U_{f_n}(t, t_0)U_{f_n}(t_0, -\tau_n)$ for n sufficiently large, we have

$$\begin{aligned} |U_f(t, t_0)\bar{u}(t_0) - u(t+\tau_n)| &\leqq |U_f(t, t_0)\bar{u}(t_0) - U_f(t, t_0)U_{f_n}(t_0, -\tau_n)u(0)| \\ &\quad + |U_f(t, t_0)U_{f_n}(t_0, -\tau_n)u(0) - U_{f_n}(t, t_0)U_{f_n}(t_0, -\tau_n)u(0)| \\ &\leqq |\bar{u}(t_0) - u(t_0+\tau_n)| + \sup\{|f_n(t) - f(t)|/|\omega| : t \in R\}; \end{aligned} \tag{13}$$

here we have used $\omega < 0$ and Propositions 4 and 5. Since

$$|\bar{u}(t_0) - u(t_0+\tau_n)| \leqq |\bar{u}(t_0) - w(t_0+\tau_n)| + |w(t_0+\tau_n) - u(t_0+\tau_n)| \tag{14}$$

and since $f_n(t) - f(t) \to 0$ as $n \to \infty$ uniformly on R, (13) shows that the first term on the right in (12) tends to zero as $n \to \infty$. But by (14) with t_0 replaced by t, we see that the second term on the right in (12) also approaches zero as $n \to \infty$. Thus $U_f(t, t_0)\bar{u}(t_0) = \bar{u}(t)$, $t \in R$; i.e., $\bar{u}$ is a generalized solution of (1) on R.

The uniqueness of $\bar{u}$ follows easily from Proposition 5; if $u_1(t)$ and $u_2(t)$ are distinct generalized a.p. solutions of (1) then $h(t) = u_1(t) - u_2(t)$ would be a.p. with $h(t_1) \neq 0$ for some $t_1 \in R$. But $h(t) \to 0$ as $t \to \infty$ by Proposition 5; since this is impossible, the uniqueness of $\bar{u}$ follows and our proof is complete.

THEOREM 2. *Let A and f be as in Theorem* 1, *and suppose X is reflexive and f is of bounded variation on each compact interval in R. Then for each $x \in D(A)$, there exists a solution $u(t)$ of* (1) *on* $[0, \infty)$ *such that $u(0) = x$ which is a.a.p. and approaches a unique generalized a.p. solution of* (1) *as $t \to \infty$.*

A proof of this theorem is an easy consequence of our Theorem 1 and Theorem 5.1(iv) in [1]; our hypotheses imply $\hat{D}(A) = D(A)$, where $\hat{D}(A)$ is defined in [1]. To show that we also have that A is closed, let $x_n \to x_0$ and $Ax_n \to y_0$ as $n \to \infty$, where $x_n \in D(A)$. Since $(I - \lambda A)^{-1}$ is defined and continuous on X (cf. [1, p. 9]) for any $\lambda > 0$,

$$(I - \lambda A)^{-1}(I - \lambda A)x_n \to (I - \lambda A)^{-1}(x_0 - \lambda y_0) \quad \text{as } n \to \infty.$$

So $(I - \lambda A)^{-1}(x_0 - \lambda y_0) = x_0$, and hence $Ax_0 = y_0$; i.e., A is closed.

It is clear that in Theorems 1 and 2, the solution $u(t)$ of (1) may be on any interval $[t_0, \infty)$ rather than on $[0, \infty)$; the latter was chosen to simplify the notation only.

We also have as a consequence of Proposition 5 that under our hypotheses, all generalized solutions of (1) are a.a.p. We stated Theorem 1 for strong solutions mainly to emphasize the fact that the definition of $U(t, s)$ as evolution operator for (1) involves strong solutions of (1).

4. An application. We consider the boundary value problem (2) stated in the introduction. We take $X = L^2[0, 1] = L^2$, the set of Lebesgue square integrable real-valued functions on $[0, 1]$ with the usual norm $|u| = \langle u, u\rangle^{1/2}$ in terms of the inner product:

$$\langle u, v\rangle = \int_0^1 u(\xi)v(\xi)\, d\xi, \qquad u, v \in L^2.$$

Since $\{L^2, \langle\ ,\ \rangle\}$ is a Hilbert space, X is now reflexive.

Let $\sigma: R \to R$ be continuously differentiable, $\sigma(0)=0$, and suppose there exist constant m and M such that $0<m \leqq \sigma'(u) \leqq M<\infty$ for $u \in R$. Let $r(t, \xi): R \times [0,1] \to R$ be in L^2 for each t and L^2-a.p.; i.e., a.p. in t in the L^2 norm.

We apply Theorems 1 and 2 of the previous section. Take $Au = (\sigma(u_\xi))_\xi$ with $D(A) = \{u \in L^2$: $u(0)=u(1)=0$, with $u(\xi)$ and $u'(\xi)$ absolutely continuous on $[0,1]$, $u''(\xi) \in L^2\}$.

It is well known that $L^2 = \overline{D(A)}$.

We next show that $A \in \mathscr{A}(\omega)$ for some $\omega < 0$. By direct calculations we find that for $u \neq v$, $u, v \in D(A)$,

$$\begin{aligned} &\lim_{\lambda \to 0+} (|u-v| - |u-v-\lambda(Au-Av)|)/\lambda|u-v| \\ &\quad = \langle u-v, Au-Av\rangle/|u-v|^2 \\ (15) \qquad &\quad = -\int_0^1 (u'(\xi)-v'(\xi))(\sigma(u'(\xi))-\sigma(v'(\xi)))\,d\xi \Big/ \int_0^1 (u(\xi)-v(\xi))^2\,d\xi \\ &\quad \leqq -m \int_0^1 (u'(\xi)-v'(\xi))^2\,d\xi \Big/ \int_0^1 (u(\xi)-v(\xi))^2\,d\xi \\ &\quad \leqq -m\pi; \end{aligned}$$

the last estimate follows from a result in [5, p. 182]; we can also easily get this with π replaced by 2 by using the Cauchy-Schwartz inequality. From (15) and the remark preceding the statement of Proposition 3, we have that $A \in \mathscr{A}(-m\pi)$.

We next show that there exists a $\lambda > 0$ such that $(I-\lambda A)D(A) = L^2$. Let $f \in L^2$ be given. Let $\{C, \|\ \|\}$ be the Banach space of real valued functions continuous on $[0,1]$ with norm $\|u\| = \sup\{|u(\xi)|: 0 \leqq \xi \leqq 1\}$. For fixed $\lambda > 0$ define $T_\lambda: C \to C$ as follows:

$$(T_\lambda u)(\xi) = \int_0^\xi \sigma^{-1}\left[\lambda^{-1}\int_0^\eta (u(s)-f(s))\,ds + c_u\right] d\eta, \qquad 0 \leqq \xi \leqq 1,$$

where c_u is the unique constant such that

$$(16) \qquad \int_0^1 \sigma^{-1}\left[\lambda^{-1}\int_0^\eta (u(s)-f(s))\,ds + c_u\right] d\eta = 0$$

and σ^{-1} denotes the inverse of σ. The fact that there exists a unique constant c_u follows since

$$(17) \qquad M^{-1} \leqq \sigma^{-1\prime}(u) \leqq m^{-1} \quad \text{for } u \in R.$$

From (16) we may also easily verify that

$$(18) \qquad |c_u - c_v| \leqq (M/m\lambda)\|u-v\|, \qquad u, v \in C;$$

we omit the details. Using (18) and (16), we see that for $\lambda > 0$ sufficiently large, T_λ is a contraction on C; let $\bar{u}$ be the unique fixed point of T_λ. It is easy to verify that $\bar{u} \in D(A)$ and satisfies $\lambda(\sigma(u'))' = u - f$. We have therefore $(I-\lambda A)D(A) = L^2$, as asserted. We have essentially proved the following theorem.

THEOREM 3. *Suppose $\sigma: R \to R$ is differentiable, $\sigma(0)=0$, and its derivative σ' satisfies $0<m \leqq \sigma'(u) \leqq M < \infty$ on R. Let $r(t,\cdot): R \to L^2$ be a.p. in t with respect to the L^2 norm. Then each solution $u(t,\xi)$ of* (2) *such that $u(t,\cdot) \in L^2$ for $t \geqq 0$ is a.a.p. and approaches a unique generalized solution of* (2) *which is L^2-a.p. as $t \to \infty$. If $r(t,\cdot)$ is of bounded variation for t in each compact interval of R, then for each $u_0 \in L^2$ with $u_0(0) = u_1(0) = 0$, u_0 and u_0' absolutely continuous on* $[0,1]$, *and $u_0'' \in L^2$, there exists a solution*

$u(t, \xi)$ of (2) *such that $u(0, \xi) = u_0(\xi)$, $0 \leqq \xi \leqq 1$, and $u(t, \xi)$ is a.a.p. and approaches a unique generalized a.p. solution of* (2).

The concept of generalized solution of (2) is in terms of the general definition given in § 2. It is not immediately clear that the generalized a.p. solution is a strong solution of (2) in the sense as defined in § 2. It is the purpose of future investigations to determine whether conditions exist sufficient for the existence of strong a.p. solutions of (1) as well as (2).

REFERENCES

[1] M. G. CRANDALL AND A. PAZY, *Nonlinear evolution equations in Banach spaces*, MRC Technical Summary Report #1191, 1972.

[2] A. M. FINK, *Almost periodic differential equations*, in Lecture Notes in Math., 377, Springer-Verlag, Berlin-Heidelberg-New York, 1974.

[3] B. M. LEVITAN AND V. V. ZHIKOV, *Almost Periodic Functions and Differential Equations*, English ed., Cambridge Univ. Press, Cambridge, 1982.

[4] R. H. MARTIN, JR., *Nonlinear Operators and Differential Equations in Banach Spaces*, Wiley-Interscience, New York, 1976.

[5] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge Univ. Press, Cambridge, 1967.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

009

APPLICATION OF ACCRETIVE OPERATORS THEORY TO THE RADIATIVE TRANSFER EQUATIONS*

B. MERCIER†

Abstract. Under some restrictive assumptions about opacities, we show that the radiative transfer equations have the form $(du/dt)+\mathscr{A}u+\mathscr{B}u=0$, where $\mathscr{A}$ is m-accretive and $\mathscr{B}$ is Lipschitz. Mathematically, this gives existence and uniqueness of the solution. We also show that the maximum principle applies. Assuming that opacities are decreasing with respect to temperature, we are able to prove that $\mathscr{B}$ itself is accretive. Finally, we derive from this analysis an algorithm for solving the radiative transfer equations which has some nice properties.

Key words. nonlinear operators, accretive operators, contraction semi-groups, radiative transfer, splitting algorithms, transport equation

AMS(MOS) subject classifications. 47, 47H06, 47H20, 85, 85A30

1. Nonlinear accretive operators theory. Let us assume we are given a real Banach space Y, which may not be reflexive. We denote by Y^* the dual space and by $\langle \cdot, \cdot \rangle$ the duality pairing between Y and Y^*. Let $\|\cdot\|$ and $\|\cdot\|_*$ denote the norms of Y and Y^*, respectively. We shall use the sets

$$s(u)=\{f\in Y^*: \|f\|_*=1, \langle u,f\rangle=\|u\|\},$$

for all $u\in Y$ such that $\|u\|\neq 0$. Let $\mathscr{A}$ denote a mapping from a nonempty subset $D(\mathscr{A})\in Y$ into Y. We shall say that $\mathscr{A}$ is an *accretive operator* if for each u_1, $u_2\in D(\mathscr{A})$ there exists an $f\in s(u_2-u_1)$ such that

$$\langle \mathscr{A}u_2-\mathscr{A}u_1, f\rangle \geqq 0.$$

In particular, if there exists a mapping $s_0: Y\rightarrow Y^*$ such that

$$s_0(u)\in s(u)$$

for all $u\in Y$, and such that

$$\langle \mathscr{A}u_2-\mathscr{A}u_1, s_0(u_2-u_1)\rangle \geqq 0 \tag{1}$$

for all u_1, $u_2\in D(\mathscr{A})$, then $\mathscr{A}$ is accretive. Furthermore, if, for some $\lambda>0$, the range of operator $\mathscr{T}+\lambda\mathscr{A}$ is equal to Y, where $\mathscr{T}$ denotes the identity of Y, we shall say that $\mathscr{A}$ is *m-accretive.* In such a case, operator $(\mathscr{T}+\lambda\mathscr{A})^{-1}$ is a contraction mapping from Y into Y (see [7]).

We now turn to the following abstract differential equation:

$$\begin{aligned}&\frac{du}{dt}+\mathscr{A}u=0\\ &u(0)=u_0\end{aligned}\qquad \text{where } u_0\in Y \text{ is given.} \tag{2}$$

* Received by the editors November 8, 1984; accepted for publication (in revised form) January 7, 1986.

† Commissariat à l'Energie Atomique, Centre d'Etudes de Limeil-Valenton, B.P. 27, 94190 Villeneuve-Saint-Georges, France.

The key idea for solving such an initial value problem is to introduce the following backward difference scheme:

$$(3)\qquad \begin{aligned} &\frac{u^{n,k+1}-u^{n,k}}{\lambda_n}+\mathscr{A}u^{n,k+1}=0,\\ &u^{n,0}=u_0 \end{aligned}$$

where $\lambda_n>0$ is a time-step that tends to zero as $n\to\infty$.

A fundamental result, due to Crandall and Liggett (see [6]), states that if $u_0\in\overline{D(\mathscr{A})}$ and $\mathscr{A}$ is m-accretive then the sequence of piecewise constant functions $u^n(t)$ defined by

$$u^n(t)=u^{n,k}\quad \text{for } k\lambda_n\leqq t<(k+1)\lambda_n$$

converges uniformly on $[0, T]$ to a limit function $u\in C^0([0, T]; Y)$, which may be considered as a *generalised solution* to problem (2). (It is actually called a mild solution (see [5]).) In particular, if problem (2) has a solution $v\in C^1$ $(0, T; X)$ such that $v(t)\in D(\mathscr{A})$ for a.e.t., then $u\equiv v$ (see [3]).

On the other hand, the mapping

$$S(t): u_0\to u(t)$$

is a contraction from $\overline{D(\mathscr{A})}$ into $\overline{D(\mathscr{A})}$ and the family $(S(t))_{t>0}$ is called the *semi-group* generated by $\mathscr{A}$.

Let us denote by F the (assumed nonempty) set of equilibrium points of $\mathscr{A}$, that is of elements $p\in\overline{D(\mathscr{A})}$ such that

$$(4)\qquad S(t)p=p$$

for all $t>0$. We have the following property:

$$(5)\qquad \|u(t)-p\|\searrow \text{ as } t\nearrow,$$

which means that the distance between the solution at time t and any equilibrium point decreases. In particular, the existence of equilibrium points ensures that the solution is bounded as time goes to infinity.

In many applications, $Y=L^1(X)$, where X is a closed subset of $\mathbb{R}^d$. In such a case, the set $s(u)$ is found to be the set of all functions $f\in L^\infty(X)$ such that

$$\begin{aligned} f(x)&=1 && \text{when } u(x)>0,\\ f(x)&=-1 && \text{when } u(x)<0,\\ |f(x)|&\leqq 1 && \text{when } u(x)=0. \end{aligned}$$

It is usually convenient to choose s_0 such that

$$(6)\qquad (s_0(u))(x)=\begin{cases} 1 & \text{when } u(x)>0,\\ -1 & \text{when } u(x)<0,\\ 0 & \text{when } u(x)=0. \end{cases}$$

We note that the sum $\mathscr{A}+\mathscr{B}$ of two accretive operators need not be accretive (see [5]). However, this is the case (see [5]) in the case where one of the operators, say $\mathscr{B}$, is continuous. This is also the case, obviously, when both operators satisfy condition (1) with the *same* s_0.

About the sum $\mathscr{A}+\mathscr{B}$ of two m-accretive operators, we shall need a result which states that if $\mathscr{A}$ is m-accretive and $\mathscr{B}$ is Lipschitz, i.e. there exists $\omega > 0$ such that

$$\|\mathscr{B}u_2-\mathscr{B}u_1\| \leqq \omega\|u_2-u_1\|$$

for all u_1, $u_2 \in Y$, then $\mathscr{A}+\mathscr{B}$ is m-accretive. This result is standard and has a simple proof; as $\mathscr{A}+\mathscr{B}$ is accretive, we need only prove that

$$u+\lambda\mathscr{A}u+\lambda\mathscr{B}u=f$$

has a solution for any $f \in X$. For this we note that

$$u \rightarrow (I+\lambda\mathscr{A})^{-1}(f-\lambda\mathscr{B}u)$$

is a strictly contractive mapping from X into itself for $\lambda > 0$ small enough.

Remark 1. When $\mathscr{B}$ is Lipschitz but not accretive, then the Crandall-Liggett result still applies. However semigroup $S(t)$ generated by $\mathscr{A}+\mathscr{B}$ is of type ω, that is

$$\|S(t)u_2-S(t)u_1\| \leqq e^{\omega t}\|u_2-u_1\|$$

where ω is the Lipschitz constant of $\mathscr{B}$.

2. Problem to be solved. Let us consider a continuous medium, assumed to be a motionless gas, in a subdomain $X \subset \mathbb{R}^3$, with boundary ∂X. In the following, we shall denote by x a generic point in the domain X and Ω a generic direction on the unit sphere S^2.

The interaction of the gas with radiation is described by the energy equation

$$\frac{\partial \varepsilon}{\partial t}+\int_0^\infty K(\nu, \varepsilon)\, d\nu \int_{S^2} (B(\nu, \varepsilon)-I)\, d\Omega=0 \tag{7}$$

where

$K \equiv K(\nu, \varepsilon)$ denotes the opacity,

$B \equiv B(\nu, \varepsilon)$ is Planck's function,

$I \equiv I(x, \Omega, \nu, t)$ is the specific intensity of radiation at frequency ν in direction Ω.

The radiation is assumed to be travelling at the speed of light c so that I satisfies the following transfer equation

$$\frac{1}{c}\frac{\partial I}{\partial t}+\Omega \cdot \frac{\partial I}{\partial x}+K(I-B)=\mathscr{D}I \tag{8}$$

where $\mathscr{D}$ denotes an integral operator defined to take into account Thomson scattering: we have

$$(\mathscr{D}I)(x, \Omega, \nu)=\int_{S^2} (K_d(\Omega', \Omega)I(x, \Omega', \nu)-K_d(\Omega, \Omega')I(x, \Omega, \nu))\, d\Omega' \tag{9}$$

where $K_d((\Omega', \Omega) \geqq 0$ denotes the scattering cross section assumed to be independent of ε and ν.

The unknown function I is subject to a *boundary condition*:

$$I(x, \Omega, \nu, t)=h(x, \Omega, \nu, t) \tag{10}$$

for $x \in \partial X$, $\nu > 0$, $t > 0$ and $\Omega \in S^2$ such that $\Omega \cdot n(x) < 0$, where $n(x)$ denotes the unit normal vector to X, directed outward. The meaning of boundary condition (10) is that the inflow specific intensity h is assumed to be known.

Finally, we supplement the system (7), (8), (10) with some *initial conditions*

$$\begin{aligned} &I(x,\Omega,\nu,0)=I_0(x,\Omega,\nu), \quad x\in X, \quad \Omega\in S^2, \quad \nu>0,\\ &\varepsilon(x,0)=\varepsilon_0(x), \qquad x\in X, \end{aligned} \tag{11}$$

where I_0 and ε_0 are given functions.

3. Choice of a functional framework. We shall introduce a Banach space Y for the couple of unknown functions

$$u(t)=\{\varepsilon(t), I(t)\}.$$

We choose

$$Y=L^1(X)\times L^1(X\times S^2\times(0,\infty)) \tag{12}$$

with

$$\|u\|\equiv\|\varepsilon\|_1+\|I\|_1 \quad \text{whenever } u\equiv\{\varepsilon, I\}. \tag{13}$$

We denote by $\|\cdot\|_1$ either the norm of $L^1(X)$ defined as

$$\|\varepsilon\|_1\equiv\int_X |\varepsilon(x)|\,dx$$

or the norm of $L^1(X\times S^2\times(0,\infty))$ defined by

$$\|I\|_1\equiv\int_X\int_{S^2}\int_0^\infty |I(x,\Omega,\nu)|\,dx\,d\Omega\,d\nu.$$

The choice (13) for the norm $\|\cdot\|$ is a natural one. Indeed, if (as they should) both $\varepsilon(t)$ and $I(t)$ are positive, then $\|u(t)\|$ is the total energy (material + radiation).

Next we shall introduce some operators $\mathscr{A}$ and $\mathscr{B}$ on space Y so that problem (7)-(11) is equivalent to

$$\frac{du}{dt}+\mathscr{A}u+\mathscr{B}u=0, \qquad u(0)=u_0 \tag{14}$$

where $u_0=\{\varepsilon_0, I_0\}$. First we define operator $\mathscr{A}$ as follows:

$$\begin{aligned} &\mathscr{A}\{\varepsilon, I\}=\left\{0, \Omega\cdot\frac{\partial I}{\partial x}-\mathscr{D}I\right\},\\ &D(\mathscr{A})=\left\{u\in X\colon u=\{\varepsilon, I\}, \Omega\cdot\frac{\partial u}{\partial x}\in L^1, I \text{ satisfies } (10)\right\}. \end{aligned} \tag{15}$$

Note that we need a trace theorem, which is proved in [4], to give meaning to the boundary condition (10). We shall assume that the inflow specific intensity (h) does not depend on t, so that $\mathscr{A}$ is independent of t. We note that $\mathscr{A}$ is unbounded.

Next, we define operator $\mathscr{B}$ as follows: for $u\in X$, $u=\{\varepsilon, I\}$, we let

$$\mathscr{B}u=\left\{-\int_0^\infty d\nu\int_{S^2} q\,d\Omega, q\right\} \tag{16}$$

where

$$q\equiv q(x,\Omega,\nu)=K(\nu,\varepsilon(x))(\varphi(\nu, I(x,\Omega,\nu))-B(\nu,\varepsilon(x)) \tag{17}$$

and

$$\varphi(\nu, I) = \max(0, \min(I, B(\nu, M))); \tag{18}$$

finally, M is a given constant.

With such a definition of operator $\mathscr{B}$, we note first that (14) coincides with the original problem (7)-(11) whenever

$$0 \leqq I(x, \Omega, \nu, t) \leqq B(\nu, M) \tag{19}$$

for $\nu > 0$, $x \in X$, $\Omega \in S^2$ and $t > 0$.

We shall prove in § 5 that the maximum principle applies to system (7)-(11) so that (19) holds for some $M > 0$ together with

$$0 \leqq \varepsilon(x, t) \leqq M. \tag{20}$$

The function $\varphi(\nu, I)$ defined in (18) is obtained from I with an appropriate truncation; if we require $\mathscr{B}$ to be Lipschitzian, then we need φ to be bounded, and if we want $\mathscr{B}$ to be accretive, we need φ to be positive, as we shall see in § 6.

4. Properties of operators $\mathscr{A}$ and $\mathscr{B}$. Let us first recall some facts about operator $\mathscr{A}$. We note that $D(\mathscr{A})$ is dense in Y (see [2]). To prove that $\mathscr{A}$ is accretive we shall need the following lemmas.

LEMMA 1. *Let W denote the space of functions I in L^1 ($\equiv L^1(X \times S^2 \times (0, \infty))$) such that $\Omega \cdot \nabla I \in L^1$. Let $\beta : \mathbb{R} \to \mathbb{R}$ be a C^1 function such that $\beta(0) = 0$ and $\beta' \in L^\infty(\mathbb{R})$. Then, for any $I \in W$, $\beta(I) \in W$ and*

$$\Omega \cdot \frac{\partial}{\partial x} \beta(I) = \beta'(I)\left(\Omega \cdot \frac{\partial I}{\partial x}\right) \quad \text{in } L^1. \tag{21}$$

To prove the result, one introduces a sequence $(I_n)_n$ of smooth functions converging to $I \in W$ (the existence of such a sequence relies on a density result; see [4]). For a given smooth I_n (21) obviously holds. The Lebesgue dominated convergence theorem is used to prove that (21) holds in the limit, after extraction of a suitable subsequence. The details are left to the reader.

LEMMA 2. *For any $I \in W$, then $|I| \in W$ and*

$$\Omega \cdot \frac{\partial}{\partial x} |I| = s_0(I)\Omega \cdot \frac{\partial I}{\partial x}.$$

Proof. For given $n \in \mathbb{N}$ we introduce the function $\beta_n : \mathbb{R} \to \mathbb{R}$ defined by

$$\beta_n(\xi) = \begin{cases} n\dfrac{\xi^2}{2} & \text{if } |\xi| \leqq \dfrac{1}{n}, \\ |\xi| - \dfrac{1}{2n} & \text{otherwise.} \end{cases}$$

We check that

$$\begin{aligned} &\text{(i)} \quad \beta_n(\xi) \to |\xi|; \\ &\text{(ii)} \quad \beta_n'(\xi) \to s_0(\xi) \end{aligned} \tag{22}$$

for all $\xi \in \mathbb{R}$ when $n \to \infty$ and that

$$|\beta_n'(\xi)| \leqq 1 \tag{23}$$

for all $\xi \in \mathbb{R}$.

Let $I \in W$ be given. From Lemma 1 we know that $\beta_n(I) \in W$ and that

$$f_n \equiv \Omega \cdot \frac{\partial}{\partial x} \beta_n(I) = \beta_n'(I)\left(\Omega \cdot \frac{\partial I}{\partial x}\right).$$

From (22)(ii) and (23) we see with Lebesgue's theorem that

$$f_n \to f \equiv s_0(I)\left(\Omega \cdot \frac{\partial I}{\partial x}\right) \quad \text{in } L^1.$$

On the other hand, using Lebesgue's theorem once again, from (22)(i) and (23) we prove that

$$\beta_n(I) \to |I| \quad \text{in } L^1.$$

If $\varphi \in C_0^\infty$ $(X \times S^2 \times (0, \infty))$ from the definition of the derivative in the distribution sense, we know that

$$\langle f_n, \varphi \rangle = -\left\langle \beta_n(I), \Omega \cdot \frac{\partial \varphi}{\partial x} \right\rangle.$$

In the limit $n \to \infty$, we get that

$$\langle f, \varphi \rangle = -\left\langle |I|, \Omega \cdot \frac{\partial \varphi}{\partial x} \right\rangle.$$

Therefore $\Omega \cdot (\partial/\partial x)|I| \in L^1$ and $f = \Omega \cdot (\partial/\partial x)|I|$ in L^1. QED

PROPOSITION 1. *Operator $\mathscr{A}$ is m-accretive.*

Proof. To show that $\mathscr{A}$ is accretive, we shall choose

$$s_0(u) = \{s_0(\varepsilon), s_0(I)\}$$

where $s_0(\varepsilon) \in L^\infty$ (X) is defined as in (6), and $s_0(I) \in L^\infty$ $(X \times S^2 \times (0, \infty))$ is defined in an analogous way.

Let $u_\alpha = \{\varepsilon_\alpha, I_\alpha\} \in D(\mathscr{A})$, $\alpha = 1, 2$. We have

$$\langle \mathscr{A}u_2 - \mathscr{A}u_1, s_0(u_2 - u_1) \rangle = \left\langle \Omega \cdot \frac{\partial}{\partial x}(I_2 - I_1) - \mathscr{D}(I_2 - I_1), s_0(I_2 - I_1) \right\rangle.$$

To prove that $\mathscr{A}$ is accretive, it suffices to prove separately that the advection operator

$$\mathscr{L} \equiv \Omega \cdot \frac{\partial}{\partial x}$$

with domain

$$D(\mathscr{L}) = \left\{ I \in L^1 : \Omega \cdot \frac{\partial I}{\partial x} \in L^1, I \text{ satisfies } (10) \right\}$$

and operator $(-\mathscr{D})$ are accretive.

To prove that $\mathscr{L}$ is accretive, we notice that from Lemma 2

$$\left[\Omega \cdot \frac{\partial}{\partial x}(I_2 - I_1)\right] s_0(I_2 - I_1) = \Omega \cdot \frac{\partial}{\partial x}|I_2 - I_1| \quad \text{in } L^1$$

provided that I_1 and $I_2 \in D(\mathscr{L})$ and that function $I \equiv |I_2 - I_1|$ satisfies

$$\Omega \cdot \frac{\partial I}{\partial x} \in L^1$$

and vanishes on the inflow boundary $\Gamma_- \times (0,\infty)$ (where $\Gamma_\pm$ denotes the subset of $\partial X \times S^2$ made with those couples $\{x, \Omega\}$ such that $\Omega \cdot n(x) \gtrless 0$).

Applying Green's formula (see [4] for a justification)

$$\int_0^\infty d\nu \int_{X\times S^2} \Omega \cdot \frac{\partial}{\partial x} I \, dx \, d\Omega = \int_0^\infty d\nu \int_{\Gamma_+} (\Omega \cdot n) I \, d\Gamma_+$$

to $I = |I_2 - I_1|$, we see that

$$\left\langle \Omega \cdot \frac{\partial}{\partial x}(I_2 - I_1), s_0(I_2 - I_1) \right\rangle \geqq 0 \text{ so that } \mathcal{L} \text{ is accretive.}$$

Since operator $\mathcal{D}$ is linear, it is sufficient to prove that

$$\langle \mathcal{D} I, s_0(I) \rangle \leqq 0. \tag{24}$$

We have

$$\int_{S^2} \left(\int_{S^2} (K_d(\Omega', \Omega) I(\Omega') - K_d(\Omega, \Omega') I(\Omega)) \, d\Omega' \right) s_0(I(\Omega)) \, d\Omega$$

$$\leqq \int_{S^2} \int_{S^2} (K_d(\Omega', \Omega) |I(\Omega')| - K_d(\Omega, \Omega') |I(\Omega)|) \, d\Omega' \, d\Omega = 0.$$

Hence from (9) we get (24).

Finally, since the problem

$$I + \lambda \Omega \cdot \frac{\partial I}{\partial x} = f,$$

with I subject to boundary condition (9), can be solved for any $f \in L^1(X)$, provided h satisfies

$$\int_0^\infty d\nu \int_{\Gamma_-} (\Omega \cdot n) h \, d\Gamma_- < +\infty,$$

which we shall assume (see [4]), operator $\mathcal{L}$ is m-accretive.

As $\mathcal{D}$ is Lipschitz, we have that $\mathcal{L} - \mathcal{D}$ is m-accretive. QED

We now turn to operator $\mathcal{B}$. To get Lipschitz continuity of $\mathcal{B}$, we shall require some regularity on the functions

$$\varepsilon \to K(\nu, \varepsilon), \qquad \varepsilon \to B(\nu, \varepsilon).$$

Actually, we know that $\varepsilon = \varepsilon(T)$ depends on the temperature T through an *equation of state* ($\varepsilon = \rho c_v T$ in the case of a perfect gas), and that

$$B(\nu, \varepsilon(T)) = \frac{2h\nu^3}{c^2} (e^{h\nu/kT} - 1)^{-1} \tag{25}$$

where h is Planck's constant, and k Boltzmann's constant. Then the function $\varepsilon \to B(\nu, \varepsilon)$ is regular provided that the equation of state $T \to \varepsilon(T)$ is regular.

More precisely, we shall assume that there exists a constant C positive such that

$$|K(\nu, \varepsilon_1) - K(\nu, \varepsilon_2)| \leqq C |\varepsilon_1 - \varepsilon_2| \quad \textit{for all } \nu > 0, \quad \varepsilon_1, \varepsilon_2 \in \mathbb{R}, \tag{H1}$$

$$0 \leqq K(\nu, \varepsilon) \leqq C \quad \textit{for all } \nu > 0, \quad \varepsilon \in \mathbb{R} \tag{H2}$$

and that there exists $\beta_0, \beta_1 > 0$ such that

$$0 < \beta_0 \leqq \frac{d\varepsilon}{dT} \leqq \beta_1. \tag{H3}$$

Furthermore, for technical reasons we shall modify function B for $\varepsilon > M$ (this is without consequence if we prove that (20) holds).

Let T_M be such that $\varepsilon(T_M) = M$, for $\varepsilon > M$ we choose

$$B(\nu, \varepsilon) = B(\nu, \varepsilon(T_M)).$$

We check that, with such a modification, we have

$$\int_0^\infty B(\nu, \varepsilon(T))\, d\nu = a \min (T^4, T_M^4) \tag{26}$$

where a is some constant.

PROPOSITION 2. *Assume* (H1) (H2) (H3); *then operator* $\mathcal{B}$ *defined in* (16)-(18) *is Lipschitzian.*

Proof. Let $u_\alpha = \{\varepsilon_\alpha, I_\alpha\} \in Y$, $\alpha = 1, 2$ be given, and

$$q_\alpha \equiv K(\nu, \varepsilon_\alpha)(\varphi(\nu, I_\alpha) - B(\nu, \varepsilon_\alpha)).$$

As

$$\mathcal{B}u_\alpha \equiv \left\{ -\int_0^\infty d\nu \int_{S^2} q_\alpha\, d\Omega, q_\alpha \right\}$$

for $\alpha = 1, 2$, we have

$$\|\mathcal{B}u_2 - \mathcal{B}u_1\| \leqq 2 \int_X dx \int_0^\infty d\nu \int_{S^2} |q_2 - q_1|\, d\Omega.$$

We notice that

$$\begin{aligned} q_2 - q_1 &= K(\varepsilon_2)(\varphi(I_2) - B(\varepsilon_2)) - K(\varepsilon_1)(\varphi(I_1) - B(\varepsilon_1)) \\ &= K(\varepsilon_2)(\varphi(I_2) - \varphi(I_1)) + (K(\varepsilon_2) - K(\varepsilon_1))\varphi(I_1) + K(\varepsilon_2)(B(\varepsilon_1) - B(\varepsilon_2)) \\ &\quad + (K(\varepsilon_1) - K(\varepsilon_2))B(\varepsilon_1). \end{aligned}$$

From (H2) and (H1) we get

$$|q_1 - q_2| \leqq C|I_2 - I_1| + C(B(M) + B(\varepsilon_1))|\varepsilon_2 - \varepsilon_1| + C|B(\varepsilon_1) - B(\varepsilon_2)|. \tag{27}$$

Applying (26) we get

$$\begin{aligned} \int_0^\infty |B(\nu, \varepsilon(T_1)) - B(\nu, \varepsilon(T_2))|\, d\nu &= a[\min (T_1^4, T_M^4) - \min (T_2^4, T_M^4)] \\ &\leqq 4aT_M^3(T_1 - T_2) \leqq C'(\varepsilon(T_1) - \varepsilon(T_2)) \end{aligned}$$

where $C' = 4aT_M^3/\beta_0$.

Finally, when we integrate (27) for $x \in X$, $\Omega \in S^2$, $\nu > 0$, we obtain

$$\|\mathcal{B}u_2 - \mathcal{B}u_1\| \leqq 2C\|I_2 - I_1\|_1 + 16\pi C\left(\int_0^\infty B(M)\, d\nu\right)\|\varepsilon_2 - \varepsilon_1\|_1 + 8\pi CC'\|\varepsilon_1 - \varepsilon_2\|_1.$$

QED

Applying the Crandall-Liggett Theorem (see Remark 1), we obtain the following existence result.

COROLLARY 1. *If assumptions* (H1) (H2) (H3) *hold, then for any given* $t_0 > 0$ *problem* (14) *has a unique solution*

$$u \in C^0([0, t_0]; Y).$$

About the asymptotic behavior of solution $u(t)$ as time $t\to\infty$, we have no information as yet. Note that in the general case of an m-accretive $\mathcal{A}$ perturbed by a Lipschitz $\mathcal{B}$, the solution $u(t)$ might grow exponentially as $t\to\infty$.

Remark 2. According to inequalities (19), (20) we can restrict ourselves to an energy interval $[0, M]$. Then assumptions (H1) and (H2) need to hold only for ε_1, ε_2, $\varepsilon\in[0, M]$.

Indeed, outside this interval we can modify opacity $K(\nu, \varepsilon)$ in order to get the desired assumptions.

5. Maximum principle. To solve the original problem ((7)-(11)), we have introduced an operator $\mathcal{B}$ which involves $\varphi(I)$ instead of I, where $\varphi(I)$ is a truncation of I defined in (18). We needed such a truncation to get a Lipschitz $\mathcal{B}$. However we have yet to prove that the solution u of (14) is also a solution of the original problem. We proceed in the following way.

Assume that the initial energy ε^0 satisfies

$$N'\leqq\varepsilon^0(x)\leqq N \tag{28}$$

for all $x\in X$, with $0\leqq N'\leqq N\leqq M$. We also assume that

$$B(\nu, N')\leqq I^0(x, \Omega, \nu)\leqq B(\nu, N) \tag{29}$$

and

$$B(\nu, N')\leqq h(x, \Omega, \nu)\leqq B(\nu, N) \tag{30}$$

for all $x\in X$, $\Omega\in S^2$, $\nu>0$. Then, we prove that the solution $u=\{\varepsilon, I\}$ of (14) satisfies

$$N'\leqq\varepsilon(x, t)\leqq N \tag{31}$$

for almost every $x\in X$, $t>0$ and

$$B(\nu, N')\leqq I(x, \Omega, \nu, t)\leqq B(\nu, N) \tag{32}$$

for almost every $x\in X$, $\Omega\in S^2$, $\nu>0$ and $t>0$. In such a case, we have $\varphi(I)=I$ so that $u=\{\varepsilon, I\}$ is also a solution of the original problem.

To prove (31), (32) we shall require K_d (appearing in definition (9) of operator $\mathcal{D}$) to be symmetric:

$$K_d(\Omega, \Omega')=K_d(\Omega', \Omega)\quad\text{for all }\Omega,\quad \Omega'\in S^2. \tag{33}$$

LEMMA 3. *Let $v=\{e, J\}\in Y$ be such that*

$$N'\leqq e(x)\leqq N$$

for (almost) every $x\in X$, and

$$B(\nu, N')\leqq J(x, \Omega, \nu)\leqq B(\nu, N)$$

for (almost) every $x\in X$, $\Omega\in S^2$, $\nu>0$. Let $u=\{\varepsilon, I\}$ denote the solution of

$$u+\lambda(\mathcal{A}u+\mathcal{B}u)=v.$$

Then, one has

$$N'\leqq\varepsilon(x)\leqq N,$$

$$B(\nu, N')\leqq I(x, \Omega, \nu)\leqq B(\nu, N)$$

for (almost) every $x\in X$, $\Omega\in S^2$, $\nu>0$.

Proof. We shall use operator $s_+: L^1(X)\to L^\infty(X)$ defined as follows, if $\varepsilon\in L^1(\mathbb{R})$, $s_+(\varepsilon)$ denotes the function equal to one where $\varepsilon>0$, and equal to zero where $\varepsilon\leqq0$.

We note that ε and I satisfy

$$(34)\quad \varepsilon - N + \lambda \int_0^\infty K(\varepsilon) \int_{S^2} ((B(\varepsilon) - B(N)) - (\varphi(I) - B(N)))\, d\Omega\, d\nu = e - N,$$

$$(35)\quad \begin{aligned} I - B(N) + \lambda\Omega \cdot \frac{\partial}{\partial x}(I - B(N)) + \lambda K(\varepsilon)(\varphi(I) - B(N)) - (B(\varepsilon) - B(N)) \\ = \lambda\mathcal{D}(I - B(N)) + J - B(N), \end{aligned}$$

since

$$\frac{\partial}{\partial x} B(N) = 0$$

and, by virtue of (33)

$$\mathcal{D}(B(N)) = B(N) \int_{S^2} (K_d(\Omega', \Omega) - K_d(\Omega, \Omega'))\, d\Omega' = 0.$$

We multiply (34) by $v \equiv s_+(\varepsilon - N)$ and integrate for $x \in X$.

On the other hand, we multiply (35) by $w = s_+(I - B(N))$ and integrate for $x \in X$, $\Omega \in S^2$ and $\nu \in (0, \infty)$.

We note that

$$(\varepsilon - N)s_+(\varepsilon - N) = (\varepsilon - N)_+$$

and since $\varepsilon \to B(\nu, \varepsilon)$ is increasing (see (25) and (H3))

$$(B(\varepsilon) - B(N))s_+(\varepsilon - N) = (B(\varepsilon) - B(N))_+.$$

On the other hand,

$$(I - B(N))s_+(I - B(N)) = (I - B(N))_+$$

and (since $M \geqq N$)

$$(\varphi(I) - B(N))s_+(I - B(N)).$$

We obtain by addition

$$\begin{aligned} &\|(\varepsilon - N)_+\|_1 + \|(I - B(N))_+\|_1 + \lambda\left\langle \Omega \cdot \frac{\partial}{\partial x}(I - B(N)), w \right\rangle \\ &+ \lambda \int_X \int_0^\infty K(\varepsilon) \int_{S^2} ((B(\varepsilon) - B(N))_+ - w(B(\varepsilon) - B(N)))\, d\Omega\, d\nu\, dx \\ &+ \lambda \int_X dx \int_0^\infty K(\varepsilon) \int_{S^2} [(\varphi(I) - B(N))_+ - v(\varphi(I) - B(N))]\, d\Omega\, d\nu \\ &= \lambda\langle \mathcal{D}(I - B(N)), w\rangle + \langle e - N, v\rangle + \langle J - B(N), w\rangle. \end{aligned}$$

We note that

$$(B(\varepsilon) - B(N))_+ - w(B(\varepsilon) - B(N)) = (1 - w)(B(\varepsilon) - B(N))_+ + w(B(\varepsilon) - B(N))_- \geqq 0$$

since $0 \leqq w \leqq 1$. Similarly,

$$(\varphi(I) - B(N))_+ - v(\varphi(I) - B(N)) = (1 - v)(\varphi(I) - B(N))_+ + v(\varphi(I) - B(N))_- \geqq 0$$

since $0 \leqq v \leqq 1$.

On the other hand, we have (using a result similar to the one proved in Lemma 2)

$$\Omega\cdot\frac{\partial}{\partial x}(I-B(N))s_+(I-B(N))=\Omega\cdot\frac{\partial}{\partial x}(I-B(N))_+;$$

therefore, by integration by parts, we obtain

$$\int_X dx\int_{S^2}\Omega\cdot\frac{\partial}{\partial x}(I-B(N))_+\,d\Omega=\int_{\Gamma_+}(\Omega\cdot n)(I-B(N))_+\,d\Gamma_+\geqq 0$$

(note that $(I-B(N))_+$ vanishes on Γ_- by assumption (see (30)).

Let $\tilde I\equiv I-B(N)$. Then we have

$$\int_{S^2}\left(\int_{S^2}(K_d(\Omega',\Omega)\tilde I(\Omega')-K_d(\Omega,\Omega')\tilde I(\Omega))\,d\Omega'\right)s_+(\tilde I(\Omega))\,d\Omega$$

$$=\int_{S^2\times S^2}K_d(\Omega',\Omega)(\tilde I_+(\Omega')-\tilde I_-(\Omega'))s_+(\tilde I(\Omega))\,d\Omega'\,d\Omega$$

$$-\int_{S^2\times S^2}K_d(\Omega,\Omega')\tilde I_+(\Omega)\,d\Omega\,d\Omega'$$

$$\leqq-\int_{S^2\times S^2}K_d(\Omega',\Omega)\tilde I_-(\Omega')\,d\Omega'\,d\Omega\leqq 0;$$

hence $\langle\mathcal{D}\tilde I, s_+(\tilde I)\rangle\leqq 0$.

Finally, we have

$$\|(\varepsilon-N)_+\|_1+\|(I-B(N))_+\|_1\leqq 0,$$

which proves that

$$\varepsilon(x)\leqq N\quad\text{for a.e. } x\in X,$$

$$I(x,\Omega,\nu)\leqq B(\nu,N)\quad\text{for a.e. } x\in X,\quad \Omega\in S^2,\quad \nu>0.$$

In an analogous way, we could prove that

$$\|(\varepsilon-N')_-\|_1+\|(I-B(N'))_-\|_1\leqq 0.$$ QED

We are now able to prove the main result.

THEOREM 1. *Assume that initial solution* $u_0=\{\varepsilon_0,I_0\}$ *satisfies* (28), (29) *and that boundary condition h satisfies* (30). *Then the solution* $u(t)=\{\varepsilon(t),I(t)\}$ *of problem* (14) *satisfies* (31), (32). *Furthermore, it is also the solution of the original problem* (7)-(11).

Proof. From the Crandall-Liggett Theorem (see § 1) we know that solution $u(t)$ is the uniform limit as $n\to\infty$ of the piecewise constant functions $u^n(t)$ defined by

$$u^n(t)=u^{n,k}\quad\text{for } k\lambda_n\leqq t\leqq(k+1)\lambda_n$$

where $u^{n,k}$ is defined inductively by

$$\frac{u^{n,k+1}-u^{n,k}}{\lambda_n}+\mathcal{A}u^{n,k+1}+\mathcal{B}u^{n,k+1}=0.$$

Let $u^{n,k}\equiv\{\varepsilon^{n,k},I^{n,k}\}$. If we have

(36) $$N'\leqq\varepsilon^{n,k}(x)\leqq N$$

for (almost every) $x\in X$, and

(37) $$B(\nu,N')\leqq I^{n,k}(x,\Omega,\nu)\leqq B(\nu,N)$$

for (almost every) $x \in X$, $\Omega \in S^2$, $\nu > 0$, then, applying Lemma 3 with $v = u^{n,k}$, $u = u^{n,k+1}$ and $\lambda = \lambda_n$, we have (36), (37) at index $k+1$. Since (36) and (37) hold for $k = 0$, then they hold for all $k > 0$. Since they also hold for all $n > 0$, making $n \to \infty$, we get (31) and (32). As noticed above, since $0 \leqq N' \leqq N \leqq M$, $u(t)$ is also a solution of (7) to (11). QED

The maximum principle we just proved has 3 applications:

(a) It shows that solution $u(t)$ of (14) is actually a solution of the original problem;

(b) It shows that solution $u(t)$ remains bounded in L^∞ as $t \to \infty$;

(c) It shows that $u(t)$ is positive.

Remark 3. Assume that $u + \lambda(\mathscr{A}u + \mathscr{B}u) = v$ with $v = \{e, J\}$ and $u = \{\varepsilon, I\}$ as in Lemma 3. We easily prove that, if $h = 0$,

$$\int_X \varepsilon\, dx + \int_X dx \int_0^\infty d\nu \int_{S^2} I\, d\Omega \leqq \int_X e\, dx + \int_X dx \int_0^\infty d\nu \int_{S^2} J\, d\Omega.$$

Using an argument similar to the one used for Theorem 1, we have

$$\int_X \varepsilon(t)\, dx + \int_X dx \int_0^\infty d\nu \int_{S^2} I(t)\, d\Omega \leqq \|u_0\|,$$

which shows that the energy is decreasing (or conserved if $X = \mathbb{R}^3$).

Remark 4. In a way similar to Remark 2, we note that assumptions (H1) and (H2) need to hold only for ε_1, ε_2, $\varepsilon \in [N', N]$. If opacities are infinite for $\varepsilon = 0$, as they should be, then our existence result applies only for $N' > 0$.

6. Accretiveness of operator $\mathscr{B}$. We shall now prove that operator $\mathscr{B}$ itself is accretive under some new assumptions (see (H4) and (H5) below). Actually, the accretiveness of $\mathscr{B}$ is not very useful information if we assume (H1) and (H2) since we already have existence, uniqueness, and uniform L^∞ bounds for $t \to \infty$. However, besides mathematical curiosity, the accretiveness of $\mathscr{B}$ is of fundamental importance for the case where $K(\nu, \varepsilon)$ is infinite either for $\nu = 0$ (see [10]) or for $\varepsilon = 0$, for which we refer to Golse and Perthame [8].

We assume the following:

(H4) for all $\nu > 0$, function $\varepsilon \to K(\nu, \varepsilon)$ is decreasing,

(H5) for all $\nu > 0$, function $\varepsilon \to K(\nu, \varepsilon)B(\nu, \varepsilon)$ is increasing.

Assumptions (H4) and (H5) should hold for all $\varepsilon \in \mathbb{R}$; however, as noticed in Remark 4, they need only be satisfied in some interval $[N', N]$ related to the initial and boundary data.

Remark 5. To understand the significance of assumptions (H4) and (H5) we note that if I is given, and is independent of t, then (7) is just an autonomous ordinary differential equation depending on a parameter $x \in X$.

Let

$$\psi(\varepsilon) \equiv 4\pi \int_0^\infty K(\nu, \varepsilon)(B(\nu, \varepsilon) - E(\nu))\, d\nu$$

where

$$E \equiv \frac{1}{4\pi} \int_{S^2} I\, d\Omega.$$

Then (7) can be written as

$$\frac{d\varepsilon}{dt} + \psi(\varepsilon) = 0$$

(for given $x \in X$). If we assume that ψ is increasing then $\psi(\varepsilon^*) = 0$ has a unique solution, and $\varepsilon(t) \to \varepsilon^*$, as $t \to \infty$, whatever ε_0. If assumptions (H4) or (H5) were violated, then limit ε^* of $\varepsilon(t)$ as $t \to \infty$ could depend on the initial value at $t = 0$, or $\varepsilon(t)$ could even diverge as $t \to \infty$.

The case where I is given is a special case, but corresponds physically to the case of an optically thin medium submitted to a constant inflow intensity h (see (10)). As opacity $K(\varepsilon)$ is small, the relaxation time of the gas is large compared to the propagation time connected with speed of light c. In other words, we get $I = h$ everywhere before $\varepsilon(t)$ has significantly changed from its starting value ε_0.

We prove the following lemma.

LEMMA 4. *If assumptions* (H4) *and* (H5) *hold, then operator* $\mathcal{B}$ *defined in* (16) *is accretive.*

Proof. We shall first prove that

$$\langle \mathcal{B}u_2 - \mathcal{B}u_1, s_+(u_2 - u_1) \rangle \geqq 0 \tag{38}$$

for all $u_1, u_2 \in Y$ where $s_+ : Y \to Y^*$ is defined by

$$s_+(u) = \{s_+(\varepsilon), s_+(I)\}$$

for all $u = \{\varepsilon, I\} \in Y$ and where s_+ is defined as before.

Let

$$q_\alpha \equiv K(\varepsilon_\alpha)(\varphi(I_\alpha) - B(\varepsilon_\alpha)), \qquad \alpha = 1, 2$$

where $\{\varepsilon_\alpha, I_\alpha\} = u_\alpha$, $\alpha = 1, 2$. We have

$$\mathcal{B}u_2 - \mathcal{B}u_1 = \left\{ \int_0^\infty d\nu \int_{S^2} (q_1 - q_2)\, d\Omega, q_2 - q_1 \right\}.$$

Then

$$\langle \mathcal{B}u_2 - \mathcal{B}u_1, s_+(u_2 - u_1) \rangle = \int_X dx \int_0^\infty d\nu \int_{S^2} (v - w)(q_1 - q_2)\, d\Omega$$

where $v \equiv s_+(\varepsilon_2 - \varepsilon_1)$ and $w \equiv s_+(I_2 - I_1)$. Thus

$$\begin{aligned} &\langle \mathcal{B}u_2 - \mathcal{B}u_1, s_+(u_2 - u_1) \rangle \\ &= \int_X dx \int_0^\infty d\nu \int_{S^2} (v - w)(K(\varepsilon_2)B(\varepsilon_2) - K(\varepsilon_1)B(\varepsilon_1))\, d\Omega \\ &+ \int_X dx \int_0^\infty d\nu \int_{S^2} (w - v)(K(\varepsilon_2)\varphi(I_2) - K(\varepsilon_1)\varphi(I_1))\, d\Omega. \end{aligned} \tag{39}$$

Let

$$A \equiv K(\varepsilon_2)B(\varepsilon_2) - K(\varepsilon_1)B(\varepsilon_1).$$

From (H5) and the definition of v, we get $vA = (A)_+$.

On the other hand, writing $A = A_+ - A_-$ we have $(v - w)A = (1 - w)A_+ + wA_-$. As function $w = s_+(I_2 - I_1)$ satisfies $0 \leqq w \leqq 1$ a.e. for $x \in X$, $\nu > 0$, $\Omega \in S^2$, we obtain $(v - w)A \geqq 0$ which shows that the first term in the right-hand side of (39) is positive. Let us show that the second term is also positive.

Let $y \equiv w - v$ and $z \equiv K(\varepsilon_2)\varphi(I_2) - K(\varepsilon_1)\varphi(I_1)$. We notice that

$$y(x, \Omega, \nu) = \begin{cases} 1 & \text{if } I_2(x, \Omega, \nu) > I_1(x, \Omega, \nu) \text{ and } \varepsilon_2(x) \leqq \varepsilon_1(x), \\ -1 & \text{if } \varepsilon_2(x) > \varepsilon_1(x) \text{ and } I_2(x, \Omega, \nu) \leqq I_1(x, \Omega, \nu), \\ 0 & \text{otherwise.} \end{cases}$$

In case $y=1$, we check that $z \geqq 0$, since, from (H4), $K(\varepsilon_2) \geqq K(\varepsilon_1)$ and $\varphi(I_2) \geqq \varphi(I_1)$. In case $y=-1$, we check that $z \leqq 0$, since (from (H4)) $K(\varepsilon_1) \leqq K(\varepsilon_2)$ and $\varphi(I_2) \leqq \varphi(I_1)$.

Finally, we have proved that $yz \geqq 0$ for all $x \in X$, $\Omega \in S^2$, $\nu > 0$, which suffices to prove that the second term in the right-hand side of (39) is positive, and proves (38).

Permuting indices 1 and 2, we also get

$$\langle \mathscr{B}u_1 - \mathscr{B}u_2, s_+(u_1-u_2)\rangle \geqq 0. \tag{40}$$

As

$$s_0(u_2-u_1) = s_+(u_2-u_1) - s_+(u_1-u_2)$$

we obtain

$$\langle \mathscr{B}u_2 - \mathscr{B}u_1, s_0(u_2-u_1)\rangle \geqq 0$$

by summation of (38) and (40). QED

Applying a result proved at the end of § 1, we obtain the following.

COROLLARY 2. *If assumptions* (H1) *to* (H5) *hold, then operator* $\mathscr{A}+\mathscr{B}$ *is m-accretive.*

As a consequence of Corollary 2, we learn that the distance between $u(t)$ and any equilibrium point u is decreasing in the energy norm.

We refer the reader to [8] and [10] for more useful applications of the accretiveness of $\mathscr{B}$.

7. Operator splitting algorithms for solving the radiative transfer equations. Each time one has to solve a problem of the form

$$\begin{aligned} &\frac{du}{dt} + \mathscr{A}u + \mathscr{B}u = 0, \\ &u(0) = u_0, \end{aligned} \tag{41}$$

one can use operator splitting methods like

$$\begin{aligned} &\text{(i)} \quad \frac{u^{n+1/2} - u^n}{\Delta t} + \mathscr{A}u^{n+1/2} = 0, \\ &\text{(ii)} \quad \frac{u^{n+1} - u^{n+1/2}}{\Delta t} + \mathscr{B}u^{n+1} = 0 \end{aligned} \tag{42}$$

or

$$\begin{aligned} &\text{(i)} \quad u^{n+1/2} = S_{\mathscr{A}}(\Delta t)u^n, \\ &\text{(ii)} \quad u^{n+1} = S_{\mathscr{B}}(\Delta t)u^{n+1/2} \end{aligned} \tag{43}$$

where $S_{\mathscr{A}}(t)$ and $S_{\mathscr{B}}(t)$ denote the semigroups generated by operators $\mathscr{A}$ and $\mathscr{B}$, respectively.

If $\mathscr{A}$ and $\mathscr{B}$ are m-accretive, then each one of these algorithms is unconditionally stable, since, in such a case, the operators

$$(\mathscr{T}+\Delta t\mathscr{A})^{-1}, \quad (\mathscr{T}+\Delta t\mathscr{B})^{-1}, \quad S_{\mathscr{A}}(\Delta t), \quad S_{\mathscr{B}}(\Delta t)$$

are contractions for all $\Delta t > 0$.

We refer the reader to [9] for convergence of such algorithms, as $\Delta t \to 0$, in the case where Y is a Hilbert space.

In the case where $\mathscr{A}$ and $\mathscr{B}$ are defined as in § 2, we see that operator splitting methods introduce a decomposition of time step Δt into two so-called "fractionary steps." The first half-step corresponds to *linear transport* of specific intensity I at speed c, coupled with Thomson diffusion. The second half-step corresponds to local *relaxation of radiation* with material energy. We note that the first half-step is global but linear, whereas the second half-step is nonlinear but local.

Practically, for solving the linear transport part, one has a choice between Monte Carlo methods, or finite element methods. Usually, such methods give a piecewise constant estimate of specific intensity I on some spatial mesh of the domain; material energy ε is chosen piecewise constant on the same mesh.

We note that solving (42)(ii) or (43)(ii) can be performed cell by cell since x is only a parameter for operator $\mathscr{B}$.

Let Q denote some cell of the given mesh: solving (42)(ii) on cell Q amounts to solving

$$(44)\quad \begin{aligned} &\text{(i)} && \varepsilon_Q^{n+1}+\Delta t\int_0^\infty d\nu\int_{S^2} K(\varepsilon_Q^{n+1})(B(\varepsilon_Q^{n+1})-I_Q^{n+1})\,d\Omega=\varepsilon_Q^n,\\ &\text{(ii)} && I_Q^{n+1}+\Delta t K(\varepsilon_Q^{n+1})(I_Q^{n+1}-B(\varepsilon_Q^{n+1}))=I_Q^n. \end{aligned}$$

From (34)(ii) (which is linear w.r.t. I_Q^{n+1}) we get

$$I_Q^{n+1}=\frac{I_Q^n+\Delta t(KB)(\varepsilon_Q^{n+1})}{1+\Delta t K(\varepsilon_Q^{n+1})},$$

which can be substituted in (44)(i). The result is a scalar nonlinear equation of the following type:

$$(45)\qquad f(\varepsilon_Q^{n+1})=\varepsilon_Q^n.$$

It can be seen that, if (H4) and (H5) hold, then f is monotone increasing and (45) has a unique solution; otherwise, for Δt large enough, f may not be monotone, so that (45) may have more than one solution. In such a case, it is not clear which one we should select. Then, from a numerical point of view, assumptions (H4) and (H5), which imply accretiveness, have some useful properties.

In the case where the discrete ordinate method is used to discretize both Ω and ν, (43)(ii) is a system of ordinary differential equations that is more difficult to solve than (42)(ii).

On the other hand, when the Monte Carlo method is used, it is *not* more difficult to solve (43)(i) than (42)(i).

This is why we prefer the following mixed algorithm:

$$(46)\quad \begin{aligned} &\text{(i)} && u^{n+1/2}=S_{\mathscr{A}}(\Delta t)u^n,\\ &\text{(ii)} && u^{n+1}=(\mathscr{T}+\Delta t\mathscr{B})^{-1}u^{n+1/2}. \end{aligned}$$

Note that each one of these algorithms is at most first order accurate with respect to Δt. It is also possible to use second order accurate algorithms (see [9]). However, these operator splitting algorithms, which are very well suited to optically thin media, are fairly inefficient when applied to optically thick media. In fact, in the latter case, radiation and material are strongly coupled since $K(\varepsilon)$ is large, which requires very small time steps Δt.

Acknowledgment. The author would like to thank the referee for many appropriate remarks which have been helpful with respect to mathematical rigor.

REFERENCES

[1] V. BARBU, *Non Linear Semi-groups and Differential Equations in Banach Spaces*, Noordhoff International Publishing, Leyden, The Netherlands, 1976.

[2] C. BARDOS, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre*, Thèse, Paris, 1969.

[3] H. BRÉZIS AND A. PAZY, *Accretive sets and differential equations in Banach spaces*, Israel J. Math., 8 (1970), pp. 367-383.

[4] M. CESSENAT, *Théorèmes de traces L^p pour des espaces de fonctions de la neutronique*, C.R. Acad. Sci. Paris, Sér. I (1984), pp. 831-834..

[5] M. G. CRANDALL, *Non linear semigroups and evolution equations governed by accretive operators*, in Proceedings of the Symposium on Nonlinear Functional Analysis and Applications, F. Browder, ed., Amer. Math. Soc., Berkeley, 1983, to appear in Vol. 45 of the Proceedings of Symposia in Pure Mathematics.

[6] M. G. CRANDALL AND T. LIGGET, *Generation of semigroups of non linear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265-298.

[7] L. C. EVANS, *Applications of non linear semigroup theory to certain partial differential equations*, in Non Linear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 163-188.

[8] F. GOLSE AND B. PERTHAME, *Existence globale d'une solution généralisée pour les équations du transfert radiatif*, C.R. Acad. Sci. Paris, Sér. I (1984), pp. 291-294.

[9] P. L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two maximal monotone operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964-979.

[10] B. MERCIER, *Influence de la troncature de l'opacité au voisinage de $\nu = 0$ en relaxation rayonnement-matière*, Internal report from the Commissariat à l'Energie Atomique (France), Note CEA 2387 (Jan. 1984).

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

010

ESTIMATES FOR CONSERVATION LAWS WITH LITTLE VISCOSITY*

CONSTANTINE M. DAFERMOS†

Abstract. We consider the parabolic system that is generated by adding "artificial viscosity" to the equations of one-dimensional nonlinear elasticity. We construct families of entropies that induce a priori bounds on solutions, independent of the viscosity. The entropies have exponential growth in the case of strain hardening and polynomial growth in the case of strain softening. In particular, we recover the standard theory of invariant regions.

Key words. vanishing viscosity, Riemann invariants, entropy

AMS (MOS) subject classification. 35L65

1. Introduction.

Consider the family of parabolic equations

$$\begin{aligned} &\partial_t u - \partial_x v + f(u, v) = \nu \partial_x^2 u, \\ &\partial_t v - \partial_x \sigma(u) + g(u, v) = \nu \partial_x^2 v, \end{aligned} \tag{1.1}$$

where $\sigma(u)$ is a given smooth, strongly monotone function on $(-\infty, \infty)$,

$$\sigma'(u) = a^2(u), \quad a(u) > 0, \quad -\infty < u < \infty. \tag{1.2}$$

$f(u, v)$ and $g(u, v)$ are given smooth functions on $(-\infty, \infty) \times (-\infty, \infty)$, and ν is a positive parameter that measures the "viscosity."

We view (1.1) as a singular perturbation of the strictly hyperbolic system

$$\begin{aligned} &\partial_t u - \partial_x v + f(u, v) = 0, \\ &\partial_t v - \partial_x \sigma(u) + g(u, v) = 0. \end{aligned} \tag{1.3}$$

The method of *vanishing viscosity* seeks to identify and construct admissible discontinuous solutions of the Cauchy problem for (1.3) as limits of solutions $\{u(x, t), v(x, t)\}$ of the Cauchy problem for (1.1), with $\nu \downarrow 0$. To carry this program out, one needs a priori bounds on $\{u(x, t), v(x, t)\}$, independent of ν, sufficiently strong to induce sequences that are convergent almost everywhere. A bound on the total variation of $\{u(x, t), v(x, t)\}$ would be ideal for that purpose, but no estimates of this type are presently available. A weaker, L^∞, estimate for $\{u(x, t), v(x, t)\}$ would induce sequences convergent in L^∞ weak $*$ but this would not suffice, in itself, to guarantee that the limit is a solution of (1.3). Nevertheless, the theory of compensated compactness yields (DiPerna [2], Rascle [6]) that when $\sigma(u)$ has isolated inflection points, any sequence of solutions of (1.1) converging in L^∞ weak $*$ converges necessarily almost everywhere.

The standard vehicle for establishing uniform L^∞ bounds for solutions of (1.1) is the method of Chueh, Conley and Smoller [1], which applies if and only if $\sigma(u)$ has a single inflection point, say at u_0, being convex on (u_0, ∞) and concave on $(-\infty, u_0)$,

* Received by the editors October 28, 1985; accepted for publication April 1, 1986. This research was supported by the National Science Foundation under grant DMS-8025355, by the U.S. Army under contract DAAG-29-83-K-0029 and the Office of Naval Research under contract N00014-83-K-0542.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

and

$$[\operatorname{sgn}(u-u_0)]a(u)f(u,v)+[\operatorname{sgn} v]g(u,v)\geqq 0 \tag{1.4}$$

holds for $|u|$, $|v|$ sufficiently large. Under these conditions, the level curves of the *Riemann invariants*

$$r(u,v):=v+\int_{u_0}^{u} a(w)\,dw, \qquad s(u,v):=v-\int_{u_0}^{u} a(w)\,dw \tag{1.5}$$

confine a nested family of open, bounded, positively invariant regions for (1.1) whose union covers the entire u-v plane. (In fact, (1.4) simply states that the vector field $\{f(u,v), g(u,v)\}$ points towards the exterior of these regions.) Then the range of the solution of the Cauchy problem for (1.1) is trapped inside one of these invariant regions, determined solely by the initial data.

It is not to be expected that solutions $\{u(x,t), v(x,t)\}$ of the Cauchy problem for (1.1) will be bounded in L^∞, uniformly in ν, for arbitrary $\sigma(u)$. For instance, considering that, in the isentropic flow of an ideal gas (the flow is described by equations of the form (1.3)), vacuum may develop over regions of space-time, one should anticipate that $u(x,t)$ will not stay bounded from above, uniformly in ν, when $\sigma(u)$ is concave for u large. Still one hopes that (perhaps weaker than L^∞) bounds for $\{u(x,t), v(x,t)\}$ hold under assumptions less stringent than those required for the existence of invariant regions. For example, in the homogeneous case $f(u,v)\equiv g(u,v)\equiv 0$, it is reasonable to expect strong bounds when $\sigma(u)$ is convex for u large and concave for u small, irrespectively of the number of inflection points in between. In fact it is plausible that estimates of a certain type apply, even when $\sigma(u)$ is concave for u large and/or convex for u small, provided $\sigma''(u)$ decays to zero, sufficiently fast, as $u\uparrow\infty$ and/or $u\downarrow-\infty$. These estimates should extend to the nonhomogeneous case, so long as the growth of $f(u,v)$ and $g(u,v)$, as $|u|$ and $|v|$ tend to infinity, is properly restricted.

The aim of this paper is to establish estimates on solutions of the Cauchy problem for (1.1), under various assumptions on $\sigma(u)$. Our approach rests on the construction of appropriate entropies for (1.3) (cf. Lax [4]). For interesting, recent, related results see Serre [7] and Venttsel' [8].

A smooth, convex function $\eta(u,v)$ on $(-\infty,\infty)\times(-\infty,\infty)$ is an *entropy* for (1.3), with *entropy flux* $q(u,v)$, if

$$\begin{aligned} q_u(u,v)&=-\sigma'(u)\eta_v(u,v),\\ q_v(u,v)&=-\eta_u(u,v) \end{aligned} \tag{1.6}$$

hold on $(-\infty,\infty)\times(-\infty,\infty)$. By eliminating $q(u,v)$ between the two equations in (1.6) and using (1.2), we deduce that $\eta(u,v)$ is an entropy for (1.3) if and only if it is a convex solution of the linear wave equation

$$\eta_{uu}(u,v)=a^2(u)\eta_{vv}(u,v), \tag{1.7}$$

on $(-\infty,\infty)\times(-\infty,\infty)$.

Assume $\eta(u,v)$ is an entropy for (1.3), with entropy flux $q(u,v)$, and let $\{u(x,t), v(x,t)\}$ be a solution of the Cauchy problem for (1.1), which tends, as $|x|\to\infty$, to a constant state $\{\bar u,\bar v\}$, for any $t\in[0,\infty)$. Upon setting

$$\begin{aligned} \bar\eta(u,v)&:=\eta(u,v)-\eta(\bar u,\bar v)-\eta_u(\bar u,\bar v)(u-\bar u)-\eta_v(\bar u,\bar v)(v-\bar v),\\ \bar q(u,v)&:=q(u,v)-q(\bar u,\bar v)+\eta_u(\bar u,\bar v)(u-\bar u)+\eta_v(\bar u,\bar v)[\sigma(u)-\sigma(\bar u)], \end{aligned} \tag{1.8}$$

we verify easily the familiar identity

$$\begin{aligned}(1.9)\qquad &\partial_t\bar{\eta}(u,v)+\partial_x\bar{q}(u,v)+\bar{\eta}_u(u,v)f(u,v)+\bar{\eta}_v(u,v)g(u,v)\\ &\quad=\nu\partial_x^2\bar{\eta}(u,v)-\nu\{\eta_{uu}(u,v)(\partial_x u)^2+2\eta_{uv}(u,v)(\partial_x u)(\partial_x v)+\eta_{vv}(u,v)(\partial_x v)^2\},\end{aligned}$$

which induces, by virtue of the convexity of $\eta(u, v)$, the inequality

$$\begin{aligned}(1.10)\qquad &\int_{-\infty}^{\infty}\bar{\eta}(u(x,t),v(x,t))\,dx+\int_0^t\int_{-\infty}^{\infty}\{\bar{\eta}_u(u,v)f(u,v)+\bar{\eta}_v(u,v)g(u,v)\}\,dx\,d\tau\\ &\quad\leqq\int_{-\infty}^{\infty}\bar{\eta}(u(x,0),v(x,0))\,dx.\end{aligned}$$

We observe that, since $\eta(u, v)$ is convex, $\bar{\eta}(u, v)$, as defined by $(1.8)_1$, is nonnegative on $(-\infty,\infty)\times(-\infty,\infty)$ and vanishes at $(\bar{u},\bar{v})$, to quadratic order.

The simplest choice of an entropy-entropy flux pair is

$$\begin{aligned}(1.11)\qquad &\eta(u,v)=\frac{1}{2}v^2+\int_0^u\sigma(w)\,dw,\\ &q(u,v)=-v\sigma(u).\end{aligned}$$

This is valid for arbitrary nondecreasing $\sigma(u)$ and yields the standard energy estimate. Our aim, however, is to derive sharper bounds and for that purpose we have to impose restrictions on $\sigma(u)$. In the homogeneous case $f(u,v)\equiv g(u,v)\equiv 0$, equations (1.3) govern, in Lagrangian coordinates, the motion of one-dimensional, nonlinear, elastic media. We need estimates that cover both the case of strain hardening and the case of strain softening.

In § 2 we assume $\sigma(u)$ is convex for u large and concave for u small (strain hardening) and construct entropies for (1.3) which grow exponentially in u and v. They induce a priori estimates on solutions of the Cauchy problem for (1.1). As a byproduct of our analysis, we get an alternative derivation of invariant regions for (1.1) and a new proof of a recent result of Roytburd and Slemrod [5]. Special entropies with exponential growth were constructed by Lax [4] and by DiPerna [3].

In § 3 we consider $\sigma(u)$ which are concave when u is large and/or convex when u is small (strain softening) and construct entropies that grow like powers of u and v. The maximal power depends on how fast $\sigma''(u)$ decays to zero, as $u\uparrow\infty$ and/or $u\downarrow-\infty$. These entropies induce L^p estimates, independent of ν, on solutions of the Cauchy problem for (1.1) with initial data in $L^p(-\infty,\infty)\cap L^2(-\infty,\infty)$.

2. Entropies with exponential growth.

In this section we assume $\sigma(u)$ is convex for u large and concave for u small, i.e., there are numbers $-\infty<u_-\leqq u_+<\infty$ such that

$$\begin{aligned}(2.1)\qquad &a'(u)\leqq 0,\qquad -\infty<u<u_-,\\ &a'(u)\geqq 0,\qquad u_+<u<\infty,\end{aligned}$$

and construct entropies for (1.3) of the form

$$(2.2)\qquad \eta(u,v)=Y(u)\cosh(kv),$$

where k is a positive constant. By virtue of (1.7), $Y(u)$ must satisfy the linear differential equation

$$(2.3)\qquad Y''(u)=k^2a^2(u)Y(u),\qquad -\infty<u<\infty.$$

For reasons to become apparent shortly, we determine $Y(u)$ as the solution of (2.3) with initial conditions

$$(2.4) \qquad Y(u_0)=1, \qquad Y'(u_0)=0,$$

where u_0 is a point at which $a(u)$ attains its maximum in the interval $[u_-, u_+]$. In particular, $Y'(u)\leqq 0$ on $(-\infty, u_0)$, $Y'(u)\geqq 0$ on (u_0, ∞) and

$$(2.5) \qquad Y(u)\geqq \cosh[k\bar{a}(u-u_0)], \qquad -\infty<u<\infty,$$

where $\bar{a}$ is the minimum of $a(u)$ over $(-\infty, \infty)$.

A simple calculation shows that $\eta(u, v)$, as defined by (2.2) with $Y(u)$ satisfying (2.3), (2.4), will be strictly convex on $(-\infty, \infty)\times(-\infty, \infty)$ if and only if

$$(2.6) \qquad |Y'(u)|\leqq ka(u)Y(u), \qquad -\infty<u<\infty.$$

As long as (2.6) holds,

$$(2.7) \qquad Y(u)\leqq \exp\left\{k\left|\int_{u_0}^{u} a(w)\,dw\right|\right\}, \qquad -\infty<u<\infty.$$

We shall test (2.6) only for $u\geqq u_0$, because the discussion for the case $u\leqq u_0$ would be completely symmetrical. We define

$$(2.8) \qquad \chi(u):=ka(u)Y(u)-Y'(u), \qquad u_0\leqq u<\infty,$$

and note that, by account of (2.3), (2.4), $\chi(u)$ is the solution of the initial value problem

$$(2.9) \qquad \chi'(u)+ka(u)\chi(u)=ka'(u)Y(u), \qquad u_0\leqq u<\infty,$$

$$(2.10) \qquad \chi(u_0)=ka(u_0).$$

It is clear that when $a'(u)\geqq 0$ on $[u_0, \infty)$, then $\chi(u)>0$ for all u in $[u_0, \infty)$. We have thus shown the following.

PROPOSITION 2.1. *If for some u_0 in $(-\infty, \infty)$*

$$(2.11) \qquad (u-u_0)a'(u)\geqq 0, \qquad -\infty<u<\infty,$$

then the function $\eta(u, v)$, defined through (2.2), (2.3), (2.4), *for any $k>0$, is a strictly convex entropy of* (1.3).

By contrast, when $a'(u)<0$ on some interval (u_0, u_1), a crude estimation using (2.9), (2.10), (2.5) shows that, for k very large, $\chi(u)<0$ on an interval $u_0<\hat{u}<u<u_1$. However, we have the following.

PROPOSITION 2.2. *Assume* (2.1) *holds. Then there is $k_0>0$ such that the function $\eta(u, v)$, defined through* (2.2), (2.3), (2.4), *for any $0<k\leqq k_0$, is a strictly convex entropy of* (1.3).

Proof. We integrate (2.9), (2.10) to get

$$(2.12) \qquad \exp\left\{k\int_{u_0}^{u} a(w)\,dw\right\}\chi(u)=ka(u_0)+k\int_{u_0}^{u} a'(w)Y(w)\exp\left\{k\int_{u_0}^{w} a(s)\,ds\right\}dw.$$

The right-hand side of (2.12) attains its minimum over $[u_0, \infty)$ at a point $u_1\in[u_0, u_+]$. We recall that u_0 is a point at which $a(u)$ attains its maximum on the interval $[u_-, u_+]$. Combining this with the observation that the function

$$Y(w)\exp\left\{k\int_{u_0}^{w} a(s)\,ds\right\}$$

is increasing on $[u_0, \infty)$, we infer that

$$\begin{aligned}\int_{u_0}^{u} a'(w)Y(w)\exp\left\{k\int_{u_0}^{w} a(s)\,ds\right\}dw \\ \geqq -\{a(u_0)-a(u_1)\}Y(u_1)\exp\left\{k\int_{u_0}^{u_1} a(s)\,ds\right\},\end{aligned} \tag{2.13}$$

for any u in $[u_0, \infty)$. From (2.12), (2.13) and (2.7) it follows that so long as χ is nonnegative on the interval (u_0, u),

$$\frac{1}{k}\exp\left\{k\int_{u_0}^{u} a(w)\,dw\right\}\chi(u) \geqq a(u_0)-\{a(u_0)-a(u_1)\}\exp\left\{2k\int_{u_0}^{u_1} a(s)\,ds\right\}. \tag{2.14}$$

Therefore, if

$$k \leqq \frac{1}{2}\left\{\int_{u_0}^{u_1} a(s)\,ds\right\}^{-1}\log\frac{a(u_0)}{a(u_0)-a(u_1)}, \tag{2.15}$$

then $\chi(u) > 0$ for all u in $[u_0, \infty)$. This completes the proof. □

We proceed to demonstrate how the entropies constructed above may induce a priori estimates on solutions $\{u(x, t), v(x, t)\}$ of the Cauchy problem for (1.1). Let us assume that, for $t \in [0, \infty)$, $\{u(x, t), v(x, t)\}$ decays, as $|x| \to \infty$, to a constant state $(\bar{u}, \bar{v})$ which is a critical point of the vector field $\{f(u, v), g(u, v)\}$, i.e.,

$$f(\bar{u}, \bar{v}) = g(\bar{u}, \bar{v}) = 0. \tag{2.16}$$

The entropy (2.2) generates, through $(1.8)_1$, a new entropy $\bar{\eta}(u, v)$, which is nonnegative on $(-\infty, \infty) \times (-\infty, \infty)$ and vanishes at $(\bar{u}, \bar{v})$ to quadratic order. We shall estimate the solution by monitoring, with the help of (1.10), the evolution of

$$\int_{-\infty}^{\infty} \bar{\eta}(u(x, t), v(x, t))\,dx.$$

As a consequence of $(1.8)_1$ and (2.16),

$$\bar{\eta}_u(u, v)f(u, v) + \bar{\eta}_v(u, v)g(u, v)$$

vanishes to quadratic order at $(\bar{u}, \bar{v})$. Furthermore, (2.2) and (2.6) yield

$$\begin{aligned}|\eta_u(u, v)| &\leqq ka(u)\eta(u, v), \\ |\eta_v(u, v)| &\leqq k\eta(u, v).\end{aligned} \tag{2.17}$$

First we consider any $a(u)$ that satisfies (2.1) and we assume

$$a(u)|f(u, v)| + |g(u, v)| \leqq L, \quad -\infty < u < \infty, \quad -\infty < v < \infty. \tag{2.18}$$

Then there is a positive constant C such that

$$|\bar{\eta}_u(u, v)f(u, v) + \bar{\eta}_v(u, v)g(u, v)| \leqq CkL\bar{\eta}(u, v). \tag{2.19}$$

From (1.10), (2.19) and Gronwall's inequality we derive the estimate

$$\int_{-\infty}^{\infty} \bar{\eta}(u(x, t), v(x, t))\,dx \leqq e^{CkLt}\int_{-\infty}^{\infty} \bar{\eta}(u(x, 0), v(x, 0))\,dx. \tag{2.20}$$

Next we turn to the special case of $a(u)$ that satisfies (2.11) and establish L^∞ estimates on $\{u(x, t), v(x, t)\}$ assuming that $\{u(x, 0), v(x, 0)\} \in L^\infty(-\infty, \infty)$ and

$$a(u)|f(u, v)| + |g(u, v)| \leqq L(|u| + |v| + 1), \quad -\infty < u < \infty, \quad -\infty < v < \infty. \tag{2.21}$$

Under this hypothesis and so long as $k \geqq \delta > 0$, we have

$$(2.22)\qquad |\bar\eta_u(u,v)f(u,v)+\bar\eta_v(u,v)g(u,v)| \leqq CkL(|u|+|v|+1)\bar\eta(u,v),$$

on $(-\infty,\infty)\times(-\infty,\infty)$, where C is a positive constant that may depend on δ but is otherwise independent of k. Upon setting

$$(2.23)\qquad W(t) := \max_x \,(|u(x,t)|+|v(x,t)|+1),$$

(1.10), (2.22) and Gronwall's inequality yield

$$(2.24)\qquad \int_{-\infty}^{\infty} \bar\eta(u(x,t),v(x,t))\,dx \leqq \exp\left\{CkL\int_0^t W(\tau)\,d\tau\right\}\int_{-\infty}^{\infty}\bar\eta(u(x,0),v(x,0))\,dx.$$

We raise both sides of (2.24) to the power $1/k$ and we pass to the limit, as $k\uparrow\infty$, taking account of (2.2), (2.5) and (2.7). In the resulting inequality we take the logarithm of both sides, thus obtaining

$$(2.25)\qquad W(t) \leqq A + CL\int_0^t W(\tau)\,d\tau, \qquad 0\leqq t<\infty,$$

where A depends solely upon the L^∞ norm of $\{u(x,0), v(x,0)\}$. Therefore, Gronwall's inequality implies that $W(t)$, and thereby also $\{u(x,t), v(x,t)\}$, are bounded, on compact time intervals, uniformly in $\nu>0$. In particular, in the homogeneous case $f(u,v)\equiv g(u,v)\equiv 0$, $\{u(x,t), v(x,t)\}$ is bounded on $(-\infty,\infty)\times[0,\infty)$, uniformly in ν.

Staying with the case of $a(u)$ that satisfies (2.11), we indicate briefly how the standard invariant regions [1] may be identified by studying the asymptotics of solutions of (2.3), as $k\uparrow\infty$. To this end, we introduce new variables:

$$(2.26)\qquad \xi := \int_{u_0}^{u} a(w)\,dw,$$

$$(2.27)\qquad Z := a(u)^{1/2}Y.$$

A straightforward calculation, using (2.3), yields

$$(2.28)\qquad \frac{d^2Z}{d\xi^2} = k^2Z - a^{-3/2}(a^{-1/2})''Z.$$

For u confined in bounded intervals, the variation of constants formula applied on (2.28) yields that, as $k\uparrow\infty$,

$$(2.29)\qquad Z(\xi) = \left[A + O\left(\frac{1}{k}\right)\right]e^{k|\xi|},$$

$$(2.30)\qquad \begin{aligned} &\frac{dZ}{d\xi}+kZ = O(1)\,e^{-k\xi}, &&\xi<0,\\ &\frac{dZ}{d\xi}-kZ = O(1)\,e^{k\xi}, &&\xi>0.\end{aligned}$$

Therefore, using (2.2), (2.27), (2.29), (2.26), (1.5) and (2.30), we obtain

$$(2.31)\qquad \lim_{k\to\infty}\eta(u,v)^{1/k} = \begin{cases} \exp[r(u,v)] & \text{if } u>u_0,\quad v>0,\\ \exp[s(u,v)] & \text{if } u<u_0,\quad v>0,\\ \exp[-s(u,v)] & \text{if } u>u_0,\quad v<0,\\ \exp[-r(u,v)] & \text{if } u<u_0,\quad v<0,\end{cases}$$

$$\frac{\eta_u(u,v)}{\eta(u,v)}=k[\operatorname{sgn}(u-u_0)]a(u)+O(1), \tag{2.32}$$
$$\frac{\eta_v(u,v)}{\eta(u,v)}=k\operatorname{sgn} v+O(1).$$

It follows from (2.32) that when k is large and (1.4) holds as a strict inequality, then the second term on the left-hand side of (1.10) is nonnegative. Hence, raising both sides of (1.10) to the power $1/k$, letting $k\uparrow\infty$, and using (2.31) we conclude that the level curves of the Riemann invariants (1.5) confine positively invariant regions for solutions of (1.1). A derivation of invariant regions that is similar in spirit to the above is given in [4].

The approach to invariant regions presented here may have some advantage over the traditional one [1] when dealing with solutions of (1.1) that are not C^2 smooth. A relevant example, arising in the theory of phase transitions, was discussed recently by Roytburd and Slemrod [5]: Assume $a(u)$ is smooth and strictly decreasing on $(-\infty,\alpha)$, it vanishes identically on the interval (α,β), and it is smooth and strictly increasing on (β,∞). Let $a(u)$ jump from a negative value to zero at $u=\alpha$ and from zero to a positive value at $u=\beta$. Thus $\sigma(u)$ is merely Lipschitz continuous on $(-\infty,\infty)$ and (1.1) does not generally have classical solutions. It is shown in [5] that, under the above hypotheses, the Cauchy problem for (1.1) has a mild solution $\{u(x,t),v(x,t)\}$ in the class of continuous functions. The standard theory of the (linear) equation of heat conduction and a straightforward "bootstrapping" argument yield that $\partial_x u$, $\partial_x v$, $\partial_t u$, $\partial_t v$, $\partial_x^2 u$, $\partial_x^2 v$ are all in $L^2((-\infty,\infty)\times(0,T))$, for any $T>0$. In particular, (1.10) is still valid under the current conditions. We may thus apply our argument, using entropies (2.2) with $k\uparrow\infty$, to infer that the level curves of the Riemann invariants (1.5) still confine positively invariant regions for solutions of (1.1). The original derivation of this result in [5] employs a mollification of $\sigma(u)$ and requires a rather lengthy argument in order to pass to the limit.

3. Entropies with polynomial growth. Here we construct entropies $\eta(u,v)$ on $(-\infty,\infty)\times(-\infty,\infty)$, which grow at infinity like the pth power of u and v and thus induce L^p estimates on solutions of the Cauchy problem for (1.1). Throughout this section we will be assuming

$$a(u)\geqq\bar{a}>0,\qquad -\infty<u<\infty, \tag{3.1}$$

but we do not impose, as yet, any conditions of convexity on $\sigma(u)$.

We shall seek convex solutions of (1.7) on $(-\infty,\infty)\times(-\infty,\infty)$ with initial conditions

$$\eta(0,v)=H(v),\quad \eta_u(0,v)=0,\quad -\infty<v<\infty, \tag{3.2}$$

where $H(v)$ is a positive, even, convex function on $(-\infty,\infty)$.

If $\eta(u,v)$ is the solution of (1.7), (3.2), we set

$$\Phi(u,v):=a(u)\eta_{vv}(u,v)-\eta_{uv}(u,v), \tag{3.3}$$
$$\Psi(u,v):=a(u)\eta_{vv}(u,v)+\eta_{uv}(u,v).$$

It follows from (1.7), (3.3) that

$$\eta_{vv}(u,v)=\frac{1}{2a(u)}\{\Phi(u,v)+\Psi(u,v)\}, \tag{3.4}$$

$$\eta_{uu}(u,v)\eta_{vv}(u,v)-\eta_{uv}^2(u,v)=\Phi(u,v)\Psi(u,v), \tag{3.5}$$

and so $\eta(u, v)$ is strictly convex on $(-\infty, \infty)\times(-\infty, \infty)$ if and only if

$$\Phi(u, v)>0, \quad \Psi(u, v)>0, \quad -\infty<u<\infty, \quad -\infty<v<\infty. \tag{3.6}$$

Combining (3.3), (3.4) and (1.7) we get

$$\begin{aligned} \Phi_u(u, v)+a(u)\Phi_v(u, v)&=\frac{a'(u)}{2a(u)}\{\Phi(u, v)+\Psi(u, v)\}, \\ \Psi_u(u, v)-a(u)\Psi_v(u, v)&=\frac{a'(u)}{2a(u)}\{\Phi(u, v)+\Psi(u, v)\}. \end{aligned} \tag{3.7}$$

$$\Phi(0, v)=\Psi(0, v)=R(v), \qquad -\infty<v<\infty, \tag{3.8}$$

where $R(v)=H_{vv}(v)$. In particular, $R(v)$ is a nonnegative even function,

$$R(-v)=R(v), \qquad -\infty<v<\infty. \tag{3.9}$$

From (3.7), (3.8) and (3.9) it follows that

$$\Psi(u, v)=\Phi(u, -v), \quad -\infty<u<\infty, \quad -\infty<v<\infty. \tag{3.10}$$

Every positive solution of (3.7) on $(-\infty, \infty)\times(-\infty, \infty)$ induces, via (3.3) and (1.7), a strictly convex entropy for (1.3). If $ua'(u)\geqq 0$ on $(-\infty, \infty)$, then the solution of (3.7), (3.8) with any positive $R(v)$ is automatically positive on $(-\infty, \infty)\times(-\infty, \infty)$. Thus, when $\sigma(u)$ is concave on $(-\infty, 0]$ and convex on $[0, \infty)$ it is easy to construct strictly convex entropies with arbitrarily prescribed growth. Our current objective, however, is to determine entropies for the case where $\sigma(u)$ may be concave for u large and/or convex for u small.

We consider solutions of (3.7), (3.8) with

$$R(v)=a(0)|v|^{\gamma}, \qquad -\infty<v<\infty, \tag{3.11}$$

where $\gamma\geqq 0$. When γ is an even integer, say $\gamma=2m$, the solution of (3.7), (3.8), (3.11) is of the form

$$\begin{aligned} \Phi(u, v)&=\sum_{k=0}^{m} A_k(u)a(u)v^{2k}+\sum_{k=0}^{m-1} B_k(u)v^{2k+1}, \\ \Psi(u, v)&=\sum_{k=0}^{m} A_k(u)a(u)v^{2k}-\sum_{k=0}^{m-1} B_k(u)v^{2k+1} \end{aligned} \tag{3.12}$$

where

$$A_m(u)=1, \qquad -\infty<u<\infty, \tag{3.13}$$

and $B_k(u)$, $A_k(u)$, $k=0, \cdots, m-1$, are determined by the recursion relations

$$\begin{aligned} B_k'(u)&=-(2k+2)a^2(u)A_{k+1}(u), \\ A_k'(u)&=-(2k+1)B_k(u), \end{aligned} \qquad k=0, \cdots, m-1, \tag{3.14}$$

$$B_k(0)=A_k(0)=0, \qquad k=0, \cdots, m-1. \tag{3.15}$$

When $m=0$, (3.12) reduces to

$$\Phi(u, v)=\Psi(u, v)=a(u), \quad -\infty<u<\infty, \quad -\infty<v<\infty. \tag{3.16}$$

This solution is uniformly positive on $(-\infty, \infty)\times(-\infty, \infty)$, by virtue of (3.1), and induces the entropy $(1.11)_1$.

When $m=1$, (3.14), (3.15), (3.13) and (1.2) yield

$$B_0(u)=-2[\sigma(u)-\sigma(0)],\qquad A_0(u)=2\int_0^u[\sigma(w)-\sigma(0)]\,dw. \tag{3.17}$$

Inserting these values into (3.12) and after an integration by parts we obtain

$$\begin{aligned}\Phi(u,v)&=a(u)\left\{v-\frac{\sigma(u)-\sigma(0)}{a(u)}\right\}^2-a(u)\int_0^u[\sigma(w)-\sigma(0)]^2[a^{-2}(w)]'\,dw,\\ \Psi(u,v)&=a(u)\left\{v+\frac{\sigma(u)-\sigma(0)}{a(u)}\right\}^2-a(u)\int_0^u[\sigma(w)-\sigma(0)]^2[a^{-2}(w)]'\,dw.\end{aligned} \tag{3.18}$$

If

$$\int_0^u[\sigma(w)-\sigma(0)]^2[a^{-2}(w)]'\,dw<K<\infty,\qquad -\infty<u<\infty, \tag{3.19}$$

then we can generate positive solutions of (3.7) with quadratic growth on $(-\infty,\infty)\times(-\infty,\infty)$ by adding to solution (3.18) K times the solution (3.16). For the case of interest, namely $ua'(u)\leqq 0$ on $(-\infty,\infty)$, and by account of (3.1), (3.19) is equivalent to

$$\int_{-\infty}^{\infty}w^2|a'(w)|\,dw<\infty. \tag{3.20}$$

Though in principle one may proceed with the study of explicit solutions (3.12) for $m=2,3,\cdots$, the calculations get so complicated that a more qualitative approach becomes necessary. We discretize the problem by the following procedure: We fix $\tau>0$, set $a_n:=a(n\tau)$, $n=0,\pm1,\pm2,\cdots$, and define the step function $\hat{a}(u)$ on $(-\infty,\infty)$ by

$$\hat{a}(u)=\begin{cases}a_n, & n\tau\leqq u<(n+1)\tau,\quad n=0,1,2,\cdots,\\ a_n, & (n-1)\tau<u\leqq n\tau,\quad n=0,-1,-2,\cdots.\end{cases} \tag{3.21}$$

We will determine the solution of (1.7), (3.2) on $(-\infty,\infty)\times(-\infty,\infty)$ by first solving

$$\hat{\eta}_{uu}(u,v)=\hat{a}^2(u)\hat{\eta}_{vv}(u,v), \tag{3.22}$$

with the same initial data, and then passing to the limit $\tau\downarrow 0$.

The solution of (3.22), (3.2) is a C^1 function with continuous second derivatives $\hat{\eta}_{uv}(u,v)$ and $\hat{\eta}_{vv}(u,v)$. By contrast, $\hat{\eta}_{uu}(u,v)$ experiences jump discontinuities across the lines $u=\pm\tau,\pm2\tau,\cdots$. The functions

$$\begin{aligned}\hat{\Phi}(u,v)&:=\hat{a}(u)\hat{\eta}_{vv}(u,v)-\hat{\eta}_{uv}(u,v),\\ \hat{\Psi}(u,v)&:=\hat{a}(u)\hat{\eta}_{vv}(u,v)+\hat{\eta}_{uv}(u,v),\end{aligned} \tag{3.23}$$

also experience jump discontinuities across the lines $u=\pm\tau,\pm2\tau,\cdots$ but are continuous, with respect to u, from the right on $[0,\infty)$ and from the left on $(-\infty,0]$.

In the sequel, we restrict attention to the upper half-plane, $0\leqq u<\infty$, $-\infty<v<\infty$, because the discussion for the lower half-plane would be entirely symmetrical. Since $a(u)$ is constant on the interval $[n\tau,(n+1)\tau)$, it follows easily from (3.23), (3.22) that

$$\begin{aligned}\hat{\Phi}_u(u,v)+a_n\hat{\Phi}_v(u,v)&=0,\\ \hat{\Psi}_u(u,v)-a_n\hat{\Psi}_v(u,v)&=0,\end{aligned}\qquad n\tau<u<(n+1)\tau. \tag{3.24}$$

Combining (3.23), (3.21) and (3.24) we deduce

$$
\tag{3.25}
\begin{aligned}
\hat{\Phi}((n+1)\tau, v) &= \frac{a_{n+1}+a_n}{2a_n}\hat{\Phi}(n\tau, v-a_n\tau)+\frac{a_{n+1}-a_n}{2a_n}\hat{\Psi}(n\tau, v+a_n\tau),\\
\hat{\Psi}((n+1)\tau, v) &= \frac{a_{n+1}-a_n}{2a_n}\hat{\Phi}(n\tau, v-a_n\tau)+\frac{a_{n+1}+a_n}{2a_n}\hat{\Psi}(n\tau, v+a_n\tau),
\end{aligned}
$$

which are the discrete analogues of (3.7).

For simple initial data it is possible to determine explicitly $\hat{\Phi}(u, v)$ and $\hat{\Psi}(u, v)$ by solving (3.25) and then using (3.24). For instance, the solution of (3.25) under initial data (3.8), (3.11), with $\gamma=0$, is

$$
\tag{3.26}
\hat{\Phi}(n\tau, v)=\hat{\Psi}(n\tau, v)=a_n, \qquad n=0,1,2,\cdots.
$$

Similarly, it is straightforward to verify by induction that the solution of (3.25) under initial data (3.8), (3.11), with $\gamma=2$, is

$$
\tag{3.27}
\begin{aligned}
\frac{1}{a_n}\hat{\Phi}(n\tau, v) &= \left\{v-\frac{a_0^2+\cdots+a_{n-1}^2}{a_n}\tau\right\}^2+\sum_{k=0}^{n-1}\{a_0^2+\cdots+a_k^2\}\{a_k^{-2}-a_{k+1}^{-2}\},\\
\frac{1}{a_n}\hat{\Psi}(n\tau, v) &= \left\{v+\frac{a_0^2+\cdots+a_{n-1}^2}{a_n}\tau\right\}^2+\sum_{k=0}^{n-1}\{a_0^2+\cdots+a_k^2\}\{a_k^{-2}-a_{k+1}^{-2}\}.
\end{aligned}
$$

As was to be expected, the above solution converges to (3.18), as $\tau \downarrow 0$.

Turning now to general initial data (3.8), (3.9), we apply (3.25) recursively, thus arriving at equations of the form

$$
\tag{3.28}
\begin{aligned}
\hat{\Phi}(n\tau, v) &= \sum_{(\varepsilon_0,\cdots,\varepsilon_{n-1})}\beta_n(\varepsilon_0,\cdots,\varepsilon_{n-1})R\left(v+\sum_{i=0}^{n-1}\varepsilon_i a_i\tau\right),\\
\hat{\Psi}(n\tau, v) &= \sum_{(\varepsilon_0,\cdots,\varepsilon_{n-1})}\beta_n(\varepsilon_0,\cdots,\varepsilon_{n-1})R\left(v-\sum_{i=0}^{n-1}\varepsilon_i a_i\tau\right).
\end{aligned}
$$

In (3.28) the ε_i take values ± 1 so the summation contains 2^n terms. The coefficients $\beta_n(\varepsilon_0,\cdots,\varepsilon_{n-1})$ are determined through the recurrence relations

$$
\tag{3.29}
\beta_0=1,
$$

$$
\tag{3.30}
\begin{aligned}
\beta_{n+1}(\varepsilon_0,\cdots,\varepsilon_{n-1},-1) &= \frac{a_{n+1}+a_n}{2a_n}\beta_n(\varepsilon_0,\cdots,\varepsilon_{n-1}),\\
\beta_{n+1}(\varepsilon_0,\cdots,\varepsilon_{n-1},1) &= \frac{a_{n+1}-a_n}{2a_n}\beta_n(\varepsilon_0,\cdots,\varepsilon_{n-1})
\end{aligned}
$$

for $n=0,1,2,\cdots$.

When $a'(u)\geqq 0$ on $[0,\infty)$, in which case the sequence $\{a_n\}$ is nondecreasing, all β_n are positive and so any positive $R(v)$ yields positive $\hat{\Phi}(u, v)$ and $\hat{\Psi}(u, v)$. Here, however, we are interested in $\sigma(u)$ that are concave on $[0,\infty)$ and convex on $(-\infty, 0]$. The main result of this section is the following proposition.

PROPOSITION 3.1. *Assume* (3.1) *holds and*

$$
\tag{3.31}
ua'(u)\leqq 0, \qquad -\infty<u<\infty.
$$

Furthermore, let

$$
\tag{3.32}
\int_{-\infty}^{\infty}|u|^{p-2}|a'(u)|\,du<\infty
$$

for some $p \geqq 2$. *Then there exists a strictly convex entropy* $\eta(u, v)$ *for* (1.3) *which satisfies the following growth conditions on* $(-\infty, \infty) \times (-\infty, \infty)$:

$$\mu(|u|^p + |v|^p) \leqq \eta(u, v) \leqq M(|u|^p + |v|^p + 1), \tag{3.33}$$

$$|\eta_u(u, v)| + |\eta_v(u, v)| \leqq L(|u|^{p-1} + |v|^{p-1} + 1), \tag{3.34}$$

with μ, M *and* L *positive constants.*

Proof. Let us estimate $\hat{\Phi}(n\tau, v)$, as computed via $(3.28)_1$, with $R(v)$ given by (3.11) for $\gamma = p - 2$. We limit our attention to $n = 0, 1, 2, \cdots$ since the discussion of the case $n = -1, -2, \cdots$ would be completely symmetrical.

Under our assumption (3.31), the sequence $\{a_n\}$ is nonincreasing and so, as it may be seen from (3.30), some of the values of β_n will be positive while others will be negative. In fact it follows easily, by induction, that

$$\begin{aligned} &\beta_n(-1, \varepsilon_1, \cdots, \varepsilon_{n-1}) \geqq 0 \quad \text{for all } (\varepsilon_1, \cdots, \varepsilon_{n-1}), \\ &\beta_n(1, \varepsilon_1, \cdots, \varepsilon_{n-1}) \leqq 0 \quad \text{for all } (\varepsilon_1, \cdots, \varepsilon_{n-1}). \end{aligned} \tag{3.35}$$

Therefore, the sign of $\hat{\Phi}(n\tau, v)$ will depend on the outcome of the competition between terms with positive coefficients and terms with negative coefficients, in the summation (3.28). Using (3.30) we can show by induction that

$$\begin{aligned} &\sum_{(\varepsilon_1, \cdots, \varepsilon_{n-1})} \beta_n(-1, \varepsilon_1, \cdots, \varepsilon_{n-1}) = \frac{a_n + a_0}{2a_0}, \\ &\sum_{(\varepsilon_1, \cdots, \varepsilon_{n-1})} \beta_n(1, \varepsilon_1, \cdots, \varepsilon_{n-1}) = \frac{a_n - a_0}{2a_0}, \end{aligned} \tag{3.36}$$

and so the gross effect of positive coefficients dominates the gross effect of negative coefficients. From (3.36) it follows that

$$\sum_{(\varepsilon_0, \cdots, \varepsilon_{n-1})} \beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1}) = \frac{a_n}{a_0}, \tag{3.37}$$

$$\sum_{(\varepsilon_0, \cdots, \varepsilon_{n-1})} |\beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1})| = 1. \tag{3.38}$$

We write down $(3.28)_1$, with $R(v)$ given by (3.11), and we separate terms with positive and negative coefficients:

$$\begin{aligned} \hat{\Phi}(n\tau, v) = &\sum_{(\varepsilon_1, \cdots, \varepsilon_{n-1})} a_0 \beta_n(-1, \varepsilon_1, \cdots, \varepsilon_{n-1}) \left| v - \sum_{i=0}^{n-1} a_i\tau + \sum_{i=0}^{n-1} (1 + \varepsilon_i) a_i \tau \right|^\gamma \\ &+ \sum_{(\varepsilon_1, \cdots, \varepsilon_{n-1})} a_0 \beta_n(1, \varepsilon_1, \cdots, \varepsilon_{n-1}) \left| v - \sum_{i=0}^{n-1} a_i\tau + \sum_{i=0}^{n-1} (1 + \varepsilon_i) a_i \tau \right|^\gamma . \end{aligned} \tag{3.39}$$

We fix $\delta > 0$ so small that

$$\begin{aligned} &(1+\delta)^{-\gamma}(a_n + a_0) + (1+\delta)^\gamma(a_n - a_0) \geqq a_n, \\ &(1+\delta)^\gamma(a_n + a_0) + (1+\delta)^{-\gamma}(a_n - a_0) \leqq 3a_n, \end{aligned} \qquad n = 1, 2, \cdots . \tag{3.40}$$

Noting the elementary inequalities

$$(1+\delta)^{-\gamma} A^\gamma - (1+\delta^{-1})^\gamma B^\gamma \leqq (A+B)^\gamma \leqq (1+\delta)^\gamma A^\gamma + (1+\delta^{-1})^\gamma B^\gamma, \tag{3.41}$$

which hold for any positive numbers A, B, and using (3.36), we obtain from (3.39)

$$\begin{aligned}\hat{\Phi}(n\tau, v) \leqq \frac{3}{2} a_n \left| v - \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} \\ + (1+\delta^{-1})^{\gamma} a_0 \sum_{(\varepsilon_0,\cdots,\varepsilon_{n-1})} |\beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1})| \left[\sum_{i=0}^{n-1} (1+\varepsilon_i) a_i\tau \right]^{\gamma},\end{aligned} \tag{3.42}$$

$$\begin{aligned}\hat{\Phi}(n\tau, v) \geqq \frac{1}{2} a_n \left| v - \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} \\ - (1+\delta^{-1})^{\gamma} a_0 \sum_{(\varepsilon_0,\cdots,\varepsilon_{n-1})} |\beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1})| \left[\sum_{i=0}^{n-1} (1+\varepsilon_i) a_i\tau \right]^{\gamma}.\end{aligned} \tag{3.43}$$

We estimate the last term on the right-hand side of (3.42) and (3.43) by rearranging the terms in the summation and using (3.30), (3.38):

$$\begin{aligned}&\sum_{(\varepsilon_0,\cdots,\varepsilon_{n-1})} |\beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1})| \left[\sum_{i=0}^{n-1} (1+\varepsilon_i) a_i\tau \right]^{\gamma} \\ &= \sum_{k=0}^{n-1} \sum_{(\varepsilon_0,\cdots,\varepsilon_{k-1},1,-1,\cdots,-1)} |\beta_n(\varepsilon_0, \cdots, \varepsilon_{n-1})| \left[\sum_{i=0}^{n-1} (1+\varepsilon_i) a_i\tau \right]^{\gamma} \\ &= \sum_{k=0}^{n-1} \left\{ \left[\prod_{j=k+1}^{n-1} \frac{a_{j+1}+a_j}{2a_j} \right] \frac{a_k - a_{k+1}}{2a_k} \right. \\ &\qquad \left. \times \sum_{(\varepsilon_0,\cdots,\varepsilon_{k-1})} |\beta_k(\varepsilon_0, \cdots, \varepsilon_{k-1})| \left[\sum_{i=0}^{k} (1+\varepsilon_i) a_i\tau \right]^{\gamma} \right\} \\ &\leqq 2^{\gamma} \sum_{k=0}^{n-1} \left[\sum_{i=0}^{k} a_i\tau \right]^{\gamma} \frac{a_k - a_{k+1}}{2a_k}.\end{aligned} \tag{3.44}$$

The right-hand side of (3.44) is a priori bounded, independently of n, by virtue of (3.32) and (3.1) (recall $\gamma = p-2$). We have thus shown that there is a constant K such that

$$\hat{\Phi}(n\tau, v) \leqq \frac{3}{2} a_n \left| v - \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} + Ka_n, \qquad n = 1, 2, \cdots, \tag{3.45}$$

$$\hat{\Phi}(n\tau, v) \geqq \frac{1}{2} a_n \left| v - \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} - Ka_n, \qquad n = 1, 2, \cdots. \tag{3.46}$$

Since $\hat{\Psi}(n\tau, v) = \hat{\Phi}(n\tau, -v)$, (3.45), (3.46) yield

$$\hat{\Psi}(n\tau, v) \leqq \frac{3}{2} a_n \left| v + \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} + Ka_n, \qquad n = 1, 2, \cdots, \tag{3.47}$$

$$\hat{\Psi}(n\tau, v) \geqq \frac{1}{2} a_n \left| v + \sum_{i=0}^{n-1} a_i\tau \right|^{\gamma} - Ka_n, \qquad n = 1, 2, \cdots. \tag{3.48}$$

Letting $\tau \downarrow 0$, $\hat{\Phi}(u, v)$, $\hat{\Psi}(u, v)$ converge to the solution $\Phi(u, v)$, $\Psi(u, v)$ of (3.7), (3.8), (3.11). Hence

$$\begin{aligned}\Phi(u, v) &\leqq \frac{3}{2} a(u) \left| v - \int_0^u a(w)\, dw \right|^{\gamma} + Ka(u), \\ \Psi(u, v) &\leqq \frac{3}{2} a(u) \left| v + \int_0^u a(w)\, dw \right|^{\gamma} + Ka(u),\end{aligned} \tag{3.49}$$

$$\begin{aligned}\Phi(u,v) &\geqq \frac{1}{2}a(u)\left|v-\int_0^u a(w)\,dw\right|^\gamma - Ka(u),\\ \Psi(u,v) &\geqq \frac{1}{2}a(u)\left|v+\int_0^u a(w)\,dw\right|^\gamma - Ka(u)\end{aligned} \tag{3.50}$$

on $(-\infty,\infty)\times(-\infty,\infty)$. Upon adding to the above solution $K+1$ times the special solution (3.16), we end up with a new solution of (3.7) which is positive on $(-\infty,\infty)\times(-\infty,\infty)$. This induces, via (3.3), (3.7) and (1.7), a strictly convex entropy $\eta(u,v)$ for (1.3), which satisfies the growth conditions (3.33), (3.34) (recall $p=\gamma+2$). This completes the proof of the proposition. □

When $p=4$, (3.32) reduces to condition (3.20), which, as we have seen, is both necessary and sufficient for the existence of positive $\Phi(u,v)$, $\Psi(u,v)$ with quadratic growth rates. Thus (3.32) appears to be sharp.

Let $(\bar{u},\bar{v})$ be any constant state and let $\bar{\eta}(u,v)$ be the positive entropy induced by $\eta(u,v)$ via (1.8). It then follows from (3.33), (3.34) that if $\{f(u,v), g(u,v)\}$ satisfies

$$f(\bar{u},\bar{v})=g(\bar{u},\bar{v})=0, \tag{3.51}$$

$$|f(u,v)|+|g(u,v)|\leqq A(|u|+|v|+1), \quad -\infty<u<\infty, \quad -\infty<v<\infty, \tag{3.52}$$

then there is a positive constant C with

$$|\bar{\eta}_u(u,v)f(u,v)+\bar{\eta}_v(u,v)g(u,v)|\leqq CA\bar{\eta}(u,v), \tag{3.53}$$

for $-\infty<u<\infty$, $-\infty<v<\infty$. Therefore, applying Gronwall's inequality to (1.10) yields

$$\int_{-\infty}^{\infty}\bar{\eta}(u(x,t),v(x,t))\,dx\leqq e^{CAt}\int_{-\infty}^{\infty}\bar{\eta}(u(x,0),v(x,0))\,dx. \tag{3.54}$$

By virtue of (3.33) and (1.8), (3.54) induces L^p bounds, independent of ν, on solutions of the Cauchy problem for (1.1) with initial data in $L^p(-\infty,\infty)\cap L^2(-\infty,\infty)$.

REFERENCES

[1] K. N. Chueh, C. C. Conley and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 372-411.

[2] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27-70.

[3] ———, *Convergence of the viscosity method for isentropic gas dynamics*, Comm. Math. Phys., 91 (1983), pp. 1-30.

[4] P. D. Lax, *Shock waves and entropy*, in Contributions to Functional Analysis, E. A. Zarantonello, ed., Academic Press, New York, 1971, pp. 603-634.

[5] V. Roytburd and M. Slemrod, *Positively invariant regions for a problem in phase transitions*, Arch. Rational Mech. Anal., 93 (1986), pp. 61-79.

[6] M. Rascle, *Un résultat de "compacité par compensation à coefficients variables". Application à l'élasticité nonlinéaire*, Compt. Rend. Acad. Sci. Paris, Série I, 302 (1986), pp. 311-314.

[7] D. Serre, *Domaines invariants pour les systèmes hyperboliques de lois de conservation*, preprint.

[8] T. D. Venttsel', *Estimates of solutions of the one-dimensional system of equations of gas dynamics with "viscosity" nondepending on "viscosity,"* Soviet Math. J., 31 (1985), pp. 3148-3153.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

011

OSCILLATING SOLUTIONS OF THE FALKNER-SKAN EQUATION FOR NEGATIVE β*

S. P. HASTINGS† AND W. C. TROY†

Abstract. The Falkner-Skan equation results from an appropriate similarity substitution in the Prandtl boundary layer equations. New solutions are found analytically, including a periodic solution and many solutions which oscillate a finite number of times and then tend to a limit.

Key words. boundary layer theory, similarity solutions

AMS(MOS) subject classifications. 76D10, 34B15, 76D30

1. Introduction. The boundary layer technique of Prandtl [13] has become one of the cornerstones of modern fluid dynamics and aerodynamics. One aspect of the method consists of finding particular solutions of approximations to the Navier-Stokes equations by assuming that a solution within the boundary layer has a "similarity" form. Using these solutions as first approximations, one then employs the method of matched asymptotic expansions to solve the boundary layer equations more generally. General references to these techniques include [2], [12], [15] and [16].

Among the first to look for similarity solutions were Blasius [1] and Falkner and Skan [3]. These authors considered laminar incompressible boundary layer flow past a flat plane or a wedge. They began with the steady-state Prandtl boundary layer equations, which we write as follows:

$$u_x + v_y = 0, \qquad uu_x + vu_y = UU_x + \gamma u_{yy}. \tag{1}$$

Here (x, y) are orthogonal coordinates in the boundary layer, with x representing arc length along the wall and y the perpendicular distance from the wall, u and v are the corresponding velocity components, and γ is the kinematic viscosity. Also, $U = U(x)$ is the assigned exterior stream velocity. The appropriate boundary conditions are

$$u = v = 0 \quad \text{when } y = 0, \qquad u \to U(x) \quad \text{as } y \to \infty. \tag{2}$$

Their investigations led to a classical fluid dynamical model which can be written in the form

$$f''' + ff'' + \beta(1 - f'^2) = 0. \tag{3}$$

Here f is proportional to the stream function and $f' = u/U$. For our purposes a similarity assumption proposed by Spalding [17] and studied extensively by Evans [2] is convenient in deriving (3) from the Prandtl equations. They assume that U satisfies the equation

$$\frac{dU}{dx} = CU^{2(\beta-1)/\beta}, \tag{4}$$

where x is the coordinate in the direction of the stream as measured along the bounding surface and C is a constant. Thus β measures the pressure gradient in the stream direction. With this interpretation negative values of β less than -0.5 and all $\beta > -0.199$ are found by Evans [2] to be of physical interest. This is to be contrasted to the original

* Received by the editors April 1, 1985, and in revised form August 1, 1986. This research was partially supported by the National Science Foundation.

† Department of Mathematics, State University of New York, Buffalo, New York 14214.

derivation of Falkner and Skan, in which U was assumed to be proportional to x^m, producing (3) with $\beta = 2m/(m+1)$. In this original setting only $0 \leqq \beta < 2$ was found to be significant. We study solutions with $\beta > 0$ in a companion paper ([9]; see also [8]).

Small negative values of β, between the "separation" value of -0.199 and 0, were introduced, with physical justification, by Hartree [6] and later by Stewartson [18]. Here we are interested primarily in values of $\beta < -1$, which is included in the range of physical interest as described by Evans. In this range there are two sets of boundary conditions for (3) which have been been considered [2], [6], [18]. These are

$$f(0) = f'(0) = 0, \qquad f'(\infty) = 1, \tag{5}$$

and

$$f(0) = f'(0) = 0, \qquad f'(\infty) = -1. \tag{6}$$

In most known applications f' is restricted to lie between -1 and $+1$, and then solutions to (3)-(5) only exist in a small range of negative β [5], [7]. However it has been suggested that physically interesting solutions may be possible in which f' exhibits "overshoot," by rising above 1, and recently some interesting numerical studies have been done of solutions of this kind [10], [11], [14].

An initial analytical investigation of solutions with overshoot, for large negative β, was done by Troy [19]. In this paper we prove the existence of many new solutions of (3), satisfying either (5) or (6). The existence of these new solutions depends on showing that (3) has a periodic solution if $\beta < -1$. This was first suggested by some numerical computations in [14].

Our method for proving the existence of these solutions is "shooting." That is, we consider an initial value problem for (3), such as

$$f(0) = f'(0) = 0, \qquad f''(0) = \gamma, \tag{7}$$

and vary γ to satisfy the desired condition at ∞. It turns out that a more fundamental initial value problem is (3) coupled with the initial conditions

$$f(0) = f''(0) = 0, \qquad f'(0) = \alpha, \tag{8}$$

with $-1 < \alpha < 0$. We denote the unique solution to (3), (8) by f_α, and will show that if $\beta < -1$, then α can be chosen so that f_α is periodic. From this we will obtain solutions of (3), (8) with either $f'(\infty) = 1$ or $f'(\infty) = -1$. Each of these conditions is of physical interest [16]. Using these solutions, we then study the initial value problem (3), (7), and again find solutions with $f'(\infty) = 1$ or $f'(\infty) = -1$. In this case, we find a solution which is not periodic but which oscillates infinitely often.

2. Statement of results. The first four results refer to the initial value problem (3), (8). They make it clear that there is a complicated bifurcation "from infinity in phase space" as β crosses -1 from above.

THEOREM 1. *For any $\beta < -1$ there is an $\bar{\alpha}$ in $(-1, 0)$ such that $f_{\bar{\alpha}}$ is periodic with some period P and $f'_{\bar{\alpha}}$ has exactly one local maximum in $(0, P)$. If $-1 \leqq \beta \leqq 0$ then there is no periodic solution of (3) except $f \equiv 0$.*

THEOREM 2. *If $\beta < -1$, then there is a sequence $\{\alpha_j\} \subset (-1, \bar{\alpha})$, $j \geqq 0$, tending to $\bar{\alpha}$ from below, such that $f'_{\alpha_j}(\eta)$ tends to -1 as η tends to infinity, and f'_{α_j} has exactly j local maxima in $0 < \eta < \infty$.*

The next result discusses solutions such that f' tends to $+1$. It is more complicated to state, because branches of solutions of (3), (5) (plotted in the β-f'' plane) move into the region $f'' > 0$, according to the numerical computations in [10] and [14], as β decreases from -1. It is convenient to begin with a result in which β is varied and the fixed initial conditions $f(0)=f'(0)=f''(0)=0$ are considered. In accordance with previous notation, we denote the corresponding solution by f_0.

THEOREM 3. *For each $j \geqq 0$, let $S_j = \{\beta \mid f'_0$ tends to $+1$ with exactly j relative maxima and minima$\}$. Then every negative β lies in some S_j, each S_j is bounded below, and there is a decreasing sequence $\{\beta_j\}$ tending to $-\infty$ such that β_j is contained in the interior of S_j.*

We then have a second result about solutions with f' tending to $+1$.

THEOREM 4. *Suppose that $\beta \in S_j$. Let $\bar{\alpha}$ be as defined in Theorem 1. Then there is a decreasing sequence α_k^*, $k \geqq j$, with $\alpha_0^* = 0$ and $\alpha_j^* \geqq \bar{\alpha}$, such that if $\alpha = \alpha_k^*$, then f' has exactly k relative maxima and minima (f'' vanishes exactly k times), and then tends to* 1 *at an exponential rate.*

It should be observed that we have not proved that the α_j^* necessarily tend to $\bar{\alpha}$.

We now consider solutions satisfying the initial conditions (7). We denote the unique solution to (3), (7) by F_γ. The results here parallel to some extent those for the initial conditions (8), except that there does not appear to be a periodic solution.

THEOREM 5. *For each $\beta < -1$ there is a $\bar{\gamma} < 0$ such that $F_{\bar{\gamma}}$, $F'_{\bar{\gamma}}$, and $F''_{\bar{\gamma}}$ all vanish infinitely often on $(0, \infty)$ and do not tend to a limit.*

THEOREM 6. *There is an increasing sequence γ_j, $j \geqq 0$, contained in $(-\infty, \bar{\gamma})$ such that F''_{γ_j} vanishes exactly $2j$ times and then $(F'_{\gamma_j}, F''_{\gamma_j})$ tends to $(-1, 0)$. Thus F_{γ_j} solves* (3), (6).

THEOREM 7. *Suppose $\beta \in S$. Then there is a decreasing sequence γ_k^*, $k \geqq j$, in $(\bar{\gamma}, 0)$ such that if $\gamma = \gamma_k^*$, then $F = F'_\gamma$ has exactly k relative maxima and minima (F'' vanishes k times) and then F' tends to* 1 *exponentially fast. Thus F_{γ_j} solves* (3), (5).

3. Proofs.

Proof of Theorem 1. We first verify that there are no periodic solutions if $-1 \leqq \beta \leqq 0$. (We are not considering any particular initial conditions at this point.) Suppose that for some β in this range there is a periodic solution, with period $P > 0$. We may assume that $f'(0) = 0$. An integration of (3) shows that

$$f'' + ff' = -\beta\eta + (\beta+1)\int_0^\eta (f'(s))^2\, ds + f''(0). \tag{9}$$

At $\eta = P$, under the restriction on β, this a contradiction, since we obtain $f''(P) > f''(0)$.

Henceforth we assume that $\beta < -1$. Let f_α be the solution of (3), (8). We try to choose $\bar{\alpha}$ so that for some $Q > 0$,

$$f(Q) = f''(Q) = 0, \tag{10}$$

where $f = f_{\bar{\alpha}}$. This implies that $f(Q+\eta) = -f(Q-\eta)$, and therefore f is periodic with period $2Q$.

The existence of $\bar{\alpha}$ is proved by shooting. Suppose first that $\alpha = 0$. Then, from (3), f, f' and f'' all are positive on some interval $(0, \varepsilon)$. Therefore for small negative α, while f is initially negative, $f = 0$ before $f'' = 0$. That is, there is an η_0 such that $f'' > 0$ on $(0, \eta_0]$ while $f(\eta_0) = 0$.

From now on we consider only values of α in $(-1, 0)$. Solutions of (3), (8) are continuous in α. Since f is negative and decreasing as long as f' is negative, it follows that f'' is positive up to the first zero of f, if this exists, and at this zero, f' is positive. Also, at the first zero of f'', f''' must be nonzero, for if f'' and and f''' vanish simultaneously, then $1 - f'^2 = 0$ and f'' is identically zero. Therefore, the set of α such that $f_\alpha = 0$ before $f''_\alpha = 0$ is open in the interval $(-1, 0)$, as well as nonempty. Similarly, the set of α such that $f'' = 0$ before $f = 0$ is also open. To obtain condition (10) for some α and $Q > 0$, we must prove that the latter set is nonempty. This follows from the following lemma.

LEMMA 1. *If α is sufficiently close to -1, then f'' becomes zero at least as soon as f.*

Proof. Suppose, instead, that for all α in $(-1, 0)$ there is a $p > 0$ such that $ff'' < 0$ on $(0, p)$ and $f(p) = 0$. Then (9) becomes

$$f''(p) = \lambda p - (\lambda - 1) \int_0^p [f'(\mu)]^2 \, d\mu \tag{11}$$

where $\lambda = -\beta > 1$. A contradiction will be obtained by showing that the right side of (11) is negative for α close to -1. For this we use the following two technical lemmas.

LEMMA 2. *Choose a fixed δ in $(0, 1)$ and assume that $\alpha < -1 + \delta$. Then there is a first $a > 0$ such that $f'(a) = -1 + \delta$ and a first $b > a$ such that $f'(b) = 0$. Further, a and $f''(a)$ tend to infinity as α tends to -1, and $b - a \leqq (1 - \delta)/\delta a$.*

Proof. The existence of a and b is assured since f_α''' is positive as long as $f' < 1$. On $(0, a)$, $\alpha < f' - 1 + \delta$. Therefore, using (9) with $\beta < -1$, we find that $f''(a) \geqq \delta a$. Also, since $f' \equiv -1$ solves (3), it is clear that $a \to \infty$ as $\alpha \to -1+$. To complete the lemma, we observe that $f'' \geqq \delta a$ on (a, b), since $f''' > 0$ there. Therefore

$$1 - \delta = f'(b) - f'(a) \geqq \delta a(b - a)$$

as desired.

In the remainder of the paper we consider a number M such that

$$M > 200\lambda/(\lambda - 1)(1 - \delta). \tag{12}$$

(This inequality is used in the proof of Theorem 2, but for Theorem 1 we only need $M > 6\lambda/(\lambda - 1)(1 - \delta)$.)

LEMMA 3. *If α is sufficiently close to -1, then there is a first $c > a$ such that $f'(c) = M$, and $c - a \to 0$ as $\alpha \to -1+$.*

Proof. Suppose that $f' < M$ on the interval $(b, b + 1)$ for all small values of $1 + \alpha$. Then (3) and Lemma 2 imply that on this interval, $f'' \geqq e^{-M}\delta a/2 - \lambda(M^2 - 1)/M$. Since a tends to infinity as α tends to -1, we see that for sufficiently small $(1 + \alpha)$,

$$f'' \geqq e^{-M}\delta a/2 > M. \tag{13}$$

But this implies that $f'(b + 1) > M$, a contradiction. Therefore for α close to -1 there is a c in $(b, b + 1)$ such that $f'(c) = M$, and (13) holds over (b, c). This implies that $c - b \leqq 2Me^M/\delta a$, and this with Lemma 2 completes the proof of Lemma 3.

We now continue with the proof of Lemma 1. We see that f is negative in $[b, c]$, so $p > c$ and $f'(p) > M$ for α close to -1. On $(0, a)$,

$$\alpha a \leqq \int_0^a f'(s) \, ds \leqq (-1 + \delta)a. \tag{14}$$

Also, $0 = \int_0^p f'(s) \, ds = \int_0^a f'(s) \, ds + \int_a^c f'(s) \, ds + \int_c^p f'(s) \, ds$. Using Lemma 2, we obtain

$$-\alpha a \geqq \int_c^p f'(s) \, ds + o(1) \geqq (1 - \delta)a \tag{15}$$

as $\alpha \to -1+$. Further, $f' > M$ on (c, p), so

$$\int_0^p f'(s)^2 \, ds \geqq M \int_c^p f'(s) \, ds \geqq M(1 - \delta)a/2 \tag{16}$$

for α close enough to -1.

Next we obtain an upper bound on p. From (15) it follows that $-\alpha a \geqq M(p - c) + o(1)$ as $\alpha \to -1+$. We also know that $c - a \to 0$ as $\alpha \to -1+$. Therefore for α sufficiently close to -1,

$$p \leqq 2a - \alpha a/M. \tag{17}$$

To complete the proof of Lemma 1 and Theorem 1, we use (11), (12), (16) and (17) to obtain

$$f''(p) \leqq 2a\lambda + \lambda a/M - (\lambda - 1)(1-\delta)aM/2 < 0.$$

Proof of Theorem 2. Let $\bar{\alpha}$ be a point chosen as above so that $f_{\bar{\alpha}}$ is periodic. (We do not claim that $\bar{\alpha}$ is unique.) Define a set W_1 as follows.

$$W_1 = \{\alpha \text{ in } (-1, 0) \text{ such that if } -1 < f'(0) < \alpha, \text{ then } f'(\bar{\eta}) = -1$$
$$\text{for some first } \bar{\eta} > 0, f''(\bar{\eta}) < 0, \text{ and } f'' = 0 \text{ exactly once in } (0, \bar{\eta})\}.$$

LEMMA 4. *The set W_1 is nonempty and open.*

Proof of Lemma 4. Since f'' and f''' cannot vanish simultaneously, unless $f'' \equiv 0$, W_1 is easily seen to be open. Let η_0 be the first zero of f''. We have shown that if $\alpha + 1$ is small, then $f'(\eta_0) > M$. In addition, define a number K by

$$K = -3\lambda + (\lambda - 1)(1-\delta)M/2.$$

Finally, choose α so close to -1 that

$$a > \max\{e^M(8\lambda + 16M)/\delta, (8\lambda + M)/K\} \tag{18}$$

where a and δ are as in the proof of Lemma 1.

We now observe that if, as previously, p is the first positive zero of f, then

$$f''(p) < -Ka. \tag{19}$$

This is seen from (9), with $\eta = p$, (16) and (17).

We now let η_1 denote the first point beyond $\eta = \eta_0$ where $f' = M$. There are now two cases to consider:

(i) $$f < 0 \quad \text{on } (\eta_0, \eta_1),$$

and

(ii) $$f(p) = 0 \quad \text{at some } p \text{ in } (\eta_0, \eta_1).$$

In considering case (i) we introduce the fundamental "energy" function

$$H = \tfrac{1}{2}f''^2 + \beta(f' - f'^3/3),$$

and we see that $H' = -ff''^2$. Therefore H is increasing in $(0, \eta_1)$, and in particular, $H(\eta_1) > H(c)$, where, as before, $f'(c) = M$. Because $M > 1$, this implies that

$$f''(\eta_1) \leqq -f''(c) \leqq -\bar{\delta}a,$$

where $\bar{\delta} = \frac{1}{2}\delta\, e^{-M}$. If $f' > 0$ on $(\eta_1, \eta_1 + 1)$, then (9) and (18) imply that $f'' < -\bar{\delta}a/2$ on this interval. Then (18) implies that, in fact, f' must equal zero for some η_2 in $(\eta_1, \eta_1 + 1)$.

Similarly, again using (9) and (18), we show that if $f' > -1$ on $(\eta_2, \eta_2 + 1)$, then on this interval

$$f'' \leqq -\bar{\delta}a/4,$$

which again leads to a contradiction. It is apparent from these inequalities that $f'' < 0$ past η_0 until $f' = -1$, and this shows W_1 is nonempty if case (i) holds.

To deal with case (ii), we use (9), (12) and (19) to show that

$$f''(\eta) \leqq -Ka \tag{20}$$

in the interval (p, η_1). This, with (18), leads to

$$f'' \leqq -Ka + \lambda \leqq -Ka/2 \leqq -M \tag{21}$$

on (η_1, η_1+1), if $f'>0$ on this interval. Since $f'(\eta_1)=M$, it is seen that $f'=0$ at some next $\eta_2<\eta_1+1$.

We now want to show that $f'=-1$ before $\eta=\eta_2+1$. This is more complicated than previously, and requires that we find a bound on $f(\eta_2)$. It will also use (12).

Integrate (9) once to give

$$f'+f^2/2 \leqq \lambda\eta^2/2.$$

This inequality, with $\eta=p$ and then with $\eta=\eta_2$, shows that

$$f'(p) \leqq \lambda p^2/2 \tag{22}$$

and

$$f(\eta_2) \leqq \lambda^{1/2}\eta_2. \tag{23}$$

But (21) holds in (p, η_2), so we find that

$$\eta_2 \leqq p+2f'(p)/Ka. \tag{24}$$

From (17), (23), and (24), and the definitions of K and M we conclude that

$$f(\eta_2) \leqq 12\lambda^{1/2}a. \tag{25}$$

Proceeding in a similar way, we can show that if $-1<f'<0$ in (η_2, η_2+1), then in that interval

$$f'' \leqq (12\lambda^{1/2}-K/4)a+\lambda \leqq -1,$$

which leads to a contradiction and so completes the proof of Lemma 4.

Continuing with the proof of Theorem 2, recall that $\bar{\alpha}$ was chosen so that $f_{\bar{\alpha}}$ is periodic, and furthermore, $f'_{\bar{\alpha}}>-1$ everywhere. Therefore if α is sufficiently close to $\bar{\alpha}$, then f'' must vanish at least twice before $f'=-1$. Hence $\alpha_1=\sup W_1$ is in the interval $(-1, \bar{\alpha})$. We wish to show that the solution f_{α_1} has the properties ascribed to it in the statement of Theorem 2. We need the following key result, which is also used later.

LEMMA 5. *Let E denote the set*

$$E=\{(f,f',f'')\,|\,f>0,\ 0\leqq f'\leqq 3^{1/2},\ (f'')^2\leqq -2\beta(f'-\tfrac{1}{3}f'^3)\}.$$

Then E is positively invariant for the natural first order system equivalent to (3), *and if f is a solution such that (f,f',f'') enters E, then $(f',f'')\to(1,0)$ as $\eta\to\infty$.*

Proof. Positive invariance follows because $H=0$ and $H'>0$ on the part of the boundary of E where $f>0$, and since $f'\geqq 0$ in E. The other part of the result is also clear, because H decreases in E and the minimum of H, and its only stationary point in E, is at $(1,0)$.

Now let $f=f_{\alpha_1}$. From the definition of α_1, and the fact that f'' and f''' cannot vanish simultaneously, we see that f'' cannot vanish more than once. On the other hand, suppose that f'' does not vanish at all. If f' becomes larger than 1, then (3) leads to a contradiction. One can therefore show that $(f',f'')\to(1,0)$, which implies that (f,f',f'') enters E. However then nearby solutions also enter E, which again contradicts the definition of α_1. Therefore f'' vanishes exactly once. However f' cannot fall below -1, and it follows easily that when $\alpha=\alpha_1$, $(f',f'')\to(-1,0)$, as desired.

Now there may be other values of α, between α_1 and $\bar{\alpha}$, such that f_α behaves in the same way. However we have seen that for α close enough to $\bar{\alpha}$, f'' must vanish more than once. Therefore the set

$$V_1=\{\alpha \text{ in } (\alpha_1,\bar{\alpha})\,|\,f'' \text{ vanishes once in } (0,\infty), \text{ and } (f',f'')\to(-1,0)\}$$

has a supremum, say $\bar{\alpha}_1$, in $[\alpha_1,\bar{\alpha})$. Let

$$W_2=\{\alpha \text{ in } (\bar{\alpha}_1,\bar{\alpha})\,|\,f'' \text{ vanish twice, after which } f' \text{ falls below } -1\}.$$

Arguing as above, we find the solution f_{α_2} of Theorem 2, and a routine induction completes the proof of this result.

Proof of Theorem 3. We first consider the solution f_0 when $\beta=-1$. A number of papers have been devoted to just this case. A recent paper contains the result we want.

LEMMA 6 [4]. *If $\beta=-1$ and $f(0)=f'(0)=f''(0)=0$, then f' increases to above* 1 *and then decreases to* 1, *while f'' has only one positive zero. Furthermore, $f'\to 1$ at an algebraic rate.*

Next we observe that for any negative β, the vector (f_0,f_0',f_0'') immediately (after $\eta=0$) enters and remains in the set E. By Lemma 5, f_0' tends to 1.

Now let

$$\bar{\beta}_1=\inf\{\beta^*\,|\,\text{if } \beta^*<\beta<-1 \text{ then } f_0'\to 1 \text{ with exactly one local maximum and no local minima}\}.$$

Since f'' and f''' cannot vanish simultaneously, we see that $\bar{\beta}_1$ lies in S_1. Furthermore, the methods of Troy in [19] show that for this value of β, $f'\to 1$ at an exponential rate.

LEMMA 7. *For β just below $\bar{\beta}_1$, f_0'' has exactly one* max *and one* min *before $f_0'\to 1$.*

Proof. Because f_0 is continuous in β, we see that for any $\varepsilon>0$, there is a $\delta>0$ such that if $0<\bar{\beta}_1-\beta<\delta$, then $|f'-1|<\varepsilon$ and $f>1/\varepsilon$ before f_0' crosses 1 for the second time. In a neighborhood of $f'=1$ solutions of (3) behave like (translates of) solutions of Weber's equation

$$y''+\eta y'+2\lambda y=0, \tag{26}$$

where $y\approx f'-1$. It is known that there is a $B>0$, depending on λ, such that no solution of (26) can have two successive zeros in the region $\eta>B$. This implies that f' cannot vanish twice in any region where $f>B$. From this we see that if $\bar{\beta}_1-\beta$ is sufficiently small, then f_0 can have no more than two crossings of $f'=1$. This proves Lemma 7.

In a similar manner one inductively completes the proof of Theorem 3.

Proof of Theorem 4. This proceeds in the same way as Theorem 3. Consider a fixed β in S_j. If $\alpha=0$, then f' tends to 1 with j relative maxima and minima. Let

$$\alpha_j=\inf\{\alpha\,|\,f_\alpha' \text{ tends to 1 after exactly } j \text{ relative maxima and minima}\}.$$

Then just as in the proof of Theorem 3 we show that f_{α_j} has the desired properties, and starting just to the left of α_j we construct α_{j+1}. We know that as $\alpha\to\bar{\alpha}$, $f_\alpha'-1$ acquires more and more zeros, and the result easily follows.

We now turn to the results which concern the initial conditions $f=f'=0$. As before, the solution to this problem, with $f''(0)=\gamma$, will be denoted by F_γ. For the sake of brevity we merely outline the proofs of Theorems 5-7.

We first consider *a* fixed β, which need only be negative.

LEMMA 8. *If $-\gamma$ is sufficiently large, then F_γ' decreases monotonically to below* -1.

This is proved by easy estimates which we omit. Let $\gamma_0=\sup\{\gamma\,|\,F_\gamma'$ decreases monotonically to below $-1\}$. Since we know $F_0'=f_0'$ and tends to $+1$, it is clear that $\gamma_0=0$. Further, F_{γ_0}'' cannot have a zero, since then F_γ'' would have a zero for γ close to γ_0. It easily follows that F_{γ_0}' tends monotonically to -1.

LEMMA 9. *F_{γ_0} is in fact the only solution to* (3), (7) *such that F' tends monotonically to* -1.

Proof. Suppose that F and G are two such solutions, and assume that $F'(0) < G'(0) < 0$. Then $F'''(0) < G'''(0)$, so initially, $F'' < G''$. Suppose that at some η, $F'' = G''$. It is easily seen that then $F''' < G'''$, a contradiction.

It is now clear that if γ is close to, but larger than, γ_0, then F'_γ comes close to, but does not cross, -1. (Otherwise there would be a second solution with the properties of F_{γ_0}.) As in the proof of Lemma 4 it then follows that for such γ, F' first has a negative local minimum, then a positive local maximum, and then falls monotonically below -1. It this way we obtain γ_1, γ_2, etc. as in the proof of Theorem 2. That is, we complete the proof of Theorem 6.

Theorem 7 is obtained in a similar manner. Finally we can obtain Theorem 5 by letting γ be the supremum of the γ_j of Theorem 6 or the infinimum of the γ_j^* of Theorem 7. Note that we cannot prove that these give the same value of γ.

REFERENCES

[1] H. BLASIUS (1908), *Grenzschichten in Flussigkeiten mit kleiner Reibung*, Z. Math. Phys., 58, pp. 1-37.

[2] H. L. EVANS (1968), *Laminar Boundary Layer Theory*, Addison-Wesley, Reading, MA.

[3] V. M. FALKNER AND S. W. SKAN (1931), *Solutions of the boundary layer equations*, Phil. Mag., 7/12, pp. 865-896.

[4] B. GABUTTI (1984), *An existence theorem for a boundary value problem related to that of Falkner and Skan*, this Journal, 15, pp. 943-995.

[5] P. HARTMAN (1964), *On the asymptotic behaviour of solutions of a differential equation in boundary layer theory*, Z. Angew. Math. Mech., 44, pp. 123-138.

[6] D. R. HARTREE (1949), *A solution of the laminar boundary layer equation for retarded flow*, ARCRM, p. 2426.

[7] S. P. HASTINGS (1971), *An existence theorem for a class of nonlinear boundary value problems including that of Falkner and Skan*, J. Differential Equations, 9, pp. 580-593.

[8] S. P. HASTINGS AND W. TROY (1984), *Oscillatory Solutions of the Falkner-Skan Equation*, Proc. Royal Soc. London, to appear.

[9] S. P. HASTINGS AND W. TROY, *Oscillatory solutions of the Falkner-Skan equation for positive β*, J. Differential Equations, to appear.

[10] C. LAINE AND L. REINHART, *Further numerical methods for the Falkner-Skan equations: shooting and continuation techniques*, Numerical Methods in Fluids, to appear.

[11] P. A. LIBBY AND T. M. LIU (1967), *Further solutions of the Falkner-Skan equation*, AIAA J., 5, pp. 1040-1042.

[12] D. MEKSYN (1961), *New Methods In Laminar Boundary-Layer Theory*, Pergamon Press, Oxford.

[13] L. L. PRANDTL (1935), *Mechanics of viscous fluids*, in Aerodynamics Theory, W. F. Daniel, ed., Berlin, 3, pp. 34-208.

[14] B. OSKAM AND A. E. P. VELDMAN (1982), *Branching of The Falkner-Skan solutions for $\lambda < 0$*, J. Engrg. Math., 36, pp. 295-307.

[15] L. ROSENHEAD (ed.) (1963), *Laminar Boundary Layers*, especially Chap. VI.

[16] H. SCHLICHTING (1979), *Boundary Layer Theory*, 7th ed., McGraw-Hill, New York, p. 29.

[17] D. B. SPALDING (1961), *Mass transfer through laminar boundary layers—1. The velocity boundary layer*, Int. J. Heat Mass Transfer, 83, p. 483.

[18] K. STEWARTSON (1954), *Further solutions of the Falkner-Skan equation*, Proc. Camb. Phil. Soc., 50, pp. 454-565.

[19] W. C. TROY (1977), *Non-monotonic solutions of the Falkner-Skan boundary layer equation*, Quart. J. Appl. Math., 37, p. 157.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

012

SOLUTION, GRADIENT, AND LAPLACIAN BOUNDS IN SOME NONLINEAR FOURTH ORDER ELLIPTIC EQUATIONS*

PHILIP W. SCHAEFER†

Abstract. In order to obtain bounds of the type in the title, a suitable function is defined on the solutions to a certain class of semilinear fourth order elliptic partial differential equations. The subharmonicity of this function under appropriate conditions on the coefficients and nonlinear terms leads one immediately to the desired bounds.

Key words. elliptic equations, maximum principles, pointwise bounds

AMS(MOS) subject classifications. 35B45, 35B50

1. Introduction. The technique of defining a function on the solution of a differential equation and deducing results about the solution of the equation by means of the auxiliary function is well known. In [4], Payne used this idea to deduce gradient bounds on the solution in the torsion problem. There he utilized the Hopf maximum principles [9] to establish that the maximum of the auxiliary function occurred at a critical point of the solution in the domain of definition. Payne and several other authors [5], [7], [8], [10], [11] and references therein have extended his idea to the determination of bounds on the gradient of the solution to more general second order elliptic partial differential equations which are subject to boundary conditions of various kinds. One can find an exposition of some of these extensions, generalizations, and applications of Payne's method in Sperb's book [13].

The aforementioned technique has also been used in the study of fourth order elliptic partial differential equations. Dunninger [2] showed that, although there is no maximum principle for the solution of fourth order elliptic equations, any nontrivial solution of

$$\Delta^2 u + cu = 0, \qquad c > 0 \quad \text{in } D \subset R^n,$$

for which $\Delta u = 0$ on ∂D, satisfies the inequality

$$|u(x)| < |u(x_0)|, \qquad x \in D,$$

where x_0 is some point on the boundary ∂D of the bounded domain D. This was accomplished by showing that a suitably defined auxiliary function was a nonconstant subharmonic function. The result was extended to allow a variable coefficient in place of c and to deduce a related result in the case of metaharmonic equations in [1].

One can give any of a number of other illustrations of this technique. Here we shall employ this method in the study of nonlinear fourth order equations of the form

$$\Delta^2 u - q(x) g(\Delta u) + p(x) f(u) = 0. \tag{1.1}$$

By means of an appropriately defined auxiliary function defined on solutions of the equation, we shall develop maximum principles and deduce bounds on the solution, gradient of the solution, and the Laplacian of the solution. The principles are presented in § 2 and some immediate applications are obtained in § 3.

* Received by the editors April 22, 1985, and in revised form July 11, 1985.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300.

2. Principles. In [6] Payne obtained some specialized results from the subharmonicity of certain auxiliary functions defined on the solutions of the equation

$$\Delta^2 u = f(u),$$

where Δ is the Laplace operator and Δ^2 is the iterated Laplacian. Here we introduce an auxiliary function for a more general equation which allows us to deduce other bounds of importance in applications.

Let D be a bounded domain in Euclidean n-space and let $u \in C^4$ be a solution of the equation

$$\Delta^2 u - q(x)g(\Delta u) + cf(u) = 0, \tag{2.1}$$

in D, where we initially assume the coefficients satisfy

$$q(x) \geqq 0, \quad c > 0, \quad c \text{ constant}, \tag{2.2}$$

and the nonlinear functions satisfy the requirements

$$sg(s) \geqq 0, \qquad f'(u) \geqq \beta > 0, \tag{2.3}$$

where the prime indicates differentiation with respect to u.

We define the function

$$V(x) = u_{,i}u_{,i} + \gamma(\Delta u)^2 + 2\gamma c F(u), \tag{2.4}$$

where γ is a positive constant to be chosen and $F(u)$ is a primitive of the function f, i.e.,

$$F(u) = \int_0^u f(t)\, dt.$$

In (2.4) we have used the comma notation to indicate partial differentiation with respect to x_i and the summation convention, i.e., repeated indices in a term signifies summation from 1 to n.

By a straightforward calculation, we have

$$\begin{aligned}\Delta V = {} & 2u_{,ij}u_{,ij} + 2u_{,i}(\Delta u)_{,i} + 2\gamma(\Delta u)(\Delta^2 u) + 2\gamma(\Delta u)_{,i}(\Delta u)_{,i} \\ & + 2\gamma cf(u)\Delta u + 2\gamma cf'(u)u_{,i}u_{,i}.\end{aligned}$$

Now by (2.1) and the addition and subtraction of $u_{,i}u_{,i}$ and $(\Delta u)_{,i}(\Delta u)_{,i}$ we can write

$$\begin{aligned}\Delta V = {} & 2u_{,ij}u_{,ij} + |(\Delta u)_{,i} + u_{,i}|^2 + 2\gamma q(x)(\Delta u)g(\Delta u) \\ & + (2\gamma - 1)(\Delta u)_{,i}(\Delta u)_{,i} + (2\gamma cf'(u) - 1)u_{,i}u_{,i},\end{aligned} \tag{2.5}$$

from which it is clear that V is subharmonic in D for $\gamma = \max\{\frac{1}{2}, \frac{1}{2c\beta}\}$.

We summarize this result in

THEOREM 1. *Let $u \in C^4(D)$ be a solution of* (2.1), *where $q(x) \geqq 0$ and $c > 0$. If $f \in C^1(R)$ satisfies $f'(u) \geqq \beta > 0$ and g satisfies* $\mathrm{sg}(s) \geqq 0$, *then for $\gamma \geqq \max\{\frac{1}{2}, \frac{1}{2c\beta}\}$ the function*

$$V(x) = u_{,i}u_{,i} + \gamma(\Delta u)^2 + 2\gamma c\int_0^{u(x)} f(t)\, dt$$

takes its maximum value on the boundary of D.

If instead of (2.1), we consider

$$\Delta^2 u - q(x)g(\Delta u) + p(x)u = 0, \tag{2.6}$$

where $q(x) \geqq 0$ and $p(x) \geqq p_0 > 0$, and define

$$W(x) = u_{,i}u_{,i} + \gamma(\Delta u)^2 + \gamma p(x)u^2, \tag{2.7}$$

then in a manner similar to the derivation of (2.5), one obtains

$$\begin{aligned}\Delta W = {}& 2u_{,ij}u_{,ij} + |(\Delta u)_{,i} + u_{,i}|^2 + 2\gamma q(x)(\Delta u)g(\Delta u) + (2\gamma - 1)(\Delta u)_{,i}(\Delta u)_{,i} \\ & + (\gamma p - 1)u_{,i}u_{,i} + \gamma p|u_{,i} + 2p^{-1}up_{,i}|^2 + \gamma u^2[\Delta p - 4p^{-1}|\nabla p|^2].\end{aligned}$$

Consequently, we deduce the following.

THEOREM 2. *Let $u \in C^4(D)$ be a solution of* (2.6), *where $q(x) \geqq 0$ and $p \in C^2(D)$ satisfies $p(x) \geqq p_0 > 0$ and $p\Delta p - 4|\nabla p|^2 \geqq 0$. If g satisfies the condition* $\operatorname{sg}(s) \geqq 0$, *then for $\gamma \geqq \max\{\frac{1}{2}, \frac{1}{p_0}\}$ the function*

$$W(x) = u_{,i}u_{,i} + \gamma(\Delta u)^2 + \gamma p(x)u^2$$

takes its maximum value on the boundary of D.

If $q(x) = 0$ in (2.1) and (2.6), then we can introduce alternative auxiliary functions defined on the solutions of the respective equations.

THEOREM 3. *If $u \in C^4(D)$ is a solution of*

$$\Delta^2 u + cf(u) = 0, \qquad c > 0, \tag{2.8}$$

where $f \in C^1$ satisfies $f'(u) \geqq \beta > 0$ and $f(0) = 0$, then for

$$\gamma \geqq \max\left\{1, \frac{1}{(n-1)^2 c\beta}\right\}$$

the function

$$T(x) = nu_{,i}u_{,i} - u\Delta u + \gamma(n-1)^2\left[\frac{1}{2}(\Delta u)^2 + c\int_0^u f(t)\,dt\right]$$

takes its maximum value on the boundary of D.

Proof. Using a straightforward calculation of the Laplacian, (2.8), and a completion of the square, one obtains the identity

$$\begin{aligned}\Delta T = {}& 2nu_{,ij}u_{,ij} - (\Delta u)^2 + cuf(u) + |(n-1)(\Delta u)_{,i} + u_{,i}|^2 \\ & + (n-1)^2(\gamma - 1)(\Delta u)_{,i}(\Delta u)_{,i} + [(n-1)^2\gamma cf' - 1]u_{,i}u_{,i}.\end{aligned}$$

Since $nu_{,ij}u_{,ij} \geqq (\Delta u)^2$, we conclude that T is subharmonic in D under the conditions cited in the theorem.

In our next special case we omit the parameter γ but encounter a fixed lower bound on $p(x)$.

THEOREM 4. *If $u \in C^4(D)$ is a solution of*

$$\Delta^2 u + p(x)u = 0, \tag{2.9}$$

where $p(x) \geqq \frac{1}{3}$ and $\Delta(p^{-1}) \leqq 0$, then

$$S(x) = \frac{1}{2}(\Delta u - 2u)^2 + \frac{(2n-1)^2}{2}(\Delta u)^2 + 4nu_{,i}u_{,i} + (2n^2 - 2n + 1)pu^2$$

takes its maximum value on the boundary of D.

Proof. Computing the Laplacian and using (2.9), we have

$$\begin{aligned}\Delta S = {}& 2pu^2 - 2(\Delta u)^2 + 4u\Delta u + (4n^2 - 4n + 2)(\Delta u)_{,i}(\Delta u)_{,i} + 4(2n-1)u_{,i}(\Delta u)_{,i} \\ & + 4u_{,i}u_{,i} + 8nu_{,ij}u_{,ij} + \tau u^2\Delta p + 2\tau pu_{,i}u_{,i} + 4\tau uu_{,i}p_{,i},\end{aligned} \tag{2.10}$$

where $\tau=\tau(n)=2n^2-2n+1$. We use the arithmetic-geometric mean inequality on the fifth term of (2.10) to deduce that

$$(4n^2-4n+2)(\Delta u)_{,i}(\Delta u)_{,i}+4(2n-1)u_{,i}(\Delta u)_{,i}+4u_{,i}u_{,i} \geqq (\Delta u)_{,i}(\Delta u)_{,i}$$

and is thus nonnegative. Moreover, using the same inequality on the last term in (2.10), we have

$$\tau u^2\Delta p+2\tau p u_{,i}u_{,i}+4\tau u u_{,i}p_{,i} \geqq \tau u^2\left[\Delta p-\frac{2|\nabla p|^2}{p}\right]$$

and hence the sum is nonnegative when the bracket is nonnegative. Consequently, from (2.10) we have

$$\Delta S \geqq 2pu^2-2(\Delta u)^2+4u\Delta u+8nu_{,ij}u_{,ij}.$$

Now since $nu_{,ij}u_{,ij} \geqq (\Delta u)^2$ and

$$4u\Delta u \geqq -\tfrac{2}{3}u^2-6(\Delta u)^2,$$

we conclude that $\Delta S \geqq 2(p-\frac{1}{3})u^2$ and S is subharmonic in D.

In our final principle we require p to be harmonic.

THEOREM 5. *Let $u \in C^4(D)$ be a solution of* (1.1), *where $q(x) \geqq 0$ and $p(x)$ is a positive harmonic function bounded below by p_0. If $f \in C^1$ is a bounded function satisfying $f'(u) \geqq \beta > 0$ and g satisfies the requirement* $\mathrm{sg}\,(s) \geqq 0$, *then for $\gamma \geqq \max\{\frac{1}{2p_0}, \frac{1}{\beta}\}$ and $\alpha=\max\,(2\gamma^2 f^2/p_0)$ the function*

$$R(x)=\frac{u_{,i}u_{,i}}{p(x)}+\gamma\frac{(\Delta u)^2}{p(x)}+2\gamma\int_0^u f(t)\,dt+\alpha p(x)$$

takes its maximum value on the boundary of D.

Proof. We calculate

$$\begin{aligned}
R_{,j} &= 2p^{-1}u_{,i}u_{,ij}-p^{-2}u_{,i}u_{,i}p_{,j}+2p^{-1}\gamma(\Delta u)(\Delta u)_{,j}\\
&\qquad -\gamma p^{-2}(\Delta u)^2p_{,j}+2\gamma f u_{,j}+\alpha p_{,j},\\
\Delta R &= 2p^{-1}u_{,ij}u_{,ij}+2p^{-1}u_{,i}(\Delta u)_{,i}-4p^{-2}u_{,i}u_{,ij}p_{,j}\\
&\qquad +2p^{-3}|\nabla u|^2|\nabla p|^2+2\gamma p^{-1}q(\Delta u)g(\Delta u)+2\gamma p^{-1}(\Delta u)_{,j}(\Delta u)_{,j}\\
&\qquad -4\gamma p^{-2}(\Delta u)(\Delta u)_{,j}p_{,j}+2\gamma p^{-3}(\Delta u)^2|\nabla p|^2+2\gamma f'u_{,j}u_{,j},
\end{aligned}$$

and then form

$$\begin{aligned}
\Delta R+2p^{-1}p_{,j}R_{,j} &= 2p^{-1}u_{,ij}u_{,ij}+2p^{-1}u_{,i}(\Delta u)_{,i}+2\gamma p^{-1}q(\Delta u)g(\Delta u)\\
&\qquad +2\gamma p^{-1}(\Delta u)_{,j}(\Delta u)_{,j}+4\gamma p^{-1}fp_{,j}u_{,j}+2\gamma f'u_{,j}u_{,j}+2\alpha p^{-1}|\nabla p|^2.
\end{aligned} \tag{2.11}$$

The need for $p(x)$ to be harmonic arises in ΔR where Δp appears with both a positive and negative coefficient. Now if we use the arithmetic-geometric mean inequality on the second and fifth terms on the right side of (2.11), we can write

$$\begin{aligned}
\Delta R+2p^{-1}p_{,j}R_{,j} &\geqq 2p^{-1}u_{,ij}u_{,ij}+(2\gamma p-1)p^{-2}(\Delta u)_{,i}(\Delta u)_{,i}\\
&\qquad +2(\gamma f'-1)u_{,i}u_{,i}+[2\alpha p-(2\gamma f)^2]p^{-2}|\nabla p|^2,
\end{aligned}$$

from which the result follows.

3. Bounds. In Theorem 1 the subharmonic function $V(x)$ attains its maximum at some point, say x_0, on the boundary of D. Thus it follows that one can obtain bounds for the gradient of the solution, the Laplacian of the solution, and (perhaps implicitly) for the solution of (2.1). For example, as a consequence of Theorem 1, we have

$$|\nabla u(x)|^2 \leqq |\nabla u(x_0)|^2 + \gamma[\Delta u(x_0)]^2 + 2\gamma c \int_{u(x)}^{u(x_0)} f(t)\,dt$$

for x in D. If in Theorem 1 we also have $f(0)=0$, such as when $f(u)=u^3+u$, then we can obtain an estimate for the gradient that does not depend on the value of the solution at the point in question, i.e.,

$$|\nabla u(x)|^2 \leqq |\nabla u(x_0)|^2 + \gamma[\Delta u(x_0)]^2 + 2\gamma c \int_{0}^{u(x_0)} f(t)\,dt.$$

If, in addition to u being a solution of (2.1), one imposes homogeneous boundary conditions on u, then it may be possible to obtain maximum principles on the gradient or the Laplacian of u. For example, if u satisfies (2.1) in D and $u=0=\partial u/\partial n$ on ∂D, then it follows from Theorem 1 when $f(0)=0$ that

$$|\Delta u(x)| \leqq |\Delta u(x_0)|, \qquad x \in \bar{D}$$

for some $x_0 \in \partial D$. However, one must be concerned with the existence question in a problem of this type as is evident from [3] and [12]. In [3] it was shown that if u is a solution of (2.1) in D, where $f'(u) \geqq 0$ and $f(0)=0$, and is subject to boundary conditions of the form $u=0=\Delta u$, then $u \equiv 0$.

We have only briefly indicated how Theorem 1 in § 2 can be utilized to obtain pointwise bounds. Many other applications of the principles in § 2 are possible.

Acknowledgment. This work was accomplished while the author was a Visiting Guest Professor in the Mathematisches Institut I at the Universität Karlsruhe, West Germany. This opportunity was made possible by a University of Tennessee Faculty Development Leave.

REFERENCES

[1] S. N. CHOW AND D. R. DUNNINGER, *A maximum principle for n-metaharmonic functions*, Proc. Amer. Math. Soc., 43 (1974), pp. 79-83.

[2] D. R. DUNNINGER, *Maximum principles for solutions of some fourth order elliptic equations*, J. Math. Anal. Appl., 37 (1972), pp. 655-658.

[3] V. B. GOYAL AND P. W. SCHAEFER, *On a subharmonic functional in fourth order nonlinear elliptic problems*, J. Math. Anal. Appl., 83 (1981), pp. 20-25.

[4] L. E. PAYNE, *Bounds for the maximum stress in the Saint Venannt torsion problem*, Indian J. Mech. Math., Special Issue (1968), pp. 51-59.

[5] L. E. PAYNE AND I. STAKGOLD, *On the mean value of the fundamental mode in the fixed membrane problem*, Appl. Anal., 3 (1973), pp. 295-303.

[6] L. E. PAYNE, *Some remarks on maximum principles*, J. Analyse Math., 30 (1976), pp. 421-433.

[7] L. E. PAYNE AND G. A. PHILIPPIN, *Some applications of the maximum principle in the problem of torsional creep*, SIAM J. Appl. Math., 33 (1977), pp. 446-455.

[8] M. H. PROTTER, *Gradient bounds for a class of second order elliptic equations*, Contemp. Math., 11 (1982), pp. 191-198.

[9] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[10] P. W. SCHAEFER AND R. P. SPERB, *Maximum principles for some functionals associated with the solution of elliptic boundary value problems*, Arch Rational Mech. Anal., 61 (1976), pp. 65-76.

[11] ———, *Maximum principles and bounds in some inhomogeneous elliptic boundary value problems*, this Journal, 8 (1977), pp. 871-878.

[12] P. W. SCHAEFER, *On a maximum principle for a class of fourth order semilinear elliptic equations*, Proc. Royal Soc. Edinburgh, 77A (1977), pp. 319-323.

[13] R. P. SPERB, *Maximum Principles and their Applications*, Academic Press, New York, 1981.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

013

HOMOGENIZATION LIMITS OF THE EQUATIONS OF ELASTICITY IN THIN DOMAINS*

A. DAMLAMIAN† AND M. VOGELIUS‡

Abstract. We study pure bending of a flat, linearly elastic three-dimensional plate with rapidly varying composition. Under appropriate coercivity—and boundedness—assumptions, it is shown that all limiting vertical displacement, coming from the equations of three-dimensional elasticity with plate thickness approaching 0, must necessarily satisfy a fourth order, two-dimensional "plate" equation. This analysis does *not* require any special structure of the composition, such as periodicity or quasiperiodicity.

Key words. elasticity, homogenization, plates

AMS(MOS) subject classifications. 73K10, 73K20, 35B30

Introduction. We study pure bending of a flat, linearly elastic three-dimensional plate with rapidly varying composition. A uniform coercivity-condition and a uniform boundedness-condition are placed on the constitutive elastic law, but we require no special structure in composition, such as periodicity or quasiperiodicity. It is shown that all limiting vertical displacements, coming from the equations of 3-D elasticity with plate-thickness approaching 0, must necessarily satisfy a fourth order equation of the form

$$\partial_{\alpha\beta}(M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta}w)=\mathcal{F} \tag{1}$$

on the (plate-) midplane Ω. The existence of a limiting equation of the form (1) is well known for plates with slowly varying composition [3], [14], [16]; recently Caillerie has studied plates with rapidly varying periodic composition [2]. The analysis presented here extends those results to plates with arbitrary variation in composition. For the periodic case there is an (essentially) unique limiting rigidity tensor $M_{\alpha\beta\gamma\delta}$. In addition to the particular form of the local variation it only depends on the limiting ratio of the thickness and the length scale of variation. The tensor $M_{\alpha\beta\gamma\delta}$ may be expressed in terms of energies of certain periodic cell problems.

Without structure assumptions about the local composition it is no longer possible to give explicit formulas for the $M_{\alpha\beta\gamma\delta}$. We believe that in interesting cases it may be possible to find upper and lower bounds for its eigenvalues much in the same way as has been (at least partially) done for certain composites [5], [12], [13], [19].

The problem studied here is also closely related to the problem of plates with rapidly varying periodic thickness considered in [7]-[10]. If voids are thought of as occupied by a material of zero strength, then a plate with rapidly varying thickness may in principle be thought of as a flat plate with rapidly varying composition. Due to our coercivity-condition, voids are not permitted in the flat plates considered here; they represent an added difficulty which we, at this point, technically do not know

* Received by the editors August 7, 1985; accepted for publication November 21, 1985. Work on this project was initiated during visits by the authors to the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota 55455, which is supported in part by a grant from the National Science Foundation.

† Centre de Mathématiques, École Polytechnique, F-91128, Palaiseau, Cedex, France, visiting the Department of Mathematics, University of Maryland, College Park, Maryland 20742.

‡ Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was partially supported by Office of Naval Research contract N00014-85-K-0169.

how to handle without extra assumptions about the local structure of the composition. Some of the ideas of Γ-convergence may be relevant to this problem, since they have successfully been applied to study limits of noncoercive functionals in other circumstances [1].

Plates with rapidly varying composition are of interest in structural optimization; in certain design contexts they are known to be stronger than plates with only slow variation in composition. We refer the reader to [10] for a more detailed discussion of the relation between an optimal design problem and plates with rapidly varying composition.

The approach taken here is a variation of the method of H-convergence introduced by Murat [15] and Tartar [17]. One major difference is that in this case both the dimension of the domain in which the limiting equation is satisfied as well as the order of the limiting equation differs from that associated to the equations with rapid variations. In the analysis this difference is probably most apparent in the construction of the isomorphism from $H^{-2}(\Omega)$ to $\overset{\circ}{H}{}^2(\Omega)$, which is the candidate for the resolvent of our limiting operator.

The organization of this paper is as follows. In the first section we briefly provide some preliminaries concerning the equations of 3-D elasticity, and in addition we give a precise statement of the convergence result to be proven later. It is very convenient to rescale the thickness variable of the plate to the interval $(-1, 1)$; the three-dimensional equations are now all in the same domain—but they become singularly perturbed as the old thickness parameter approaches zero. In § 2 we apply Korn's and Poincare's inequalities to the solutions of the rescaled equations, and this leads directly to estimates of various expressions and then to statements about weakly convergent subsequences and the structure of their limits. A major part of any convergence argument, using the method of H-convergence, is to construct the isomorphism which is the candidate for the resolvent of the limit operator. In this case such an operator must necessarily map $H^{-2}(\Omega)$ onto $\overset{\circ}{H}{}^2(\Omega)$ and it turns out that it may be constructed from the three-dimensional equations by restricting attention to external loads that are uniform throughout the thickness. Section 4 contains the final step of the convergence argument, namely the verification of the right constitutive relationship between curvature and bending moments. This is accomplished by integration by parts of the energy bilinear form. The trial-functions are the solutions to the three-dimensional elastic problem with the prescribed loads, and the test functions are picked so that they satisfy the three-dimensional elastic equations with an external load which is uniform throughout the thickness, and so that they furthermore have constant curvatures in the limit as the thickness approaches zero. It is possible to select such test functions because of the aforementioned isomorphism. This last part of the proof is a variation of Murat and Tartar's div-curl lemma, which again is a special case of the so-called method of compensated compactness [18].

1. Preliminaries and statement of the main result. We shall write $\underset{\sim}{x} = (x_1, x_2, x_3)$ for vectors in $\mathbb{R}^3$ and $\underset{\approx}{x} = (x_1, x_2)$ for vectors in $\mathbb{R}^2$. Latin indices will usually range from 1 to 3, and Greek ones from 1 to 2; the summation convention applies whenever indices are repeated. We write $\partial_i = \partial/\partial x_i$ and $\partial_{ij} = \partial^2/\partial x_i\, \partial x_j$. The three-dimensional flat plate of thickness 2ε is given by

$$R(\varepsilon) = \{\underset{\sim}{x} \colon \underset{\approx}{x} \in \Omega, |x_3| < \varepsilon\}$$

where Ω is a smooth bounded domain in $\mathbb{R}^2$ and ε denotes a small parameter, with say $0 < \varepsilon \leqq 1$. We shall denote by $\partial_+ R(\varepsilon)$ and $\partial_- R(\varepsilon)$ the upper and lower faces of the plate

$$\partial_\pm R(\varepsilon)=\{\underset{\sim}{x}\colon \underset{\sim}{x}\in\Omega,\ x_3=\pm\varepsilon\}$$

and by $\partial_0 R(\varepsilon)$ we denote the outer edge of the plate

$$\partial_0 R(\varepsilon)=\{\underset{\sim}{x}\colon \underset{\sim}{x}\in\partial\Omega,\ |x_3|<\varepsilon\}.$$

Associated with any displacement $\underline{u}=(u_1, u_2, u_3)$ of $R(\varepsilon)$ is its strain tensor

$$e_{ij}(\underline{u})=\tfrac{1}{2}(\partial_i u_j+\partial_j u_i)$$

and the corresponding stress tensor

$$\sigma^\varepsilon_{ij}(\underline{u})=b^\varepsilon_{ijkl}e_{kl}(\underline{u}).$$

We are concerned with spatially inhomogeneous materials and so the components b^ε_{ijkl} of the elastic tensor are bounded measurable functions in $\underset{\sim}{x}$. These functions are permitted to change with the thickness parameter ε, and (except for certain symmetries) they will only be restricted by the following two requirements:

$$b^\varepsilon_{ijkl}(\underset{\sim}{x})e_{ij}e_{kl}\geqq c_1\sum_{i,j}|e_{ij}|^2, \tag{2}$$

$$\left(\sum_{i,j}|b^\varepsilon_{ijkl}(\underset{\sim}{x})e_{kl}|^2\right)^{1/2}\leqq C_2\left(\sum_{i,j}|e_{ij}|^2\right)^{1/2}, \tag{3}$$

a.e. in $R(\varepsilon)$ for any symmetric 2 tensor e, with constants $c_1>0$ and C_2 that are independent of ε. Note: throughout this paper c_1 and C_2 always refer to these same constants whereas the letters c and C will denote generic positive constants (independent of ε).

Remark 1. One simple example of rapid variation in composition, which falls within the framework of our study and which has received quite a bit of attention, is the case

$$b^\varepsilon_{ijkl}(\underset{\sim}{x})=b^1_{ijkl}(\underset{\sim}{x}/\varepsilon),$$

where $b^1_{ijkl}(\underset{\sim}{y})$ is periodic in $\underset{\sim}{y}$ and satisfies (2), (3) (cf. [2]). We also mention the work in [7]-[10] about plates with rapidly varying periodic thickness

$$\tilde{R}(\varepsilon)=\{\underset{\sim}{x}\colon \underset{\sim}{x}\in\Omega,\ |x_3|<\varepsilon h(\underset{\sim}{x}/\varepsilon^a)\},\qquad 0<a<\infty.$$

If we denote $h_{\max}=\max h$ then formally, at least, such a plate corresponds to an inhomogeneous material

$$b^\varepsilon_{ijkl}(\underset{\sim}{x})=b^1_{ijkl}(\underset{\sim}{x}/\varepsilon^a, x_3/\varepsilon)$$

with

$$b^1_{ijkl}(\underline{y})=\begin{cases} c_{ijkl}, & |y_3|<h(\underset{\sim}{y}),\\ 0, & |y_3|\geqq h(\underset{\sim}{y}),\end{cases}$$

occupying the flat domain

$$\{\underset{\sim}{x}\colon \underset{\sim}{x}\in\Omega,\ |x_3|<\varepsilon h_{\max}\}.$$

The possibility that b^ε_{ijkl} may vanish on $R(\varepsilon)$ is technically a significant extra difficulty which we shall not include in our analysis of arbitrary, rapidly varying composition. □

We always assume that the elastic tensor obeys the symmetries

$$b^\varepsilon_{ijkl}=b^\varepsilon_{jikl}=b^\varepsilon_{ijlk}=b^\varepsilon_{klij} \tag{4}$$

and that the horizontal planes are planes of elastic symmetry, which means (cf. [11])

$$(5)\qquad b^\varepsilon_{\alpha\beta\gamma 3}=0,\qquad b^\varepsilon_{\alpha 333}=0.$$

Finally we assume that b^ε is even with respect to x_3:

$$(6)\qquad b^\varepsilon_{ijkl}(\underset{\sim}{x}, x_3)=b^\varepsilon_{ijkl}(\underset{\sim}{x}, -x_3).$$

The equations of elastostatic equilibrium for the clamped, vertically loaded, three-dimensional plate $R(\varepsilon)$ are

$$(7)\qquad \begin{aligned} &-\partial_j[\sigma^\varepsilon_{ij}(\underline{u}^\varepsilon)]=\begin{cases}0, & i=1,2\\ \varepsilon^2F^\varepsilon, & i=3\end{cases}\quad \text{in } R(\varepsilon),\\ &\sigma^\varepsilon_{ij}(\underline{u}^\varepsilon)\nu_j=\begin{cases}0, & i=1,2\\ \varepsilon^3 f^\varepsilon_\pm, & i=3\end{cases}\quad \text{on } \partial_\pm R(\varepsilon),\\ &\underline{u}^\varepsilon=0\quad \text{on } \partial_0R(\varepsilon),\end{aligned}$$

where $\underline{\nu}=(0,0,\pm1)$ denotes the outward normal to $\partial_\pm R(\varepsilon)$. The loads are scaled in order to insure that $\underline{u}^\varepsilon$ stays bounded as ε goes to zero. For convenience we shall assume that F^ε is even in x_3 and that $f^\varepsilon_+(\underset{\sim}{x})=f^\varepsilon_-(\underset{\sim}{x})$; the common boundary load is denoted $f^\varepsilon(\underset{\sim}{x})$. Because of the linearity of the problem, this represents no loss of generality in the limit $\varepsilon\to0$, as the energy corresponding to odd loading is negligible compared with that of even loading for an elastic law with the symmetries (5) and (6) (cf. [8]). Notice that from the assumptions about the loads and (5), (6) it follows that

$$u^\varepsilon_1, u^\varepsilon_2 \text{ are odd},\quad u^\varepsilon_3 \text{ is even},$$

$$\sigma^\varepsilon_{\alpha\beta}, \sigma^\varepsilon_{33} \text{ are odd},\quad \sigma^\varepsilon_{\alpha 3} \text{ is even},$$

with respect to x_3; X_ε will denote the space of all admissible displacements that obey these symmetries:

$$X_\varepsilon=\{\underline{u}\in H^1(R(\varepsilon)) \colon \underline{u}|_{\partial_0R(\varepsilon)}=0,\ u_1,\ u_2 \text{ are odd and } u_3 \text{ is even in } x_3\},$$

where $H^1(R(\varepsilon))$ is the space of (vector-valued) functions with square integrable first derivatives. The problem (7) now has the following variational formulation:

$$(8)\qquad \underline{u}^\varepsilon\in X_\varepsilon,\qquad \int_{R(\varepsilon)}\sigma^\varepsilon_{ij}(\underline{u}^\varepsilon)e_{ij}(\underline{v})\,d\underline{x}=\varepsilon^2\int_{R(\varepsilon)}F^\varepsilon v_3\,d\underline{x}+2\varepsilon^3\int_{\partial_+R(\varepsilon)}f^\varepsilon v_3\,d\underset{\sim}{x}$$

for any $\underline{v}\in X_\varepsilon$. Unless explicitly stated otherwise we shall always assume that

$$(9)\qquad F^\varepsilon\in L^2(R(\varepsilon)),\qquad f^\varepsilon\in L^2(\Omega)$$

with

$$(10)\qquad \begin{aligned}&F^\varepsilon(\underset{\sim}{x}, \varepsilon y)\to F^0(\underset{\sim}{x}, y)\quad \text{in } L^2(\Omega\times(-1,1)),\\ &f^\varepsilon(\underset{\sim}{x})\to f^0(\underset{\sim}{x})\quad \text{in } L^2(\Omega),\end{aligned}$$

as $\varepsilon\to0$ (the regularity assumptions on the loads can be somewhat relaxed, see Remark 4, but we do not feel that this serves any purpose in the present context).

Since we are not imposing any requirements on the structure of the rapid variation in the b^ε_{ijkl}, we will not in general obtain convergence of $\underline{u}^\varepsilon$ as ε approaches zero. Instead our main result concerns convergent subsequences (which corresponds to the compactness property in the theory of Γ-convergence, cf. [4]).

THEOREM. *Let $\{\varepsilon_k\}_{k=1}^{\infty}$ be any given sequence converging to zero. There exist a subsequence $\{\varepsilon_{k(l)}\}_{l=1}^{\infty}$, for simplicity denoted $\{\varepsilon_l\}_{l=1}^{\infty}$, and a tensor-valued function $M_{\alpha\beta\gamma\delta}(\underset{\sim}{x})$ such that*

(i) $M_{\alpha\beta\gamma\delta}=M_{\beta\alpha\gamma\delta}=M_{\alpha\beta\delta\gamma}=M_{\gamma\delta\alpha\beta}$,

(ii) *$M_{\alpha\beta\gamma\delta}$ lies in $L^{\infty}(\Omega)$ with*

$$M_{\alpha\beta\gamma\delta}(\underset{\sim}{x})t_{\alpha\beta}t_{\gamma\delta}\geqq\frac{2}{3}c_1\sum_{\alpha,\beta}|t_{\alpha\beta}|^2,$$

$$\left(\sum_{\alpha,\beta}|M_{\alpha\beta\gamma\delta}(\underset{\sim}{x})t_{\gamma\delta}|^2\right)^{1/2}\leqq\frac{2}{3}C_2\left(\sum_{\alpha,\beta}|t_{\alpha\beta}|^2\right)^{1/2},$$

a.e. in Ω, for any symmetric 2 tensor t;

(iii) *for any even external load F^{ε_l} and any boundary loads $f_+^{\varepsilon_l}=f_-^{\varepsilon_l}=f^{\varepsilon_l}$, satisfying* (9) *and* (10), *the solution to the problem* (7), *$\underset{\sim}{u}^{\varepsilon_l}$, as ε_l approaches zero, converges to*

$$(-x_3\partial_1 w(\underset{\sim}{x}),\ -x_3\partial_2 w(\underset{\sim}{x}),\ w(\underset{\sim}{x})),$$

where $w\in H^2(\Omega)$ solves the problem

$$\partial_{\alpha\beta}(M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta}w)=\mathscr{F}^0\quad in\ \Omega,$$

$$w=\frac{\partial w}{\partial n}=0\quad on\ \partial\Omega.$$

The effective load $\mathscr{F}^0(\underset{\sim}{x})$ is given by

$$\mathscr{F}^0(\underset{\sim}{x})=\int_{-1}^{1}F^0(\underset{\sim}{x},y)\,dy+2f^0(\underset{\sim}{x}),$$

where F^0 and f^0 are the limits from (10). *The convergence is in the weak topologies*

$$u_3^{\varepsilon_l}(\underset{\sim}{x},\varepsilon_l y)\rightharpoonup w(\underset{\sim}{x})\quad in\ H^1(\Omega\times(-1,1)),$$

$$\frac{1}{\varepsilon_l}u_\alpha^{\varepsilon_l}(\underset{\sim}{x},\varepsilon_l y)\rightharpoonup -y\partial_\alpha w(\underset{\sim}{x})\quad in\ H^1(\Omega\times(-1,1)).$$

Remark 2. It is possible to prove a similar theorem without the symmetry requirement that $b_{ijkl}^{\varepsilon}=b_{klij}^{\varepsilon}$ (i.e., without assuming that b_{ijkl}^{ε} is a symmetric operator on 2 tensors). Of course, the resulting tensor $M_{\alpha\beta\gamma\delta}$ will not possess this symmetry either. The coercivity estimate stays the same, but the operator norm estimate for $M_{\alpha\beta\gamma\delta}$, $\frac{2}{3}C_2$, is replaced by $\frac{2}{3}((C_2)^2/c_1)$. We consider the symmetric case, since it is the only physically interesting one in the context of elastostatics.

Remark 3. It follows immediately from the statement of our theorem that

$$\frac{1}{2\varepsilon_l}\int_{-\varepsilon_l}^{\varepsilon_l}u_3^{\varepsilon_l}\,dx_3\rightharpoonup w(\underset{\sim}{x})\quad\text{in }\mathring{H}^1(\Omega);$$

it is actually shown in the proof of the theorem that $1/(2\varepsilon_l)\int_{-\varepsilon_l}^{\varepsilon_l}u_3^{\varepsilon_l}\,dx_3$ converges strongly towards w in $\mathring{H}^1(\Omega)$.

Remark 4. As stated earlier, the assumptions (10) on the loads are not the weakest possible. For the solution of (7) (or rather of (8)) to make sense it is necessary and sufficient that the functional

$$v\to\varepsilon^2\int_{R(\varepsilon)}F^{\varepsilon}v\,d\underset{\sim}{x}+2\varepsilon^3\int_{\partial_+R(\varepsilon)}f^{\varepsilon}v\,d\underset{\sim}{x}$$

be in the dual of $\{v \in H^1(R(\varepsilon)) \colon v|_{\partial_0 R(\varepsilon)} = 0,\ v$ is even in $x_3\}$, with the integrals representing appropriate duality pairings.

Define a rescaled functional $\mathscr{F}^\varepsilon$ as follows:

$$\langle \mathscr{F}^\varepsilon, v\rangle = \varepsilon^{-1} \int_{R(\varepsilon)} F^\varepsilon v(\underset{\sim}{x}, x_3/\varepsilon)\, d\underset{\sim}{x} + 2 \int_{\partial_+ R(\varepsilon)} f^\varepsilon v(\underset{\sim}{x}, x_3/\varepsilon)\, d\underset{\sim}{x}.$$

$\mathscr{F}^\varepsilon$ is then an element of the dual of $\{v \in H^1(R(1)) \colon v|_{\partial_0 R(1)} = 0,\ v$ is even in $x_3\}$ and our theorem still holds provided $\mathscr{F}^\varepsilon$ converge strongly in this dual space. The effective load $\mathscr{F}^0$ is given by the functional

$$\langle \mathscr{F}^0, w\rangle = \langle \lim_{\varepsilon \to 0} \mathscr{F}^\varepsilon, w\rangle,$$

where the function $w(\underset{\sim}{x})$ in the second expression is interpreted as a function of all three variables $(x_1, x_2, x_3) = (\underset{\sim}{x}, x_3)$ (only independent of x_3). We note that it is not possible to obtain all elements of the dual of $\overset{\circ}{H}{}^2(\Omega)$ as effective loads by this construction. The solution to the limit problem

$$\partial_{\alpha\beta}(M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta} w) = \mathscr{F}^0 \quad \text{in } \Omega,$$
$$w = \frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

may be defined variationally for any $\mathscr{F}^0$ in the dual of $\overset{\circ}{H}{}^2(\Omega)$, but for certain (very unsmooth) $\mathscr{F}^0$, w is not related to solutions of the 3-D equations of elasticity through the limiting process discussed in this paper.

2. The rescaled problem—a priori estimates and limit behaviour. In this paragraph we study a rescaling of the problem (7) to the fixed domain $R(1) = \Omega \times (-1, 1)$. Independent variables in $R(1)$ will be denoted $(\underset{\sim}{x}, y)$ and we define the new dependent variables as follows:

$$U^\varepsilon_\alpha(\underset{\sim}{x}, y) = \frac{1}{\varepsilon} u^\varepsilon_\alpha(\underset{\sim}{x}, \varepsilon y), \qquad U^\varepsilon_3(\underset{\sim}{x}, y) = u^\varepsilon_3(\underset{\sim}{x}, \varepsilon y),$$

$$E^\varepsilon_{ij}(\underset{\sim}{x}, y) = \frac{1}{\varepsilon} e_{ij}(\underline{u}^\varepsilon)(\underset{\sim}{x}, \varepsilon y), \qquad \Sigma^\varepsilon_{ij}(\underset{\sim}{x}, y) = \frac{1}{\varepsilon} \sigma^\varepsilon_{ij}(\underline{u}^\varepsilon)(\underset{\sim}{x}, \varepsilon y).$$

The strain tensor of $\underline{U}^\varepsilon$ respective to the variables $(\underset{\sim}{x}, y)$ is given by

$$\left(\begin{array}{c|c} E^\varepsilon_{\alpha\beta} & \varepsilon E^\varepsilon_{\alpha 3} \\ \hline \varepsilon E^\varepsilon_{3\beta} & \varepsilon^2 E^\varepsilon_{33} \end{array}\right), \tag{11}$$

and from an application of Korn's inequality in the domain $R(1)$ (cf. [6]), it now follows that

$$\|\underline{U}^\varepsilon\|_{H^1(R(1))} \leqq C\left(\sum_{\alpha,\beta} \|E^\varepsilon_{\alpha\beta}\|^2_{L^2(R(1))} + \varepsilon^2 \sum_{\alpha} \|E^\varepsilon_{\alpha 3}\|^2_{L^2(R(1))} + \varepsilon^4 \|E^\varepsilon_{33}\|^2_{L^2(R(1))}\right)^{1/2} \tag{12}$$

(remember $\underline{U}^\varepsilon$ vanishes on $\partial_0 R(1)$). By rescaling the system (7) we see that

$$\begin{gathered} -\partial_\beta \Sigma^\varepsilon_{\alpha\beta} + \frac{1}{\varepsilon} \frac{\partial}{\partial y} \Sigma^\varepsilon_{\alpha 3} = 0, \\ -\frac{1}{\varepsilon} \partial_\beta \Sigma^\varepsilon_{3\beta} - \frac{1}{\varepsilon^2} \frac{\partial}{\partial y} \Sigma^\varepsilon_{33} = F^\varepsilon(\underset{\sim}{x}, \varepsilon y) \quad \text{in } R(1), \\ \Sigma^\varepsilon_{\alpha 3} = 0, \quad \frac{1}{\varepsilon^2} \Sigma^\varepsilon_{33} = \pm f^\varepsilon(\underset{\sim}{x}) \quad \text{on } \partial_\pm R(1), \end{gathered} \tag{13}$$

or in a variational form

$$\begin{aligned}(14)\qquad &\int_{R(1)}\left[\Sigma^\varepsilon_{\alpha\beta}e_{\alpha\beta}(\underline{V})+\frac{2}{\varepsilon}\Sigma^\varepsilon_{\alpha3}e_{\alpha3}(\underline{V})+\frac{1}{\varepsilon^2}\Sigma^\varepsilon_{33}e_{33}(\underline{V})\right]d\underset{\sim}{x}\,dy\\ &=\int_{R(1)}F^\varepsilon(\underset{\sim}{x},\varepsilon y)V_3\,d\underset{\sim}{x}\,dy+2\int_{\partial_+R(1)}f^\varepsilon V_3\,d\underset{\sim}{x}\end{aligned}$$

for any $\underline{V}\in X_1=\{\underline{V}\in H^1(R(1))\colon \underline{V}|_{\partial_0R(1)}=0,\ V_1,\ V_2$ are odd and V_3 is even in $y\}$. Here we have used the notation $e(\underline{V})$ for the strain of the displacement field $\underline{V}(\underset{\sim}{x},y)$ relative to the variables $(\underset{\sim}{x},y)$.

LEMMA 1. *The norms* $\|\underline{U}^\varepsilon\|_{H^1(R(1))}$, $\|E^\varepsilon_{ij}\|_{L^2(R(1))}$, *and* $\|\Sigma^\varepsilon_{ij}\|_{L^2(R(1))}$ *are all bounded by* $C(\|F^\varepsilon(\underset{\sim}{x},\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)})$.

Proof. Inserting $\underline{V}=\underline{U}^\varepsilon$ into (14) and using the formula (11) for $e(\underline{U}^\varepsilon)$ we get

$$(15)\qquad \int_{R(1)}\Sigma^\varepsilon_{ij}E^\varepsilon_{ij}\,d\underset{\sim}{x}\,dy=\int_{R(1)}F^\varepsilon(\underset{\sim}{x},\varepsilon y)U^\varepsilon_3\,d\underset{\sim}{x}\,dy+2\int_{\partial_+R(1)}f^\varepsilon U^\varepsilon_3\,d\underset{\sim}{x}.$$

The coercivity of the elastic tensor and the estimate (12) now leads to

$$\begin{aligned}c_1\sum_{i,j}\|E^\varepsilon_{ij}\|^2_{L^2(R(1))}&\leqq C(\|F^\varepsilon(\underset{\sim}{x},\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)})\|\underline{U}^\varepsilon\|_{H^1(R(1))}\\ &\leqq C(\|F^\varepsilon(x,\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)})\left(\sum_{i,j}\|E^\varepsilon_{ij}\|^2_{L^2(R(1))}\right)^{1/2}\end{aligned}$$

so that

$$(16)\qquad \left(\sum_{i,j}\|E^\varepsilon_{ij}\|^2\right)^{1/2}\leqq C(\|F^\varepsilon(\underset{\sim}{x},\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)}).$$

The desired estimates follow directly from (16) in combination with (3) and (12). □

Remark 5. Based on Lemma 1 and the formulas for $E^\varepsilon_{\alpha3}$ and E^ε_{33} we get that

$$(17)\qquad \left\|\frac{\partial}{\partial y}U^\varepsilon_\alpha+\partial_\alpha U^\varepsilon_3\right\|_{L^2(R(1))}\leqq C\varepsilon(\|F^\varepsilon(\underset{\sim}{x},\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)}),$$

$$(18)\qquad \left\|\frac{\partial}{\partial y}U^\varepsilon_3\right\|_{L^2(R(1))}\leqq C\varepsilon^2(\|F^\varepsilon(\underset{\sim}{x},\varepsilon y)\|_{L^2(R(1))}+\|f^\varepsilon\|_{L^2(\Omega)}).$$

According to our assumption (10) $F^\varepsilon(\underset{\sim}{x},\varepsilon y)$ converges in $L^2(R(1))$ and f^ε in $L^2(\Omega)$; Lemma 1 now gives that $\|\underline{U}^\varepsilon\|_{H^1(R(1))}$, $\|E^\varepsilon_{ij}\|_{L^2(R(1))}$ and $\|\Sigma^\varepsilon_{ij}\|_{L^2(R(1))}$ are bounded independently of ε. From any sequence $\{\varepsilon_k\}^\infty_{k=1}$ converging to zero it is thus possible to extract a subsequence $\{\varepsilon_l\}^\infty_{l=1}$ such that

$$(19)\qquad \underline{U}^{\varepsilon_l}\rightharpoonup\underline{U}^0\quad\text{in } H^1(R(1)),$$

$$(20)\qquad E^{\varepsilon_l}_{ij}\rightharpoonup E^0_{ij}\quad\text{in } L^2(R(1)),$$

$$(21)\qquad \Sigma^{\varepsilon_l}_{ij}\rightharpoonup\Sigma^0_{ij}\quad\text{in } L^2(R(1)),$$

as ε_l approaches zero.

LEMMA 2. *The third component of* $\underline{U}^0$, U^0_3, *is independent of* y *and belongs to* $\overset{\circ}{H}{}^2(\Omega)$. *Furthermore*

$$U^0_\alpha=-y\partial_\alpha U^0_3,\qquad E^0_{\alpha\beta}=-y\partial_{\alpha\beta}U^0_3.$$

Proof. From (18) we get that $(\partial/\partial y)U^0_3=0$, U^0_3 is therefore independent of y. From (17) we get $(\partial/\partial y)U^0_\alpha=-\partial_\alpha U^0_3$ and since U^0_α is odd with respect to y, $U^0_\alpha=-y\partial_\alpha U^0_3$.

$E^{\varepsilon_l}_{\alpha\beta}=e_{\alpha\beta}(\underset{\sim}{U}^{\varepsilon_l})=\frac{1}{2}(\partial_\alpha U^{\varepsilon_l}_\beta+\partial_\beta U^{\varepsilon_l}_\alpha)$ converges as a distribution towards $\frac{1}{2}(\partial_\alpha U^0_\beta+\partial_\beta U^0_\alpha)=-y\partial_{\alpha\beta}U^0_3$; on the other hand, it also converges towards $E^0_{\alpha\beta}$, and consequently $E^0_{\alpha\beta}=-y\partial_{\alpha\beta}U^0_3$. We already know that $U^0_3\in H^1(\Omega)$ with $U^0_3=0$ on $\partial\Omega$. From the fact that $-y\,\partial_\alpha U^0_3=U^0_\alpha\in H^1(R(1))$ we conclude that $U^0_3\in H^2(\Omega)\cap\mathring{H}^1(\Omega)$. It only remains to prove that $(\partial/\partial n)U^0_3=0$ on $\partial\Omega$, where $\partial/\partial n$ is the outward normal derivative.

Since $U^{\varepsilon_l}_\alpha=0$ on $\partial_0R(1)=\partial\Omega\times(-1,1)$ and $U^{\varepsilon_l}_\alpha$ converges weakly towards U^0_α in $H^1(R(1))$, it follows that $-y\partial_\alpha U^0_3=U^0_\alpha=0$ on $\partial\Omega\times(-1,1)$. This necessarily implies that $\partial_\alpha U^0_3=0$ on $\partial\Omega$, and so $(\partial/\partial n)U^0_3=n_\alpha\partial_\alpha U^0_3=0$ on $\partial\Omega$. □

If $v(\underset{\sim}{x},y)$ is a function on $R(1)$ then we define

$$\bar{v}(\underset{\sim}{x})=\frac{1}{2}\int_{-1}^{1}v(\underset{\sim}{x},y)\,dy.$$

From (19) and the fact that U^0_3 is independent of y it follows that $\bar{U}^{\varepsilon_l}_3\rightharpoonup U^0_3$ in $H^1(\Omega)$; this result may be slightly improved.

LEMMA 3. *$\bar{U}^{\varepsilon_l}_3$ converges strongly towards U^0_3 in $\mathring{H}^1(\Omega)$ as ε_l approaches zero.*

Proof. We may write

$$\begin{aligned}\partial_\alpha\bar{U}^\varepsilon_3(\underset{\sim}{x})&=\frac{1}{2}\int_{-1}^{1}\partial_\alpha U^\varepsilon_3(\underset{\sim}{x},y)\,dy\\&=\frac{1}{2}\int_{-1}^{1}\left(\varepsilon E^\varepsilon_{\alpha3}(\underset{\sim}{x},y)-\frac{\partial}{\partial y}U^\varepsilon_\alpha(\underset{\sim}{x},y)\right)dy\\&=\varepsilon\bar{E}^\varepsilon_{\alpha3}(\underset{\sim}{x})-\frac{1}{2}(U^\varepsilon_\alpha(\underset{\sim}{x},1)-U^\varepsilon_\alpha(\underset{\sim}{x},-1)).\end{aligned}$$

From the estimate of $\|E^\varepsilon_{\alpha3}\|_{L^2(R(1))}$ in Lemma 1 it therefore follows that

$$\partial_\alpha\bar{U}^\varepsilon_3(\underset{\sim}{x})+\tfrac{1}{2}(U^\varepsilon_\alpha(\underset{\sim}{x},1)-U^\varepsilon_\alpha(\underset{\sim}{x},-1))\to0\quad\text{in } L^2(\Omega)\quad\text{as }\varepsilon\to0. \tag{22}$$

We also know that $U^{\varepsilon_l}_\alpha\rightharpoonup U^0_\alpha$ in $H^1(R(1))$, and this implies that $U^{\varepsilon_l}_\alpha(\underset{\sim}{x},\pm1)$ converges weakly in $H^{1/2}(\Omega)$, and thus strongly in $L^2(\Omega)$, towards $U^0_\alpha(\underset{\sim}{x},\pm1)=\mp\partial_\alpha U^0_3(\underset{\sim}{x})$. It follows immediately from (22) that

$$\partial_\alpha(\bar{U}^{\varepsilon_l}_3-U^0_3)\to0\quad\text{in } L^2(\Omega),$$

which shows that $\bar{U}^{\varepsilon_l}_3$ converges strongly towards U^0_3 in $\mathring{H}^1(\Omega)$, as ε_l approaches zero. □

Integration in y on both sides of (21) gives $\int_{-1}^{1}\Sigma^{\varepsilon_l}_{\alpha3}\,dy\rightharpoonup\int_{-1}^{1}\Sigma^0_{\alpha3}\,dy$ in $L^2(\Omega)$. It turns out that $\int_{-1}^{1}\Sigma^0_{\alpha3}\,dy=0$, and furthermore that it is possible to find the weak limit of $(1/\varepsilon_l)\int_{-1}^{1}\Sigma^{\varepsilon_l}_{\alpha3}\,dy$ in $H^{-1}(\Omega)$ (our notation for the dual of $\mathring{H}^1(\Omega)$).

LEMMA 4. *$(1/\varepsilon_l)\int_{-1}^{1}\Sigma^{\varepsilon_l}_{\alpha3}\,dy$ converges weakly in $H^{-1}(\Omega)$ towards $\partial_\beta\int_{-1}^{1}y\Sigma^0_{\alpha\beta}\,dy$ as ε_l approaches zero.*

Proof. Performing an integration by parts and using the first and third equation in (13) we obtain

$$\frac{1}{\varepsilon}\int_{-1}^{1}\Sigma^\varepsilon_{\alpha3}\,dy=-\frac{1}{\varepsilon}\int_{-1}^{1}y\frac{\partial}{\partial y}\Sigma^\varepsilon_{\alpha3}\,dy=\int_{-1}^{1}y\partial_\beta\Sigma^\varepsilon_{\alpha\beta}\,dy=\partial_\beta\int_{-1}^{1}y\Sigma^\varepsilon_{\alpha\beta}\,dy. \tag{23}$$

Since $\Sigma^{\varepsilon_l}_{\alpha\beta}\rightharpoonup\Sigma^0_{\alpha\beta}$ in $L^2(R(1))$ we know that

$$\int_{-1}^{1}y\Sigma^{\varepsilon_l}_{\alpha\beta}\,dy\rightharpoonup\int_{-1}^{1}y\Sigma^0_{\alpha\beta}\,dy\quad\text{in } L^2(\Omega)$$

and therefore

$$\frac{1}{\varepsilon_l}\int_{-1}^{1}\Sigma^{\varepsilon_l}_{\alpha3}\,dy\rightharpoonup\partial_\beta\int_{-1}^{1}y\Sigma^0_{\alpha\beta}\,dy\quad\text{in } H^{-1}(\Omega).$$

□

The tensor $-\int_{-1}^{1} y\Sigma_{\alpha\beta}^0\, dy$ will play a significant role in the construction of the tensor $M_{\alpha\beta\gamma\delta}$; it eventually turns out that

$$-\int_{-1}^{1} y\Sigma_{\alpha\beta}^0\, dy = M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta}U_3^0.$$

At this point we only observe that

$$-\partial_{\alpha\beta}\int_{-1}^{1} y\Sigma_{\alpha\beta}^0\, dy = \int_{-1}^{1} F^0(\underset{\sim}{x}, y)\, dy + 2f^0(\underset{\sim}{x}) \quad \text{in } \Omega, \tag{24}$$

where F^0 and f^0 are the limits of $F^\varepsilon(\underset{\sim}{x}, \varepsilon y)$ and f^ε, respectively (cf. (10)). To get (24) one uses (23) and the second and fourth identities in (13) to write

$$\begin{aligned} -\partial_{\alpha\beta}\int_{-1}^{1} y\Sigma_{\alpha\beta}^\varepsilon\, dy &= -\frac{1}{\varepsilon}\partial_\alpha \int_{-1}^{1} \Sigma_{\alpha 3}^\varepsilon\, dy \\ &= \frac{1}{\varepsilon^2}\int_{-1}^{1} \frac{\partial}{\partial y}\Sigma_{33}^\varepsilon\, dy + \int_{-1}^{1} F^\varepsilon(\underset{\sim}{x}, \varepsilon y)\, dy \\ &= 2f^\varepsilon(\underset{\sim}{x}) + \int_{-1}^{1} F^\varepsilon(\underset{\sim}{x}, \varepsilon y)\, dy. \end{aligned} \tag{25}$$

Passing to the limits in this identity as ε approaches zero along the sequence $\{\varepsilon_l\}_{l=1}^\infty$ we are led to (24).

Remark 6. So far we have obtained a number of convergence results for subsequences $\underline{U}^{\varepsilon_l}$, $E_{ij}^{\varepsilon_l}$ and $\Sigma_{ij}^{\varepsilon_l}$ corresponding to a specific set of loads $(F^\varepsilon, f^\varepsilon)$ converging to (F^0, f^0). Since the appropriate norms of the differences between the U^ε's, the E^ε's and the Σ^ε's corresponding to different loads $(F^\varepsilon, f^\varepsilon)$ and $(G^\varepsilon, g^\varepsilon)$ are bounded by $C(\|(F^\varepsilon - G^\varepsilon)(\underset{\sim}{x}, \varepsilon y)\|_{L^2(R(1))} + \|f^\varepsilon - g^\varepsilon\|_{L^2(\Omega)})$ (Lemma 1), it follows that we may pick the same index sequence $\{\varepsilon_l\}_{l=1}^\infty$ for *any* loads $(F^\varepsilon, f^\varepsilon)$ that converge to this (F^0, f^0) in the sense of (10).

$L^2(R(1)) \times L^2(\Omega)$ is a separable Hilbert space; let $\{(F_N^0, f_N^0)\}_{N=1}^\infty$ be a basis. Following the previous argument we may for each N find a subsequence

$$\{\varepsilon_l^N\}_{l=1}^\infty \subseteq \{\varepsilon_l^{N-1}\}_{l=1}^\infty \subseteq \cdots \subseteq \{\varepsilon_l^1\}_{l=1}^\infty \subseteq \{\varepsilon_k\}_{k=1}^\infty$$

so that the convergence results listed above hold for solutions to the problem (7) for any $(F^\varepsilon, f^\varepsilon)$ converging to (F_N^0, f_N^0). By taking the diagonal subsequence of all the $\{\varepsilon_l^N\}_{l=1}^\infty$ we obtain a subsequence $\{\varepsilon_l\}_{l=1}^\infty$ for which the convergence results hold simultaneously for all N. Since linear combinations of the $\{(F_N^0, g_N^0)\}_{N=1}^\infty$ are dense in $L^2(R(1)) \times L^2(\Omega)$, and since appropriate norms of the differences between the U^{ε_l}'s, the E^{ε_l}'s and the Σ^{ε_l}'s, in the limit, are bounded by the norm of the difference between the limits of the loads in $L^2(R(1)) \times L^2(\Omega)$ (Lemma 1), it follows that the *convergence results of this section hold for the fixed subsequence* $\{\varepsilon_l\}_{l=1}^\infty$ *for any loads* $(F^\varepsilon, f^\varepsilon)$ *that converge in the sense of* (10).

3. An auxiliary isomorphism. If the homogenized limit operator is to have the form $\partial_{\alpha\beta}(M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta})$ then it must necessarily be an isomorphism between $\mathring{H}^2(\Omega)$ and $H^{-2}(\Omega)$ (our notation for the dual of $\mathring{H}^2(\Omega)$). We shall now study in more detail a particular case of the boundary value problem (7), where the exterior load is independent of x_3 and ε and where the boundary loads vanish. We show that, in the limit as ε_l approaches 0, this naturally leads to an isomorphism between $\mathring{H}^2(\Omega)$ and $H^{-2}(\Omega)$. We owe the initial suggestion, that it might be easier to obtain an isomorphism using vanishing boundary loads, to L. Tartar. For the remainder of this paper $\{\varepsilon_l\}_{l=1}^\infty$ always refers to the "universal" subsequence of $\{\varepsilon_k\}_{k=1}^{\infty}$ selected by the diagonalization process discussed in Remark 6.

Let G be in $L^2(\Omega)$ and let $\underline{v}^\varepsilon \in X_\varepsilon$ be the solution to

$$
\begin{aligned}
&-\partial_j[\sigma_{ij}^\varepsilon(\underline{v}^\varepsilon)] = \begin{cases} 0, & i=1,2 \\ \varepsilon^2 G(\underset{\sim}{x}), & i=3 \end{cases} \quad \text{in } R(\varepsilon), \\
&\sigma_{ij}^\varepsilon(\underline{v}^\varepsilon)\nu_j = 0, \qquad i=1,2,3 \quad \text{on } \partial_\pm R(\varepsilon), \\
&\underline{v}^\varepsilon = 0 \quad \text{on } \partial_0 R(\varepsilon).
\end{aligned} \tag{26}
$$

As before we introduce rescaled variables

$$
V_\alpha^\varepsilon(\underset{\sim}{x}, y) = \frac{1}{\varepsilon} v_\alpha^\varepsilon(\underset{\sim}{x}, \varepsilon y), \qquad V_3^\varepsilon(\underset{\sim}{x}, y) = v_3^\varepsilon(\underset{\sim}{x}, \varepsilon y),
$$

$$
\tilde{E}_{ij}^\varepsilon = \frac{1}{\varepsilon} e_{ij}(\underline{v}^\varepsilon)(\underset{\sim}{x}, \varepsilon y), \qquad \tilde{\Sigma}_{ij}^\varepsilon = \frac{1}{\varepsilon} \sigma_{ij}^\varepsilon(\underline{v}^\varepsilon)(\underset{\sim}{x}, \varepsilon y).
$$

From the analysis in the previous section we know among other things that

$$
\bar{V}_3^{\varepsilon_l} \to V_3^0 \quad \text{in } H^1(\Omega)
$$

as ε_l approaches zero. We furthermore know that $V_3^0 \in \mathring{H}^2(\Omega)$.

LEMMA 5. *For any* $G \in L^2(\Omega)$,

$$
\|V_3^0\|_{\mathring{H}^2(\Omega)}^2 \leqq C \int_\Omega G V_3^0 \, d\underset{\sim}{x}, \qquad \|G\|_{H^{-2}(\Omega)}^2 \leqq C \int_\Omega G V_3^0 \, d\underset{\sim}{x}.
$$

Proof. The identity corresponding to (15) in this case ($F^\varepsilon(\underset{\sim}{x}, \varepsilon y) = G(\underset{\sim}{x})$, $f^\varepsilon = 0$) reads

$$
\int_{R(1)} \tilde{\Sigma}_{ij}^\varepsilon \tilde{E}_{ij}^\varepsilon \, d\underset{\sim}{x}\, dy = \int_{R(1)} G V_3^\varepsilon \, d\underset{\sim}{x}\, dy;
$$

because of the coercivity assumption (2) and the fact that G is independent of y it follows that

$$
c_1 \sum_{\alpha,\beta} \|\tilde{E}_{\alpha\beta}^\varepsilon\|_{L^2(R(1))}^2 \leqq c_1 \sum_{i,j} \|\tilde{E}_{ij}^\varepsilon\|_{L^2(R(1))}^2 \leqq 2 \int_\Omega G \bar{V}_3^\varepsilon \, d\underset{\sim}{x}. \tag{27}
$$

From Lemma 2 we know

$$
\tilde{E}_{\alpha\beta}^{\varepsilon_l} \rightharpoonup -y \partial_{\alpha\beta} V_3^0 \quad \text{in } L^2(R(1)).
$$

Passing to the limit in (27) along the sequence $\{\varepsilon_l\}$, and using the weak lower semicontinuity of the norm, we thus obtain

$$
\frac{2}{3} c_1 \sum_{\alpha,\beta} \|\partial_{\alpha\beta} V_3^0\|_{L^2(\Omega)}^2 = c_1 \sum_{\alpha,\beta} \|y \partial_{\alpha\beta} V_3^0\|_{L^2(R(1))}^2 \leqq 2 \int_\Omega G V_3^0 \, d\underset{\sim}{x}.
$$

This proves the first inequality of our statement, since

$$
\left(\sum_{\alpha,\beta} \|\partial_{\alpha\beta} \cdot \|_{L^2(\Omega)}^2 \right)^{1/2}
$$

is one of the equivalent norms on $\mathring{H}^2(\Omega)$.

The identity corresponding to (25) in the present situation reads

$$\partial_{\alpha\beta}\int_{-1}^{1} y\tilde{\Sigma}^{\varepsilon}_{\alpha\beta}\,dy=-2G(\underset{\sim}{x}).$$

Consequently,

$$\|G\|^2_{H^{-2}(\Omega)}\leqq C\sum_{\alpha,\beta}\left\|\int_{-1}^{1} y\tilde{\Sigma}^{\varepsilon}_{\alpha\beta}\,dy\right\|^2_{L^2(\Omega)}\leqq C\sum_{\alpha,\beta}\|\tilde{\Sigma}^{\varepsilon}_{\alpha\beta}\|^2_{L^2(R(1))}. \tag{28}$$

The $\tilde{E}^{\varepsilon}_{ij}$ and $\tilde{\Sigma}^{\varepsilon}_{ij}$ are related by

$$\tilde{\Sigma}^{\varepsilon}_{ij}=b^{\varepsilon}_{ijkl}(\underset{\sim}{x},\varepsilon y)\tilde{E}^{\varepsilon}_{kl},$$

and this in combination with (3) and (28) leads to

$$\|G\|^2_{H^{-2}(\Omega)}\leqq C\sum_{i,j}\|\tilde{E}^{\varepsilon}_{ij}\|^2_{L^2(R(1))}.$$

The last inequality of (27) then gives

$$\|G\|^2_{H^{-2}(\Omega)}\leqq C\int_{\Omega} G\bar{V}^{\varepsilon}_3\,d\underset{\sim}{x},$$

which in the limit as ε approaches zero along the sequence $\{\varepsilon_l\}_{l=1}^{\infty}$ yields the desired second inequality. □

We define an operator from $L^2(\Omega)$ into $\mathring{H}^2(\Omega)$ by $G\to V^0_3$. It follows directly from the two inequalities in Lemma 5 that

$$c\|G\|_{H^{-2}(\Omega)}\leqq\|V^0_3\|_{\mathring{H}^2(\Omega)}\leqq C\|G\|_{H^{-2}(\Omega)},$$

i.e., the above operator may be extended as an injective and bounded linear operator $\mathcal{S}$: $H^{-2}(\Omega)\to\mathring{H}^2(\Omega)$. From the second inequality in Lemma 5 it now follows, using the Lax-Milgram lemma, that $\mathcal{S}$ maps $H^{-2}(\Omega)$ onto $\mathring{H}^2(\Omega)$. In summary

(29) The operator $G\to V^0_3$ may be extended as an isomorphism $\mathcal{S}$ between $H^{-2}(\Omega)$ and $\mathring{H}^2(\Omega)$.

Let G_i, $i=1,2$, be two elements of $L^2(\Omega)$, and let $\underset{\sim}{v}^{\varepsilon}_{(i)}$ denote the solution of (26), corresponding to $G=G_i$, $i=1,2$. According to the variational formulation (8) and the symmetry of the elastic tensor b^{ε}_{ijkl}

$$\begin{aligned}\varepsilon^2\int_{R(\varepsilon)} G_1v^{\varepsilon}_{(2),3}\,d\underset{\sim}{x}&=\int_{R(\varepsilon)}\sigma^{\varepsilon}_{ij}(\underset{\sim}{v}^{\varepsilon}_{(1)})e_{ij}(\underset{\sim}{v}^{\varepsilon}_{(2)})\,d\underset{\sim}{x}\\&=\int_{R(\varepsilon)}\sigma^{\varepsilon}_{ij}(\underset{\sim}{v}^{\varepsilon}_{(2)})e_{ij}(\underset{\sim}{v}^{\varepsilon}_{(1)})\,d\underset{\sim}{x}=\varepsilon^2\int_{R(\varepsilon)} G_2v^{\varepsilon}_{(1),3}\,d\underset{\sim}{x}.\end{aligned}$$

From this we immediately conclude that

$$\int_{\Omega} G_1\bar{V}^{\varepsilon}_{(2),3}\,d\underset{\sim}{x}=\int_{\Omega} G_2\bar{V}^{\varepsilon}_{(1),3}\,d\underset{\sim}{x},$$

and thus in the limit, as ε approaches zero along the sequence $\{\varepsilon_l\}_{l=1}^{\infty}$, we obtain

$$\int_{\Omega} G_1\mathcal{S}(G_2)\,d\underset{\sim}{x}=\int_{\Omega} G_2\mathcal{S}(G_1)\,d\underset{\sim}{x}. \tag{30}$$

By continuity the identity (30) is satisfied for any $G_i\in H^{-2}(\Omega)$, i.e., we have shown that $\mathcal{S}$ is selfadjoint.

Finally let us consider the operator that takes $G \in L^2(\Omega)$ to the tensor $-\int_{-1}^{1} y\tilde{\Sigma}^0_{\alpha\beta}\, dy \in L^2(\Omega)$. The $\tilde{E}^\varepsilon_{ij}$ and $\tilde{\Sigma}^\varepsilon_{ij}$ are related by

$$\tilde{\Sigma}^\varepsilon_{ij} = b^\varepsilon_{ijkl}(x, \varepsilon y)\tilde{E}^\varepsilon_{kl},$$

and using (3) and the last inequality in (27), we thus get

$$\begin{aligned} \sum_{\alpha,\beta} \left\| \int_{-1}^{1} y\tilde{\Sigma}^\varepsilon_{\alpha\beta}\, dy \right\|^2_{L^2(\Omega)} &\leqq C \sum_{\alpha,\beta} \|\tilde{\Sigma}^\varepsilon_{\alpha\beta}\|^2_{L^2(R(1))} \\ &\leqq C \sum_{i,j} \|\tilde{E}^\varepsilon_{ij}\|^2_{L^2(R(1))} \leqq C \int_\Omega G\bar{V}^\varepsilon_3\, dx. \end{aligned} \tag{31}$$

Because of the weak lower semicontinuity of the norm it follows, by passing to the limit in (31) along the sequence $\{\varepsilon_l\}_{l=1}^\infty$, that

$$\sum_{\alpha,\beta} \left\| \int_{-1}^{1} y\tilde{\Sigma}^0_{\alpha\beta}\, dy \right\|^2_{L^2(\Omega)} \leqq C \int_\Omega GV^0_3\, dx. \tag{32}$$

We just proved that $\|V^0_3\|_{\mathring{H}^2(\Omega)} \leqq C\|G\|_{H^{-2}(\Omega)}$ and from (32) we therefore get

$$\sum_{\alpha,\beta} \left\| \int_{-1}^{1} y\tilde{\Sigma}^0_{\alpha\beta}\, dy \right\|^2_{L^2(\Omega)} \leqq C\|G\|^2_{H^{-2}(\Omega)},$$

or

(33) The operator $G \to -\int_{-1}^{1} y\tilde{\Sigma}^0_{\alpha\beta}\, dy$ may be extended as a bounded linear operator from $H^{-2}(\Omega)$ into $L^2(\Omega)$.

We shall refer to this extension as $\mathcal{T}_{\alpha\beta}$.

4. The proof of the main result. We already proved that the rescaled displacements

$$U^{\varepsilon_l} = \left(\frac{1}{\varepsilon} u_1^{\varepsilon_l}(x, \varepsilon_l y), \frac{1}{\varepsilon} u_2^{\varepsilon_l}(x, \varepsilon_l y), u_3^{\varepsilon_l}(x, \varepsilon_l y)\right)$$

converge weakly towards

$$(-y\partial_1 U^0_3(x), -y\partial_2 U^0_3(x), U^0_3(x))$$

in $H^1(R(1))$, with $U^0_3 \in \mathring{H}^2(\Omega)$.

In this section we verify that there exists a tensor $M_{\alpha\beta\gamma\delta}$ (independent of the loads $F^\varepsilon, f^\varepsilon$), with the properties (i) and (ii) listed in our theorem, for which

$$-\int_{-1}^{1} y\Sigma^0_{\alpha\beta}\, dy = M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta}U^0_3. \tag{34}$$

From (24) we know that

$$-\partial_{\alpha\beta} \int_{-1}^{1} y\Sigma^0_{\alpha\beta}\, dy = \int_{-1}^{1} F^0(x, y)\, dy + 2f^0(x),$$

and by combining with (34) we therefore get that U^0_3 satisfies

$$\partial_{\alpha\beta}(M_{\alpha\beta\gamma\delta}\partial_{\gamma\delta}U^0_3) = \int_{-1}^{1} F^0(x, y)\, dy + 2f^0(x) \quad \text{in } \Omega,$$

with

$$U^0_3 = \frac{\partial}{\partial n} U^0_3 = 0 \quad \text{in } \partial\Omega.$$

Except for a change of notation (replace U_3^0 by the simpler w) this will complete the proof of our theorem.

Our verification of the existence of the tensor $M_{\alpha\beta\gamma\delta}$ proceeds by the method of compensated compactness (cf. [18]); specifically we adapt the so-called div-curl lemma of Murat and Tartar to the present problem. $\underline{u}^\varepsilon$ as previously denotes the solution of (7) (or (8)) with loads $F^\varepsilon \in L^2(R(\varepsilon))$ and $f_+^\varepsilon = f_-^\varepsilon = f^\varepsilon \in L^2(\Omega)$, and $\underline{v}^\varepsilon$ denotes the solution of (26) with $G \in L^2(\Omega)$. $\underline{U}^\varepsilon$, E_{ij}^ε, Σ_{ij}^ε and $\underline{V}^\varepsilon$, $\tilde{E}_{ij}^\varepsilon$, $\tilde{\Sigma}_{ij}^\varepsilon$ denote the rescaled variables corresponding to $\underline{u}^\varepsilon$ and $\underline{v}^\varepsilon$, respectively. Letting ϕ be an arbitrary but fixed function in $\mathscr{D}(\Omega)$, we shall then compute the limit of

$$\int_{R(1)} \Sigma_{ij}^\varepsilon \tilde{E}_{ij}^\varepsilon \phi \, d\underset{\sim}{x} \, dy = \int_{R(1)} \tilde{\Sigma}_{ij}^\varepsilon E_{ij}^\varepsilon \phi \, d\underset{\sim}{x} \, dy$$

in two different ways as ε goes to zero along the sequence $\{\varepsilon_l\}_{l=1}^\infty$.

Inserting the test field $\phi \underline{V}^\varepsilon$ into (14), and using the fact that $e(\underline{V}^\varepsilon)$ has the form (11), with E^ε replaced by $\tilde{E}^\varepsilon$, we get

$$\begin{aligned}\int_{R(1)} \Sigma_{ij}^\varepsilon \tilde{E}_{ij}^\varepsilon \phi \, d\underset{\sim}{x} \, dy &= \int_{R(1)} F^\varepsilon(\underset{\sim}{x}, \varepsilon y) V_3^\varepsilon \phi \, d\underset{\sim}{x} \, dy + 2 \int_{\partial_+ R(1)} f^\varepsilon V_3^\varepsilon \phi \, d\underset{\sim}{x} \\ &\quad - \int_{R(1)} \Sigma_{\gamma\delta}^\varepsilon V_\gamma^\varepsilon \partial_\delta \phi \, d\underset{\sim}{x} \, dy - \frac{1}{\varepsilon} \int_{R(1)} \Sigma_{\gamma 3}^\varepsilon V_3^\varepsilon \partial_\gamma \phi \, d\underset{\sim}{x} \, dy \end{aligned} \tag{35}$$

(other terms vanish because ϕ is independent of y). From (19), compactness and Lemma 2 we get that

$$\begin{aligned} &V_3^{\varepsilon_l} \phi \to V_3^0 \phi \quad \text{in } L^2(R(1)), \\ &V_3^{\varepsilon_l}(\underset{\sim}{x}, 1) \phi \to V_3^0 \phi \quad \text{in } L^2(\Omega), \\ &V_\gamma^{\varepsilon_l} \partial_\delta \phi \to -y \partial_\gamma V_3^0 \partial_\delta \phi \quad \text{in } L^2(R(1)); \end{aligned}$$

at the same time F^{ε_l}, f^{ε_l} and $\Sigma_{\alpha\beta}^{\varepsilon_l}$ are all weakly convergent in L^2 (indeed the first two converge strongly). We therefore conclude that as ε approaches 0 along the sequence $\{\varepsilon_l\}_{l=1}^\infty$, the first three terms on the right-hand side of (35) approach

$$\int_\Omega \left(\int_{-1}^1 F^0(\underset{\sim}{x}, y) \, dy + 2 f^0(\underset{\sim}{x}) \right) V_3^0 \phi \, d\underset{\sim}{x} + \int_\Omega \int_{-1}^1 y \Sigma_{\gamma\delta}^0 \, dy \, \partial_\gamma V_3^0 \partial_\delta \phi \, d\underset{\sim}{x}. \tag{36}$$

The last term $-1/\varepsilon \int_{R(1)} \Sigma_{\gamma 3}^\varepsilon V_3^\varepsilon \partial_\gamma \phi \, d\underset{\sim}{x} \, dy$ requires special attention. Using Poincaré's inequality on vertical lines and the fact that $(\partial/\partial y) V_3^\varepsilon = \varepsilon^2 \tilde{E}_{33}^\varepsilon$ (cf. (11)) we get

$$\begin{aligned} \left| \frac{1}{\varepsilon} \int_{R(1)} \Sigma_{\gamma 3}^\varepsilon (V_3^\varepsilon - \bar{V}_3^\varepsilon) \partial_\gamma \phi \, d\underset{\sim}{x} \, dy \right| &\leqq C \frac{1}{\varepsilon} \| \Sigma_{\gamma 3}^\varepsilon \|_{L^2(R(1))} \| V_3^\varepsilon - \bar{V}_3^\varepsilon \|_{L^2(R(1))} \\ &\leqq C \frac{1}{\varepsilon} \| \Sigma_{\gamma 3}^\varepsilon \|_{L^2(R(1))} \left\| \frac{\partial}{\partial y} V_3^\varepsilon \right\|_{L^2(R(1))} \\ &= C \varepsilon \| \Sigma_{\gamma 3}^\varepsilon \|_{L^2(R(1))} \| \tilde{E}_{33}^\varepsilon \|_{L^2(R(1))}, \end{aligned}$$

and since both $\|\Sigma_{\gamma 3}^\varepsilon\|_{L^2(R(1))}$ and $\|\tilde{E}_{33}^\varepsilon\|_{L^2(R(1))}$ are bounded this last term is of order ε. It thus suffices to study the limiting behavior of

$$-\frac{1}{\varepsilon} \int_{R(1)} \Sigma_{\gamma 3}^\varepsilon \bar{V}_3^\varepsilon \partial_\gamma \phi \, d\underset{\sim}{x} \, dy = \int_\Omega \left(-\frac{1}{\varepsilon} \int_{-1}^1 \Sigma_{\gamma 3}^\varepsilon \, dy \right) \bar{V}_3^\varepsilon \partial_\gamma \phi \, d\underset{\sim}{x}.$$

According to Lemma 3, $\bar{V}_3^{\varepsilon_l}$ converges strongly in $\mathring{H}^1(\Omega)$ and according to Lemma 4 $(1/\varepsilon_l)\int_{-1}^1 \Sigma_{\gamma 3}^{\varepsilon_l}\,dy$ converges weakly in $H^{-1}(\Omega)$ towards $\partial_\delta \int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy$, hence

$$\int_\Omega \left(-\frac{1}{\varepsilon_l}\int_{-1}^1 \Sigma_{\gamma 3}^{\varepsilon_l}\,dy\right)\bar{V}_3^{\varepsilon_l}\partial_\gamma\phi\,d\underset{\sim}{x}$$

converges to

$$-\int_\Omega \partial_\delta \Bigg(\Bigg(\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\Bigg) V_3^0\partial_\gamma\phi\,d\underset{\sim}{x} \tag{37}$$

as ε_l approaches 0. Collecting the terms in (36) and (37) we get

$$\begin{aligned}\lim_{\varepsilon_l\to 0}\int_{R(1)} \Sigma_{ij}^{\varepsilon_l}\tilde{E}_{ij}^{\varepsilon_l}\phi\,d\underset{\sim}{x}\,dy = \int_\Omega \mathscr{F}^0 V_3^0\phi\,d\underset{\sim}{x} + \int_\Omega \Bigg(\Bigg(\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\Bigg)\partial_\gamma V_3^0\partial_\delta\phi\,d\underset{\sim}{x}\\ -\int_\Omega \partial_\delta\Bigg(\Bigg(\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\Bigg) V_3^0\partial_\gamma\phi\,d\underset{\sim}{x}.\end{aligned}$$

We integrate the last two terms in the right-hand side by parts to remove derivatives from ϕ. Using the fact that ϕ vanishes on $\partial\Omega$ and the identity (24) we thus obtain

$$\lim_{\varepsilon_l\to 0}\int_{R(1)} \Sigma_{ij}^{\varepsilon_l}\tilde{E}_{ij}^{\varepsilon_l}\phi\,d\underset{\sim}{x}\,dy = -\int_\Omega \Bigg(\Bigg(\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\Bigg)\partial_{\gamma\delta}V_3^0\phi\,d\underset{\sim}{x}. \tag{38}$$

Exchanging the roles of $\underset{\sim}{U}^\varepsilon$ and $\underset{\sim}{V}^\varepsilon$ in the above argument we would similarly obtain

$$\lim_{\varepsilon_l\to 0}\int_{R(1)} \tilde{\Sigma}_{ij}^{\varepsilon_l}E_{ij}^{\varepsilon_l}\phi\,d\underset{\sim}{x}\,dy = -\int_\Omega \Bigg(\Bigg(\int_{-1}^1 y\tilde{\Sigma}^0_{\gamma\delta}\,dy\Bigg)\partial_{\gamma\delta}U_3^0\phi\,d\underset{\sim}{x}. \tag{39}$$

Because of the symmetry of the elastic law

$$\int_{R(1)} \Sigma_{ij}^{\varepsilon}\tilde{E}_{ij}^{\varepsilon}\phi\,d\underset{\sim}{x}\,dy = \int_{R(1)} \tilde{\Sigma}_{ij}^{\varepsilon}E_{ij}^{\varepsilon}\phi\,d\underset{\sim}{x}\,dy,$$

and it then follows from (38) and (39) that

$$\int_\Omega \left(-\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\right)\partial_{\gamma\delta}V_3^0\phi\,d\underset{\sim}{x} = \int_\Omega \left(-\int_{-1}^1 y\tilde{\Sigma}^0_{\gamma\delta}\,dy\right)\partial_{\gamma\delta}U_3^0\phi\,d\underset{\sim}{x}.$$

In terms of the operators $\mathscr{S}$ and $\mathscr{T}$ defined in the previous section, this may be stated

$$\int_\Omega \left(-\int_{-1}^1 y\Sigma^0_{\gamma\delta}\,dy\right)\partial_{\gamma\delta}\mathscr{S}(G)\phi\,d\underset{\sim}{x} = \int_\Omega \mathscr{T}_{\gamma\delta}(G)\partial_{\gamma\delta}U_3^0\phi\,d\underset{\sim}{x}. \tag{40}$$

The identity (40) has so far only been verified for G in $L^2(\Omega)$ (and any $\phi\in\mathscr{D}(\Omega)$), but because of the continuity of the operators $\mathscr{S}$ and $\mathscr{T}_{\gamma\delta}$ (cf. (29), (33)) it follows immediately that (40) holds for any G in $H^{-2}(\Omega)$. Let $\Omega'\subset\subset\Omega$ and pick $\psi\in\mathscr{D}(\Omega)$ with $\psi\equiv 1$ in Ω'. $\mathscr{S}$ is an isomorphism between $H^{-2}(\Omega)$ and $\mathring{H}^2(\Omega)$ (cf. (29)); insertion of $G=\mathscr{S}^{-1}(\frac{1}{2}x_\alpha x_\beta\psi)$ into (40) yields

$$\int_\Omega \left(-\int_{-1}^1 y\Sigma^0_{\alpha\beta}\,dy\right)\phi\,d\underset{\sim}{x} = \int_\Omega \mathscr{T}_{\gamma\delta}(\mathscr{S}^{-1}(\tfrac{1}{2}x_\alpha x_\beta\psi))\partial_{\gamma\delta}U_3^0\phi\,d\underset{\sim}{x} \tag{41}$$

for all ϕ in $\mathscr{D}(\Omega')$ (here we use that $\psi\equiv 1$ on supp (ϕ), $\phi\in\mathscr{D}(\Omega')$). Since $\Omega'\subset\subset\Omega$ is arbitrary we conclude from (41) that there exists $M_{\alpha\beta\gamma\delta}(\underset{\sim}{x})$ with

$$-\int_{-1}^1 y\Sigma^0_{\alpha\beta}\,dy = M_{\alpha\beta\gamma\delta}(\underset{\sim}{x})\partial_{\gamma\delta}U_3^0,$$

as stated in (34). $M_{\alpha\beta\gamma\delta}$ is given by

$$M_{\alpha\beta\gamma\delta} = \mathscr{T}_{\gamma\delta}(\mathscr{S}^{-1}(\tfrac{1}{2}x_\alpha x_\beta\psi)) \quad \text{in } \Omega', \tag{42}$$

where ψ is any element of $\mathcal{D}(\Omega)$, $\psi\equiv 1$ in Ω'. It is clear from the formula (42) for $M_{\alpha\beta\gamma\delta}$ that it obeys the symmetries

$$M_{\alpha\beta\gamma\delta}=M_{\beta\alpha\gamma\delta}=M_{\alpha\beta\delta\gamma}.$$

By taking the U_3^0 (and $\Sigma_{\gamma\delta}^0$), that correspond to loads $F^\varepsilon=F(\underset{\sim}{x})\in L^2(\Omega)$, $f^\varepsilon=0$ and inserting in (40) we obtain

$$\int_\Omega \mathcal{T}_{\gamma\delta}(F)\partial_{\gamma\delta}\mathcal{S}(G)\phi\,d\underset{\sim}{x}=\int_\Omega \mathcal{T}_{\gamma\delta}(G)\partial_{\gamma\delta}\mathcal{S}(F)\phi\,d\underset{\sim}{x}. \tag{43}$$

Because of continuity (43) holds for all F, $G\in H^{-2}(\Omega)$. Pick $F=\mathcal{S}^{-1}(\frac{1}{2}x_\alpha x_\beta\psi)$ and $G=\mathcal{S}^{-1}(\frac{1}{2}x_\rho x_\sigma\psi)$ with $\psi\equiv 1$ in Ω', it then follows from (42) and (43) that

$$M_{\alpha\beta\rho\sigma}=\mathcal{T}_{\rho\sigma}(\mathcal{S}^{-1}(\tfrac{1}{2}x_\alpha x_\beta\psi))=\mathcal{T}_{\alpha\beta}(\mathcal{S}^{-1}(\tfrac{1}{2}x_\rho x_\sigma\psi))=M_{\rho\sigma\alpha\beta}$$

in Ω'. Since Ω' is arbitrary this verifies the last symmetry of $M_{\alpha\beta\gamma\delta}$.

At this point we only know that $M_{\alpha\beta\gamma\delta}\in L^2_{\mathrm{loc}}(\Omega)$; we now verify that the $M_{\alpha\beta\gamma\delta}$ are indeed L^∞-functions. Consider the identity (38) corresponding to $F^\varepsilon=G(\underset{\sim}{x})$ and $f^\varepsilon=0$, and replace ϕ by ϕ^2:

$$\lim_{\varepsilon_l\to 0}\int_{R(1)}\tilde{\Sigma}^{\varepsilon_l}_{ij}\tilde{E}^{\varepsilon_l}_{ij}\phi^2\,d\underset{\sim}{x}\,dy=-\int_\Omega\left(\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right)\partial_{\alpha\beta}V_3^0\phi^2\,d\underset{\sim}{x}. \tag{44}$$

It follows directly from (3) and the fact that b^ε_{ijkl} is symmetric that

$$C_2^{-1}\sum_{i,j}|\tilde{\Sigma}^\varepsilon_{ij}|^2\leqq\tilde{\Sigma}^\varepsilon_{ij}\tilde{E}^\varepsilon_{ij},$$

consequently,

$$C_2^{-1}\int_{R(1)}\sum_{i,j}|\tilde{\Sigma}^\varepsilon_{ij}|^2\phi^2\,d\underset{\sim}{x}\,dy\leqq\int_{R(1)}\tilde{\Sigma}^\varepsilon_{ij}\tilde{E}^\varepsilon_{ij}\phi^2\,d\underset{\sim}{x}\,dy.$$

Because of the weak lower semicontinuity of the norm, (44) now yields

$$C_2^{-1}\int_{R(1)}\sum_{i,j}|\tilde{\Sigma}^0_{ij}|^2\phi^2\,d\underset{\sim}{x}\,dy\leqq-\int_\Omega\left(\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right)\partial_{\alpha\beta}V_3^0\phi^2\,d\underset{\sim}{x}. \tag{45}$$

Hölder's inequality gives

$$\int_\Omega\sum_{\alpha,\beta}\left|\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right|^2\phi^2\,d\underset{\sim}{x}\leqq\frac{2}{3}\int_{R(1)}\sum_{\alpha,\beta}|\tilde{\Sigma}^0_{\alpha\beta}|^2\phi^2\,d\underset{\sim}{x}\,dy,$$

and therefore in combination with (45) it gives

$$\begin{aligned}\frac{3}{2}C_2^{-1}\int_\Omega\sum_{\alpha,\beta}\left|\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right|^2\phi^2\,d\underset{\sim}{x}&\leqq-\int_\Omega\left(\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right)\partial_{\alpha\beta}V_3^0\phi^2\,d\underset{\sim}{x}\\&\leqq\left(\int_\Omega\sum_{\alpha,\beta}\left|\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right|^2\phi^2\,d\underset{\sim}{x}\right)^{1/2}\\&\quad\cdot\left(\int_\Omega\sum_{\alpha,\beta}|\partial_{\alpha\beta}V_3^0|^2\phi^2\,d\underset{\sim}{x}\right)^{1/2}.\end{aligned}$$

From this we conclude that

$$\int_\Omega\sum_{\alpha,\beta}\left|\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right|^2\phi^2\,d\underset{\sim}{x}\leqq\left(\frac{2}{3}C_2\right)^2\int_\Omega\sum_{\alpha,\beta}|\partial_{\alpha\beta}V_3^0|^2\phi^2\,d\underset{\sim}{x},$$

or in terms of the operators $\mathcal{S}$ and $\mathcal{T}$

$$\int_\Omega\sum_{\alpha,\beta}|\mathcal{T}_{\alpha\beta}(G)|^2\phi^2\,d\underset{\sim}{x}\leqq\left(\frac{2}{3}C_2\right)^2\int_\Omega\sum_{\alpha,\beta}|\partial_{\alpha\beta}\mathcal{S}(G)|^2\phi^2\,d\underset{\sim}{x}. \tag{46}$$

If we pick $G=\mathcal{S}^{-1}(\frac{1}{2}x_\gamma x_\delta t_{\gamma\delta}\psi)$ for some constant symmetric 2 tensor t and some $\psi\in\mathcal{D}(\Omega)$, $\psi\equiv 1$ in $\Omega'\subset\subset\Omega$, then

$$\mathcal{T}_{\alpha\beta}(\mathcal{S}^{-1}(\tfrac{1}{2}x_\gamma x_\delta t_{\gamma\delta}\psi))=\mathcal{T}_{\alpha\beta}(\mathcal{S}^{-1}(\tfrac{1}{2}x_\gamma x_\delta\psi))t_{\gamma\delta}=M_{\alpha\beta\gamma\delta}t_{\gamma\delta}$$

in Ω'. Inserting in (46) we get for any $\phi\in\mathcal{D}(\Omega')$

$$(47)\qquad \int_{\Omega'}\sum_{\alpha,\beta}|M_{\alpha\beta\gamma\delta}t_{\gamma\delta}|^2\phi^2\,d\underset{\sim}{x}\leqq\left(\frac{2}{3}C_2\right)^2\int_{\Omega'}\sum_{\alpha,\beta}|t_{\alpha\beta}|^2\phi^2\,d\underset{\sim}{x}$$

(since $\mathcal{S}(G)=\frac{1}{2}x_\gamma x_\delta t_{\gamma\delta}\psi$). Equation (47) says that $(\sum_{\alpha,\beta}|M_{\alpha\beta\gamma\delta}t_{\gamma\delta}|^2)^{1/2}$ is an L^2 multiplier of norm $\leqq\frac{2}{3}C_2(\sum_{\alpha,\beta}|t_{\alpha\beta}|^2)^{1/2}$, consequently, $(\sum_{\alpha,\beta}|M_{\alpha\beta\gamma\delta}t_{\gamma\delta}|^2)^{1/2}$ is in $L^\infty(\Omega')$ and

$$\left(\sum_{\alpha,\beta}|M_{\alpha\beta\gamma\delta}t_{\gamma\delta}|^2\right)^{1/2}\leqq\frac{2}{3}C_2\left(\sum_{\alpha,\beta}|t_{\alpha\beta}|^2\right)^{1/2}$$

a.e. in Ω'. Since $\Omega'\subset\subset\Omega$ is arbitrary this proves that $M_{\alpha\beta\gamma\delta}\in L^\infty(\Omega)$ and it also verifies the second inequality in (ii). It remains to show that $M_{\alpha\beta\gamma\delta}$ is coercive. Due to the coercivity of b^ε_{ijkl} (cf. (2))

$$c_1\sum_{i,j}|\tilde{E}^\varepsilon_{ij}|^2\leqq\tilde{\Sigma}^\varepsilon_{ij}\tilde{E}^\varepsilon_{ij},$$

so

$$c_1\int_{R(1)}\sum_{i,j}|\tilde{E}^\varepsilon_{ij}|^2\phi^2\,d\underset{\sim}{x}\,dy\leqq\int_{R(1)}\tilde{\Sigma}^\varepsilon_{ij}\tilde{E}^\varepsilon_{ij}\phi^2\,d\underset{\sim}{x}\,dy.$$

Passing to the limit in ε along the sequence $\{\varepsilon_l\}_{l=1}^\infty$, using the relation $\tilde{E}^0_{\alpha\beta}=-y\partial_{\alpha\beta}V^0_3$ (Lemma 2), the identity (44) and the weak lower semicontinuity of the norm, we get

$$\begin{aligned}c_1\int_{R(1)}y^2\sum_{\alpha,\beta}|\partial_{\alpha\beta}V^0_3|^2\phi^2\,d\underset{\sim}{x}\,dy&\leqq c_1\int_{R(1)}\sum_{i,j}|\tilde{E}^0_{ij}|^2\phi^2\,d\underset{\sim}{x}\,dy\\&\leqq-\int_\Omega\left(\int_{-1}^1 y\tilde{\Sigma}^0_{\alpha\beta}\,dy\right)\partial_{\alpha\beta}V^0_3\phi^2\,d\underset{\sim}{x}.\end{aligned}$$

Because of the constitutive relation (34) (which has already been verified) this yields

$$\frac{2}{3}c_1\int_\Omega\sum_{\alpha,\beta}|\partial_{\alpha\beta}V^0_3|^2\phi^2\,d\underset{\sim}{x}\leqq\int_\Omega M_{\alpha\beta\gamma\delta}\partial_{\alpha\beta}V^0_3\partial_{\gamma\delta}V^0_3\phi^2\,d\underset{\sim}{x},$$

or in terms of the operator $\mathcal{S}$

$$(48)\qquad \frac{2}{3}c_1\int_\Omega\sum_{\alpha,\beta}|\partial_{\alpha\beta}\mathcal{S}(G)|^2\phi^2\,d\underset{\sim}{x}\leqq\int_\Omega M_{\alpha\beta\gamma\delta}\partial_{\alpha\beta}\mathcal{S}(G)\partial_{\gamma\delta}\mathcal{S}(G)\phi^2\,d\underset{\sim}{x}.$$

If we pick $G=\mathcal{S}^{-1}(\frac{1}{2}x_\gamma x_\delta t_{\gamma\delta}\psi)$ for some constant symmetric 2 tensor t and some ψ in $\mathcal{D}(\Omega)$, $\psi\equiv 1$ in $\Omega'\subset\subset\Omega$ then

$$\partial_{\alpha\beta}\mathcal{S}(G)=t_{\alpha\beta}$$

in Ω'. Inserting in (48) we get for any $\phi\in\mathcal{D}(\Omega')$

$$\frac{2}{3}c_1\int_{\Omega'}\sum_{\alpha,\beta}|t_{\alpha\beta}|^2\phi^2\,d\underset{\sim}{x}\leqq\int_{\Omega'}M_{\alpha\beta\gamma\delta}t_{\alpha\beta}t_{\gamma\delta}\phi^2\,d\underset{\sim}{x},$$

from which it immediately follows that

$$\frac{2}{3}c_1\sum_{\alpha,\beta}|t_{\alpha\beta}|^2\leqq M_{\alpha\beta\gamma\delta}t_{\alpha\beta}t_{\gamma\delta}$$

a.e. in Ω'. Since $\Omega' \subset\subset \Omega$ is arbitrary this establishes the first inequality in (ii). We have thus completed the proof of our theorem.

REFERENCES

[1] G. Buttazzo and G. Dal Maso, *Γ-limits of integral functionals*, J. Analyse Math., 37 (1980), pp. 145-185.

[2] D. Caillerie, *Thin elastic and periodic plates*, Math. Meth. Appl. Sci., 6 (1984), pp. 159-191.

[3] P. G. Ciarlet and P. Destuynder, *A justification of the two-dimensional linear plate model*, J. Mécanique, 18 (1979), pp. 315-344.

[4] E. De Giorgi, *Quelques problèmes de Γ-convergence*, Proceedings of the conference "Computing Methods in Applied Sciences and Engineering," Versailles 1979, R. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam, 1980, pp. 637-643.

[5] G. A. Francfort and F. Murat, *Homogenization and optimal bounds in linear elasticity*, preprint.

[6] J. Gobert, *Une inequation fondamentale de la théorie de l'élasticité*, Bull. Soc. Roy. Sci. Liège, 31 (1962), pp. 182-191.

[7] R. V. Kohn and M. Vogelius, *A new model for thin plates with rapidly varying thickness*, Int. J. Solids Structures, 20 (1984), pp. 333-350.

[8] ———, *A new model for thin plates with rapidly varying thickness.* II: *A convergence proof*, Quart. Appl. Math., 43 (1985), pp. 1-22.

[9] ———, *A new model for thin plates with rapidly varying thickness.* III: *Comparison of different scalings*, Quart. Appl. Math., 44 (1986), pp. 35-48.

[10] ———, *Thin plates with rapidly varying thickness and their relation to structural optimization*, in Homogenization and Effective Moduli of Materials and Media, J. Ericksen, D. Kinderlehrer, R. Kohn and J. L. Lions, eds., IMA Volumes in Math. and Appl., Springer-Verlag, Berlin-New York-Heidelberg-Tokyo, 1986, to appear.

[11] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., Dover, New York, 1944.

[12] K. A. Lurie and A. V. Cherkaev, *G-closure of a set of anisotropically conducting media in the two-dimensional case*, J. Optim. Theory Appl., 42 (1984), pp. 283-304.

[13] ———, *G-closure of some particular sets of admissible material characteristics for the problem of bending of thin plates*, J. Optim. Theory Appl., 42 (1984), pp. 305-316.

[14] D. Morgenstern and I. Szabo, *Vorlesungen über Theoretische Mechanik*, Springer-Verlag, Berlin, 1961.

[15] F. Murat, *H-convergence*, Seminaire d'analyse fonctionnelle et numérique 1977/78, Département de Mathématiques, Université d'Alger.

[16] R. P. Nordgren, *A bound on the error in plate theory*, Quart. Appl. Math., 28 (1971), pp. 587-595.

[17] L. Tartar, Cours Peccot, Collège de France, 1977.

[18] ———, *Compensated compactness and applications to partial differential equations*, in Nonlinear Mechanics and Analysis, Heriot-Watt Symposium, R. J. Knops, ed., Pitman Research Notes in Mathematics 39, Pitman, London, 1979, pp. 136-212.

[19] L. Tartar, *Estimations fines de coéfficients homogénéisés*, in Ennio de Giorgi's Colloquium, P. Kree, ed., Pitman Research Notes in Mathematics, Pitman, London, to appear.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

014

LINEAR RECURSIVE SCHEMES ASSOCIATED WITH SOME NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS IN ONE DIMENSION AND THE TAU METHOD*

E. L. ORTIZ† AND A. PHAM NGOC DINH‡

Abstract. Standard compactness arguments for Volterra's equation are used to prove the existence of the solution of some nonlinear partial differential equations in one dimension by associating them to linear recursive schemes. Sufficient conditions for the quadratic convergence of hyperbolic and parabolic types are given in this paper. By using a best approximation perturbation technique—the Tau Method—we show that these recursive schemes lead to accurate numerical approximations in concrete problems.

Key words. Tau Method, nonlinear PDE, hyperbolic equations, parabolic equations, singular perturbation of PDE

AMS(MOS) subject classifications. A07, 35J60, 35K55, B25, 45D05, 65M99

1. Introduction. Kalaba has shown in [4] that the solutions of certain classes of *nonlinear* ordinary and partial differential equations may be represented in terms of maximal operations applied to the solution of associated *linear* equations. The technique of quasilinearization was introduced by Bellman [2], who used it successfully in the analysis of the initial value problem associated with Riccati's equation.

The nonlinear partial differential equations considered in this paper are of the form

$$L(u)=\varepsilon f(t, u(x,t)), \tag{0}$$

where u is an unknown function defined on a domain D and $\varepsilon: 0<\varepsilon\leqq 1$ is a small parameter. The linear differential operators L considered here are the wave and heat operators in one dimension. The function $f(t,u)$ is assumed to be continuous in (t,u) and to satisfy a Lipschitz condition in u.

In the hyperbolic case this paper can be regarded as a generalization of Pham Ngoc Dinh [13], where stronger assumptions on f are required. We associate with our problem $L(u)=\varepsilon f$ a linear recursive scheme for which the existence of solution is proved by using standard compactness arguments related to Volterra's equation. If $f(t,u)$ is twice differentiable in u, quadratic convergence is shown for the case of a parabolic operator; this result extends a previous result of Kalaba [4]. A similar result is proved for the hyperbolic case. Both results are local in time.

The linear recursive schemes developed in this paper enabled us to use a perturbation technique based on the ideas of best uniform approximation, which extends very considerably the classical Tau Method of Lanczos, and which has attracted considerable attention in the last few years. We give examples in which by using the recursive formulation of the Tau Method of Ortiz [8] we are able to construct very accurate numerical approximations to two nonlinear partial differential equations: one of hyperbolic and one of parabolic type.

2. A weakly nonlinear hyperbolic problem in one dimension. In this section we are concerned with the solution of the nonhomogeneous one-dimensional wave equation with a right-hand side depending on the unknown function $u(x,t)$.

* Received by the editors April 7, 1983, and in revised form December 18, 1984.

† Mathematics Department, Imperial College, London SW7 2BZ, England.

‡ Université d'Orleans, Département de Mathématique et d'Informatique, 45046 Orleans Cedex, France.

Let us consider the problem of finding a function $u(x, t)$ satisfying the equation:

$$\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x^2}=\varepsilon f(t, u) \quad \text{for } 0<x<1, \quad 0<t<T, \tag{1}$$

and such that

$$\begin{aligned} &u(0, t)=u(1, t)=0, \qquad u(x, 0)=\tilde{u}_0(x), \\ &\frac{\partial u}{\partial t}(x, 0)=\tilde{u}_1(x) \quad \text{for } 0<x<1, \quad 0<t<T. \end{aligned} \tag{2}$$

We make the following assumptions on the function f in (1):

(3) (i) f is locally Lipschitz with respect to u, i.e. for each $T>0$, there exist $A(t)$ such that $|f(t, u_1)-f(t, u_2)| \leqq A(t)|u_1-u_2|$ for each $t \in]0, T[$ and with $A(t) \in L^2(0, T)$.

(ii) $f(t, u)$ is a continuous function with respect to the two variables t and u.

(3′) $\tilde{u}_0(x)$ is given in $H_0^1(\Omega)$ and $\tilde{u}_1(x)$ in $L^2(\Omega)$, where $\Omega =]0, 1[$, ε is a small parameter $(0<\varepsilon \leqq 1)$.

We write $f(u) := f(t, u)$; $u(t) := u(x, t)$; $\dot{u}(t) := \partial u/\partial t$.

2.1. Definition of a bounded sequence $u_n(t)$. Let us introduce the sequence of functions $\{u_n(t)\}$, $u_n(t) := u_n(x, t)$, defined by the linear recursive relations:

$$\begin{aligned} &\frac{\partial^2 u_1}{\partial t^2}-\frac{\partial^2 u_1}{\partial x^2}=\varepsilon f(u_0), \\ &u_1(0, t)=u_1(1, t)=0, \\ &u_1(x, 0)=\tilde{u}_0(x); \qquad \dot{u}_1(x, 0)=\tilde{u}_1(x). \end{aligned} \tag{4}$$

$$\begin{aligned} &\frac{\partial^2 u_{n+1}}{\partial t^2}-\frac{\partial^2 u_{n+1}}{\partial x^2}=\varepsilon f(u_n), \\ &u_{n+1}(0, t)=u_{n+1}(1, t)=0, \\ &u_{n+1}(x, 0)=\tilde{u}_0(x); \qquad \dot{u}_{n+1}(x, 0)=\tilde{u}_1(x); \end{aligned} \tag{5}$$

the first function $u_0(t)$ will be determined for

$$u_0(t) \in L^\infty(0, T; H_0^1(\Omega)), \qquad \dot{u}_0(t) \in L^\infty(0, T; L^2(\Omega)).$$

LEMMA 1. *The solution $u_{n+1}(t)$ of* (5) *exists and belongs to a bounded set of $L^\infty(0, T; H_0^1(\Omega))$, and $\dot{u}_{n+1}(t)$ belongs to a bounded set of $L^\infty(0, T; L^2(\Omega))$; these sets are bounded independently of n and ε.*

Proof. Let us suppose that u_n is in a bounded set of $L^\infty(0, T; H_0^1(\Omega))$ and that $\dot{u}_n$ is in a bounded set of $L^\infty(0, T; L^2(\Omega))$, i.e.:

Let M be a constant independent of n and ε such that

$$\|u_n\|_{H_0^1(\Omega)} \leqq M, \quad \|\dot{u}_n\|_{L^2(\Omega)} \leqq M, \quad \text{a.e. } t \in [0, T]. \tag{6}$$

Then, let us show that $u_{n+1}(t)$ and $\dot{u}_{n+1}(t)$ are in the same bounded sets of $L^\infty(0, T; H_0^1(\Omega)$ (and $L^\infty(0, T; L^2(\Omega))$ respectively.

Equation (5) is equivalent to the following variational formulation:

$$\begin{aligned} &a(u_{n+1}, v)+\frac{d}{dt}\langle \dot{u}_{n+1}, v\rangle_{L^2(\Omega)}=\varepsilon\langle f(u_n), v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega), \\ &u_{n+1}(0)=\tilde{u}_0, \qquad \dot{u}_{n+1}(0)=\tilde{u}_1 \end{aligned} \tag{7}$$

where $a(u, v) = \int_0^1 (\partial u/\partial x) \cdot (\partial v/\partial x)\, dx$ and $\langle u, v\rangle_{L^2(\Omega)}$ is the scalar product in $L^2(\Omega)$. Let $u^j_{n+1}(t)$ be the sequence defined by:

$$u^j_{n+1}(t) = \sum_{k=1}^{j} \xi^{n+1}_{k,j}(t) \cdot v_k(x) \tag{8}$$

($\{v_k(x)\}$ is a "basis" of $H^1_0(\Omega)$ i.e. a countable and everywhere dense subset of $H^1_0(\Omega)$; the $\{v_k(x)\}$ can be the eigenfunctions of $\partial^2/\partial x^2$), which is the solution of

$$\begin{aligned} &a(u^j_{n+1}, v_p) + \frac{d}{dt}\langle \dot{u}^j_{n+1}, v_p\rangle_{L^2(\Omega)} = \varepsilon\langle f(u_n), v_p\rangle_{L^2(\Omega)}, \\ &u^j_{n+1}(0) = \tilde{u}_{0j}, \qquad \dot{u}^j_{n+1}(0) = \tilde{u}_{1j}, \end{aligned} \tag{9}$$

where

$$\tilde{u}_{0j}(x) = \sum_{k=1}^{j} \eta_{kj} \cdot v_k(x) \to \tilde{u}_0(x) \quad \text{in } H^1_0(\Omega) \text{ strongly,}$$

$$\tilde{u}_{1j}(x) = \sum_{k=1}^{j} \gamma_{kj} \cdot v_k(x) \to \tilde{u}_1(x) \quad \text{in } L^2(\Omega) \text{ strongly.}$$

The coefficients $\xi^{n+1}_{kj}(\mathrm{t})$ verify the linear differential equation of second order [3]:

$$\begin{aligned} &\ddot{\xi}^{n+1}_{k,j}(t) + k^2\pi^2\xi^{n+1}_{k,j}(t) = 2\varepsilon\langle f(u_n), v\rangle_{L^2(\Omega)}, \\ &\xi^{n+1}_{k,j}(0) = \eta_{k,j}, \qquad \dot{\xi}^{n+1}_{k,j}(0) = \gamma_{kj}, \quad 1 \leqq k \leqq j. \end{aligned} \tag{10}$$

Let us multiply (9) by $\dot{\xi}^{n+1}_{k,j}(t)$ and sum up with respect to the index k,

$$\frac{1}{2}\frac{d}{dt}\|u^j_{n+1}\|^2_{H^1_0(\Omega)} + \frac{1}{2}\frac{d}{dt}\|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)} = \varepsilon\langle f(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}.$$

Therefore by integration we obtain

$$\|u^j_{n+1}\|^2_{H^1_0(\Omega)} + \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)} = \|\tilde{u}_{0j}\|^2_{H^1_0(\Omega)} + \|\tilde{u}_{1j}\|^2_{L^2(\Omega)} + 2\varepsilon\int_0^t \langle f(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta. \tag{11}$$

Since $\tilde{u}_{0j}$ and $\tilde{u}_{1j}$ converge to $\tilde{u}_0$ and $\tilde{u}_1$ in $H^1_0(\Omega)$ and $L^2(\Omega)$ respectively, there exists a constant C_1 independent of n and j such that:

$$\|\tilde{u}_{0j}\|^2_{H^1_0(\Omega)} + \|\tilde{u}_{1j}\|^2_{L^2(\Omega)} \leqq C_1. \tag{12}$$

On the other hand:

$$\left|2\varepsilon\int_0^t \langle f(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta\right| \leqq \int_0^t \|f(u_n)\|^2_{L^2(\Omega)}\, d\theta + \int_0^t \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}\, d\theta.$$

But in one dimension the imbedding of $H^1(\Omega) = \{u \mid u, \partial u/\partial x \in L^2(\Omega)\}$, Sobolev space of order 1, in $\mathscr{C}^0(\bar{\Omega})$ space of continuous functions defined in $\bar{\Omega}$, is continuous [1]. Hence

$$|u_n(t)| \leqq \|u_n\|_{\mathscr{C}^0(\bar{\Omega})} \leqq \sqrt{2}\|u_n\|_{H^1_0(\Omega)} \leqq \sqrt{2M}. \tag{13}$$

Therefore

$|f(u_n)| \leqq C_2$, where C_2 is a constant independent of n, ε

on account of the continuity of the function f on a compact set.

Let us consider:

$$S^j_{n+1}(t) = \|u^j_{n+1}\|^2_{H^1_0(\Omega)} + \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}. \tag{14}$$

From (11), (12) and (13):

$$S^j_{n+1}(t) \leqq C_1 + \int_0^t (C_2^2 + \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)})\, d\theta.$$

Hence

$$S^j_{n+1}(t) \leqq C_1 + \int_0^t (C_2^2 + S^j_{n+1}(\theta))\, d\theta. \tag{15}$$

The theory of Volterra integral equations [5] implies that

$$S^j_{n+1}(t) \leqq S(t) \quad \text{a.e. } t \in [0, \bar{T}[, \quad \bar{T} \leqq T \tag{16}$$

where $S(t)$ is the maximal solution of Volterra's equation:

$$S(t) = C_1 + \int_0^t (C_2^2 + S(\theta))\, d\theta,$$

and the solution is defined in $[0, \bar{T}[$. Hence

$$S^j_{n+1}(t) \leqq S(t) \quad \text{in } [0, \hat{T}] \quad \text{with } \hat{T} < \bar{T} \leqq T. \tag{17}$$

Let us then choose the bound M of (6) such that

$$2C_1 < M^2 \qquad (C_1 \text{ given by (12)}). \tag{18}$$

But $S(0) = C_1$; $S(t)$ being a continuous function, there exists an interval $[0, \hat{\hat{T}}]$($\hat{\hat{T}} \leqq \hat{T}$) such that

$$S(t) \leqq M^2 \quad \text{for each } t \in [0, T] \qquad (\hat{\hat{T}} \text{ will now be called } T). \tag{19}$$

From (17) and (19)

$$\|u^j_{n+1}\|_{H_0^1(\Omega)} \leqq M; \qquad \|\dot{u}^j_{n+1}\|_{L^2(\Omega)} \leqq M \quad \text{a.e. } t \in [0, T], \tag{20}$$

that is

$$\begin{aligned} &u^j_{n+1} \text{ is in a bounded set of } L^\infty(0, T; H_0^1(\Omega)),\\ &\dot{u}^j_{n+1} \text{ is in a bounded set of } L^\infty(0, T; L^2(\Omega)),\\ &\text{sets bounded independently of } j, n, \varepsilon. \end{aligned} \tag{21}$$

From (21) we can extract a subsequence, still denoted by $\{u^j_{n+1}\}$, such that

$$\begin{aligned} &\text{(i)} \quad u^j_{n+1} \to u_{n+1} \quad \text{in } L^\infty(0, T, H_0^1(\Omega)) \text{ weakly} * \text{ when } j \to \infty,\\ &\text{(ii)} \quad \dot{u}^j_{n+1} \to \dot{u}_{n+1} \quad \text{in } L^\infty(0, T; L^2(\Omega)) \text{ weakly} * \text{ when } j \to \infty. \end{aligned} \tag{22}$$

We can easily check from (9) and (22) that u_{n+1} satisfies (7) in $L^\infty(0, T)$ weakly $*$. From (20) we deduce that

(23) the solution $u_{n+1}(t)$ of (7) remains in a bounded set of $L^\infty(0, T; H_0^1(\Omega))$ and $\dot{u}_{n+1}(t)$ in a bounded set of $L^\infty(0, T; L^2(\Omega))$.

2.2. Solution of the initial problem.

1. *Existence of the solution of* (1)-(2). We shall show that $\{u_n(t)\}$ is a Cauchy sequence in $L^\infty(0, T; L^2(\Omega))$. Let d_{n+1} be $d_{n+1} = u_{n+1} - u_n$ with $d_{n+1}(0) = \dot{d}_{n+1}(0) = 0$. d_{n+1} satisfies the equation

$$\begin{aligned} &\ddot{d}_{n+1} - \frac{\partial^2}{\partial x^2}(d_{n+1}) = \varepsilon[f(u_n) - f(u_{n-1})],\\ &d_{n+1}(0) = \dot{d}_{n+1}(0) = 0, \quad d_{n+1} \in L^\infty(0, T; H_0^1(\Omega)), \quad \dot{d}_{n+1} \in L^\infty(0, T; L^2(\Omega)). \end{aligned} \tag{24}$$

Let us consider $s \in]0, T[$ and define $\psi_{n+1}(t)$ [6] by

$$(25) \qquad \psi_{n+1}(t) = \begin{cases} -\int_t^s d_{n+1}(\sigma)\,d\sigma, & s \geqq t, \\ 0, & s < t. \end{cases}$$

Let us set

$$(25') \qquad w_{n+1}(t) = \int_0^t d_{n+1}(\sigma)\,d\sigma.$$

Multiplying (24) by $\psi_{n+1}(t)$, we obtain, after integrating by parts

$$\|d_{n+1}(s)\|^2_{L^2(\Omega)} + \|w_{n+1}(s)\|^2_{H_0^1(\Omega)} = -2\varepsilon \int_0^s \langle [f(u_n) - f(u_{n-1})], \psi_{n+1}\rangle_{L^2(\Omega)}\,dt.$$

Using the assumption (3)(i) it follows that

$$(26) \qquad \|d_{n+1}(s)\|^2_{L^2(\Omega)} + \|w_{n+1}(s)\|^2_{H_0^1(\Omega)} \leqq 2\varepsilon \int_0^s A(t)\|d_n\|_{L^2(\Omega)} \cdot \|\psi_{n+1}\|_{L^2(\Omega)}\,dt.$$

But

$$(27) \qquad \|\psi_{n+1}\|^2_{L^2(\Omega)} = \int_0^1 \left|\int_t^s d_{n+1}(\sigma)\right|^2 dx \leqq T^2\|d_{n+1}\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

Finally, due to (26) and (27):

$$(28) \qquad \|d_{n+1}\|_{L^\infty(0,T;L^2(\Omega))} \leqq \left(2\varepsilon T \int_0^T A(t)\,dt\right)\|d_n\|_{L^\infty(0,T;L^2(\Omega))}.$$

Let

$$k = 2\varepsilon T \int_0^T A(t)\,dt < 1$$

(a condition which is always satisfied by taking a sufficiently small ε or a suitable T).

Therefore

(29) $\{u_n(t)\}$ is a Cauchy sequence in $L^\infty(0, T; L^2(\Omega))$;
hence in $L^2(0, T; L^2(\Omega)) = L^2(Q)$ where $Q =]0, T[\times]0, 1[$.

From (23) it is possible to extract from $\{u_n(t)\}$ a subsequence, still denoted by $\{u_n(t)\}$, such that

$$u_{n+1} \to u \quad \text{in } L^\infty(0, T; H_0^1(\Omega)) \text{ weakly} * \text{ when } n \to \infty,$$
$$\dot{u}_{n+1} \to \dot{u} \quad \text{in } L^\infty(0, T; L^2(\Omega)) \text{ weakly} * \text{ when } n \to \infty.$$

From (20) we can deduce that

(30) the limit u of the subsequence $\{u_n\}$ is such that
$u_n \to u$ (a.e. in Q).

From (30) we can show, by using a lemma on weak convergence [7], that

$$\langle f(t, u_n), v_p\rangle_{L^2(\Omega)} \to \langle f(t, u), v_p\rangle_{L^2(\Omega)} \quad \text{in } L^\infty(0, T) \text{ weakly} *.$$

Then we can take limits in (7) and find that u satisfies the equation

$$\begin{aligned} &a(u, v)+(d/dt)\langle \dot{u}, v\rangle_{L^2(\Omega)} = \varepsilon\langle f(u), v\rangle_{L^2(\Omega)} \quad \text{for each } v \in H_0^1(\Omega), \\ &u(0) = \tilde{u}_0, \qquad u(0) = \tilde{u}_1. \end{aligned} \tag{31}$$

Expression (31) is the variational formulation of (1)-(2).

By using the same argument as in § 2.1 we obtain

$$\|u_{n+1} - u\|_{L^\infty(0,T,L^2(\Omega))} \leqq k\|u_n - u\|_{L^\infty(0,T;L^2(\Omega))} \qquad (k<1).$$

Remark 1. By using the function $\psi_{n+1}(t)$ defined in (25) we can show the uniqueness of the problem (7).

2. *Uniqueness of the solution of the problem* (1)-(2). Let u and b be two solutions of (1)-(2); if we set $w = u - v$, then w verifies the following equation:

$$\begin{aligned} &\ddot{w} - \frac{\partial^2 w}{\partial x^2} = \varepsilon[f(t, u) - f(t, v)], \\ &w(0) = \dot{w}(0) = 0, \quad w \in L^\infty(0, T; H_0^1(\Omega)), \quad \dot{w} \in L^\infty(0, T; L^2(\Omega)). \end{aligned} \tag{32}$$

As before let us introduce the function $\psi(t)$ defined by

$$\psi(t) = \begin{cases} -\displaystyle\int_t^s w(\sigma)\, d\sigma, & s \geqq t, \\ 0, & s < t. \end{cases}$$

Multiplying (32) by $\psi(t)$, and after integrating by parts we find that

$$\|w(s)\|^2_{L^2(\Omega)} + \|w_1(s)\|^2_{H_0^1(\Omega)} = -2\varepsilon \int_0^s \langle [f(t, u) - f(t, v)], \psi \rangle_{L^2(\Omega)}\, dt,$$

where $w_1(t) = \int_0^t w(\sigma)\, d\sigma$. If we set

$$\sigma(s) = \|w(s)\|^2_{L^2(\Omega)} + \|w_1(s)\|^2_{H_0^1(\Omega)},$$

we obtain, by using hypothesis (3)(i) and the inequality $2ab \leqq (1/\alpha)a^2 + \alpha b^2$ for each $\alpha > 0$, that

$$\sigma(s) \leqq 2\varepsilon \max\left(\alpha, \frac{1}{\alpha} + \frac{1}{\alpha'}\right) \int_0^s A(t)\sigma(t)\, dt + \varepsilon\alpha' \|w_1(s)\|^2_{H_0^1(\Omega)} \int_0^s A(t)\, dt \tag{33}$$

for each α, $\alpha' > 0$.

Then let us choose α' such that $\varepsilon\alpha' \int_0^T A(t)\, dt < 1$. Therefore we obtain finally

$$\sigma(s) \leqq C(T)\left(\int_0^T A^2(t)\, dt\right)^{1/2} \left(\int_0^s \sigma^2(t)\, dt\right)^{1/2}, \tag{34}$$

where $C(T)$ is a constant only depending on T. By using Gronwall's lemma we have:

$$\sigma(s) = 0 \quad \text{i.e.} \quad u = v.$$

Hence, we have the following.

THEOREM 1. *The problem* (1)-(2) *under the assumptions* (3)(i), (ii) *and* (3') *has one and only one solution* $u \in L^\infty(0, T; H_0^1(\Omega))$ *and such that* $\dot{u} \in L^\infty(0, T; L^2(\Omega))$.

Remark 2. From the uniqueness of the solution of the problem (1)-(2) it follows that the total sequence $\{u_n(t)\}$ converges to the solution $u(t)$.

3) *Limit when ε approaches* 0. Let us denote u_ε the solution of (1)-(2). On account of (21) u_ε and $\dot{u}_\varepsilon$ are in a bounded set of $L^\infty(0, T; H_0^1(\Omega))$ and $L^\infty(0, T; L^2(\Omega))$ respectively; therefore we can prove as before the following result:

THEOREM 2. *When $\varepsilon \to 0$, $u_\varepsilon \to \bar{u}$ unique solution of*

$$(35)\qquad \begin{aligned} &a(u, v)+\frac{d}{dt}\langle \dot{u}, v\rangle_{L^2(\Omega)}=0,\\ &u(0)=\tilde{u}_0, \quad \dot{u}(0)=\tilde{u}_1 \quad \textit{in } L^2(Q) \textit{ strongly.}\end{aligned}$$

3. A weakly nonlinear parabolic problem in one dimension. In this section we consider the partial differential equation

$$(36)\qquad \begin{aligned} &\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=\varepsilon f(u), \qquad 0<x<1, \quad 0<t<T,\\ &u(0, t)=u(1, t)=0,\\ &u(x, 0)=\tilde{u}_0(x).\end{aligned}$$

3.1. Case of simple convergence. Let us assume the f satisfies hypothesis (3)(i), (ii). Let us define the sequence of functions $\{u_n(t)\}$ by the linear recursive expression

$$(37)\qquad \begin{aligned} &\frac{\partial u_{n+1}}{\partial t}-\frac{\partial^2 u_{n+1}}{\partial x^2}=\varepsilon f(u_n), \quad n\geqq 0, \quad 0<x<1, \quad 0<t<T,\\ &u_{n+1}(0, t)=u_{n+1}(1, t)=0,\\ &u_{n+1}(x, 0)=\tilde{u}_0(x)\in H_0^1(\Omega),\end{aligned}$$

with $u_0(t)\in L^\infty(0, T; H_0^1(\Omega))$ and such that $\dot{u}_0(t)\in L^2(0, T; L^2(\Omega))$.

By using the same arguments as in § 2 it is possible to show that

LEMMA 2. *The solution $u_{n+1}(t)$ of the equation* (37) *exists, is unique, belongs to a bounded set of $L^\infty(0, T; H_0^1(\Omega))$ and $\dot{u}_{n+1}(t)$ is in a bounded set of $L^2(0, T; L^2(\Omega))$; the sets are bounded independently of n and ε.*

THEOREM 3. *The problem* (36) *under the assumptions* (3)(i), (ii) *has one and only one solution $u\in L^\infty(0, T; H_0^1(\Omega))$ and is such that $\dot{u}\in L^2(0, T; L^2(\Omega))$. We also have*

$$\lim_{n\to\infty} \|u_{n+1}-u\|_{L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H_0^1(\Omega))}=0,$$

where $u_{n+1}(t)$ is the solution of (37).

Let us denote $u_\varepsilon(t)$ the solution of (36). Then, we have the following.

THEOREM 4. *When $\varepsilon\to 0$, $u_\varepsilon\to\bar{u}$, unique solution of*:

$$\begin{aligned} &a(u, v)+\langle \dot{u}, v\rangle_{L^2(\Omega)}=0 \quad \textit{for all } v\in H_0^1(\Omega),\\ &u(0)=\tilde{u}_0 \quad \textit{in } L^2(0, T; H_0^1(\Omega)) \textit{ strongly.}\end{aligned}$$

3.2. Case of quadratic convergence. Let us assume that f satisfies the conditions 3(i), (ii) and also that

(38) the first and second derivatives f'_u and f''_{u^2} with respect to u exist
and are continuous with respect to (t, u).

Let $\dot{u}$ stand for the partial derivative of u with respect to t. As before, we will associate with (36) a sequence of functions $\{u_n(t)\}$ defined now by the linear recursive scheme

$$(39)\qquad \begin{aligned} &\frac{\partial u_{n+1}}{\partial t}-\frac{\partial^2 u_{n+1}}{\partial x^2}=\varepsilon[f(u_n)+(u_{n+1}-u_n)f'(u_n)], \qquad n\geqq 0,\\ &u_{n+1}(0, t)=u_{n+1}(1, t)=0,\\ &u_{n+1}(x, 0)=\tilde{u}_0(x).\end{aligned}$$

Let us set

$$F_n(t, u)=f(u_n)+(u-u_n)f'(u_n).$$

LEMMA 3. *The solution $u_{n+1}(t)$ of* (39) *exists, is unique and belongs to a bounded set of $L^\infty(0, T; H_0^1(\Omega))$, and $\dot{u}_{n+1}(t)$ is in a bounded set of $L^2(0, T; L^2(\Omega))$.*

Proof. Let us suppose that u_n and $\dot{u}_n$ are in bounded sets of $L^\infty(0, T; H_0^1(\Omega))$ and $L^2(0, T; L^2(\Omega))$ respectively,

$$\text{(40)} \qquad \begin{aligned} &\|u_n\|_{H_0^1(\Omega)} \leqq M \quad \text{a.e. } t\in[0, T], \\ &\int_0^T \|\dot{u}_n\|^2_{L^2(\Omega)}\, d\theta \leqq M^2, \end{aligned}$$

where M is a constant independent of n and ε. Let $u^j_{n+1}(t)$ be the sequence defined by

$$\text{(41)} \qquad u^j_{n+1}(t)=\sum_{k=1}^{j} \xi^{n+1}_{k,j}(t)\cdot v_k(x) \qquad (\{v_k(x)\} \text{ is a “basis” of } H_0^1(\Omega)),$$

solution of the equation

$$\text{(42)} \qquad \begin{aligned} &a(u^j_{n+1}, v_p)+\langle \dot{u}^j_{n+1}, v_p\rangle_{L^2(\Omega)}=\varepsilon\langle F_n(t, u^j_{n+1}), v_p\rangle_{L^2(\Omega)}, \\ &u^j_{n+1}(0)=\tilde{u}_{0j}, \end{aligned}$$

where $\tilde{u}_{0j}=\sum_{k=1}^{j} \eta_{kj}\cdot v_k(x)\rightarrow \tilde{u}_0(x)$ in $H_0^1(\Omega)$ strongly. The coefficients $\xi^{n+1}_{k,j}$ satisfy the following first order linear differential equation

$$\text{(43)} \qquad \begin{aligned} &\dot{\xi}^{n+1}_{k,j}(t)+(k^2\pi^2-\varepsilon f'(u_n))\xi^{n+1}_{k,j}(t)=2\varepsilon\langle[f(u_n)-u_nf'(u_n)], v_k\rangle_{L^2(\Omega)}, \\ &\xi^{n+1}_{k,j}(0)=\eta_{kj}, \qquad 1\leqq k\leqq j. \end{aligned}$$

Let us multiply (42) by $\dot{\xi}^{n+1}_{k,j}$ and sum with respect to the index k:

$$\text{(44)} \quad \|u^j_{n+1}\|^2_{H_0^1(\Omega)}+2\int_0^t \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}\, d\theta=2\varepsilon\int_0^t \langle F_n(\theta, u^j_{n+1}), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta+\|\tilde{u}_{0j}\|^2_{H_0^1(\Omega)}$$

with

$$\text{(45)} \qquad \begin{aligned} &2\varepsilon\int_0^t \langle F(\theta, u^j_{n+1}), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta \\ &\quad=2\varepsilon\int_0^t \langle u^j_{n+1}f'(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta+2\varepsilon\int_0^t \langle f(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta \\ &\quad-2\varepsilon\int_0^t \langle u_nf'(u_n), \dot{u}^j_{n+1}\rangle_{L^2(\Omega)}\, d\theta. \end{aligned}$$

By using the hypothesis of recurrence (40), assumption (38) and the inequality $2ab\leqq a^2/\alpha+\alpha b^2$ for each $\alpha>0$ we obtain

$$\text{(46)} \qquad \begin{aligned} &\|u^j_{n+1}\|^2_{H_0^1(\Omega)}+2\int_0^t \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}\, d\theta \\ &\quad\leqq C_1+\int_0^t [C_2+(\alpha+\beta+\gamma)\|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}+C_3\|u^j_{n+1}\|^2_{L^2(\Omega)}]\, d\theta \end{aligned}$$

for each α, β, $\gamma>0$ and where C_1 (C_1 such that $\|\tilde{u}_{0j}\|^2_{H_0^1(\Omega)}\leqq C_1$), C_2 and C_3 are constants independent of n and ε.

Let us set

$$S^j_{n+1}(t) = \|u^j_{n+1}\|^2_{H^1_0(\Omega)} + \int_0^t \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}\, d\theta \tag{47}$$

and let us choose $\alpha = \beta = \gamma = \frac{1}{3}$. Then we finally obtain

$$S^j_{n+1}(t) \leqq C_1 + \int_0^t (C_2 + C_3 S^j_{n+1}(\theta))\, d\theta. \tag{48}$$

As in § 2, if we choose, for instance, M such that $M^2 > 2C_1$ we can deduce that there exists an interval $[0, T]$ such that

$$\begin{aligned} &\|u^j_{n+1}(t)\|_{H^1_0(\Omega)} \leqq M \qquad \text{a.e. } t \in [0, T], \\ &\int_0^T \|\dot{u}^j_{n+1}\|^2_{L^2(\Omega)}\, d\theta \leqq M^2. \end{aligned} \tag{49}$$

From (49) and if we take limits in equation (42), we find that there exists one and only one $u_{n+1}(t)$ which solves

$$\begin{aligned} &a(u_{n+1}, v) + (\dot{u}_{n+1}, v)_{L^2(\Omega)} = \varepsilon\langle F_n(t, u_{n+1}), v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega), \\ &u_{n+1}(0) = \tilde{u}_0. \end{aligned} \tag{50}$$

Quadratic convergence. Let us show now that $\{u_n(t)\}$ is a Cauchy sequence in $L^\infty(0, T; L^2(\Omega))$. Let d_{n+1} be such that $d_{n+1} = u_{n+1} - u_n$, with $d_{n+1}(0) = 0$. $d_{n+1}(t)$ satisfies the equation

$$a(d_{n+1}, v) + \langle \dot{d}_{n+1}, v\rangle_{L^2(\Omega)} = \varepsilon\langle [F_n(u_{n+1}) - F_{n-1}(u_n)], v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega) \tag{51}$$

with

$$\begin{aligned} F_n(u_{n+1}) - F_{n-1}(u_n) &= \varepsilon[f(u_n) + (u_{n+1} - u_n)f'(u_n) - f(u_{n-1}) - (u_n - u_{n-1})f'(u_{n-1})] \\ &= \varepsilon\left[\frac{d_n^2}{2} f''(q_n) + d_{n+1} \cdot f'(u_n)\right] \end{aligned} \tag{52}$$

and where $q_n = \lambda u_n + (1-\lambda)u_{n-1} (0 < \lambda < 1)$. By using a lemma on regularity [7] we obtain from (51)

$$\begin{aligned} &2\int_0^t \|d_{n+1}\|^2_{H^1_0(\Omega)}\, d\theta + \|d_{n+1}\|^2_{L^2(\Omega)} \\ &\qquad = 2\varepsilon \int_0^t \left\langle \frac{d_n^2}{2} f''(q_n) + f'(u_n) \cdot d_{n+1}, d_{n+1}\right\rangle_{L^2(\Omega)} d\theta. \end{aligned} \tag{53}$$

The continuous imbedding of $H^1_0(\Omega)$ in $\mathscr{C}^0(\bar{\Omega})$ enables us to show that

$$\left|\left\langle \frac{d_n^2}{2} f''(q_n), d_{n+1}\right\rangle_{L^2(\Omega)}\right| \leqq \frac{\alpha c^2}{16}\|d_{n+1}\|^2_{H^1_0(\Omega)} + \frac{1}{\alpha}\|d_n\|^4_{L^2(\Omega)} \quad \text{for all } \alpha > 0 \tag{54}$$

where C is a constant independent of n and ε.

If we choose α and T sufficiently small, we can deduce that there exist two constants β and $\gamma > 0$ such that:

$$\beta \int_0^T \{\|d_{n+1}\|^2_{H^1_0(\Omega)} + \gamma\|d_{n+1}\|^2_{L^\infty(0,T;L^2(\Omega))}\}\, d\theta \leqq \frac{2\varepsilon T}{\alpha}\|d_n\|^4_{L^\infty(0,T;L^2(\Omega))}.$$

Finally

$$\|d_{n+1}\|_{L^\infty(0,T;L^2(\Omega))} \leqq k_T^{(0)}\|d_n\|^2_{L^\infty(0,T;L^2(\Omega))} \tag{55}$$

with $(k_T^{(0)})^2=2\varepsilon T/\alpha$. Therefore, for a suitable value of T, $\{u_n\}$ is a Cauchy sequence in $L^\infty(0, T; L^2(\Omega))$ (ε is not necessarily small).

Taking limits in (50), we find that the limit u is the solution of:

$$\tag{56} \begin{aligned} &a(u, v)+\langle \dot u, v\rangle_{L^2(\Omega)}=\varepsilon\langle f(u), v\rangle_{L^2(\Omega)} \qquad \text{for all } v\in H_0^1(\Omega).\\ &u(0)=\tilde u_0. \end{aligned}$$

Equation (56) is the variational formulation of (36).

The convergence of $\{u_n\}$ to u is such that

$$\tag{57} \|u_{n+1}-u\|_{L^\infty(0,T;L^2(\Omega))}\leqq k_T^{(0)}\|u_n-u\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

Hence, we have the following:

THEOREM 5. *Problem* (36), *with assumptions* (3)(ii) *and* (38), *has one and only one solution* $u\in L^\infty(0, T; H_0^1(\Omega))$. *The sequence* $\{u_n\}$ *defined by* (37) *converges quadratically to it in the meaning defined by* (57).

Remark 3. Let us go back to the hyperbolic case. If the function $f(t, u)$ satisfies the assumptions (3)(ii) and (38), we can show that there exists an inequality similar to (55): let us consider the sequence $\{u_n\}$ defined by

$$\tag{58} \frac{\partial^2 u_{n+1}}{\partial t^2}-\frac{\partial^2 u_{n+1}}{\partial x^2}=\varepsilon F_n(t, u_{n+1})$$

with $F_n(t, u)=f(u_n)+(u-u_n)f'(u_n)$, where $u_{n+1}(t)$ satisfies condition (2). As in the parabolic case we show that the sequences $\{u_n\}$ and $\{\dot u_n\}$ are in bounded sets of $L^\infty(0, T; H_0^1(\Omega))$ and $L^\infty(0, T; L^2(\Omega))$ respectively.

Let $d_{n+1}=u_{n+1}-u_n$. $d_{n+1}(t)$ satisfy the equation

$$\tag{59} \begin{aligned} &\ddot d_{n+1}(t)-\frac{\partial^2}{\partial x^2}d_{n+1}=\varepsilon[F_n(t, u_{n+1})-F_{n-1}(t, u_n)],\\ &d_{n+1}(0)=\dot d_{n+1}(0)=0. \end{aligned}$$

Let us introduce the functions $\psi_{n+1}(t)$ and $w_{n+1}(t)$ defined by (25) and (25'). We have

$$\tag{60} \begin{aligned} &\|d_{n+1}(s)\|^2_{L^2(\Omega)}+\|w_{n+1}(s)\|^2_{H_0^1(\Omega)}\\ &\quad=-2\varepsilon\int_0^s\langle[F_n(t, u_{n+1})-F_{n-1}(t, u_n)], \psi_{n+1}\rangle_{L^2(\Omega)}\,dt\\ &\quad=-2\varepsilon\int_0^s\left\langle\frac{d_n^2}{2}f''(q_n)+f'(u_n)d_{n+1}, \psi_{n+1}\right\rangle_{L^2(\Omega)}dt \end{aligned}$$

where $q_n=\lambda u_n+(1-\lambda)u_{n-1}$ $(0<\lambda<1)$.

But using the continuous imbedding of $H_0^1(\Omega)$ in $\mathscr{C}^0(\bar\Omega)$ and (27), we have

$$\tag{61} \left|2\varepsilon\int_0^s\langle d_{n+1}f'(u_n), \psi_{n+1}\rangle_{L^2(\Omega)}\right|\leqq 2\varepsilon T^2C_1\|d_{n+1}\|^2_{L^\infty(0,T;L^2(\Omega))}$$

where C_1 is a constant independent of n and ε.

Writing $\psi_{n+1}(t)=w_{n+1}(t)-w_{n+1}(s)$, we have

$$\tag{62} |\psi_{n+1}(t)|\leqq|w_{n+1}(t)|+|w_{n+1}(s)| \qquad (t\leqq s).$$

This enables us to evaluate

$$\left|-2\varepsilon\int_0^s\left\langle\frac{d_n^2}{2}f''(q_n), \psi_{n+1}\right\rangle_{L^2(\Omega)}dt\right|.$$

On account of the continuous imbedding of $H_0^1(\Omega)$ in $\mathscr{C}^0(\bar{\Omega})$:

$$\left|-2\varepsilon\int_0^s\left\langle\frac{d_n^2}{2}f''(q_n),\,\psi_{n+1}\right\rangle_{L^2(\Omega)}dt\right|\leqq 2\varepsilon C_2T\|w_{n+1}\|_{L^\infty(0,T;H_0^1(\Omega))}\|d_n\|^2_{L^\infty(0,T;L^2(\Omega))}\tag{63}$$

where C_2 is a constant independent of n and ε.

Finally by (60), (61) and (63) we can find for a suitable $T>0$ a number $\alpha>0$ such that

$$\begin{aligned}\alpha^2\|d_{n+1}\|^2_{L^\infty(0,T;L^2(\Omega))}&+\|w_{n+1}\|^2_{L^\infty(0,T;H_0^1(\Omega))}\\&\leqq 2T\varepsilon C_2\|w_{n+1}\|_{L^\infty(0,T;H_0^1(\Omega))}\|d_n\|^2_{L^\infty(0,T;L^2(\Omega))}.\end{aligned}\tag{64}$$

But the inequality $2ab\leqq a^2+b^2$ implies that

$$\|d_{n+1}\|_{L^\infty(0,T;L^2(\Omega))}\leqq k_T^{(1)}\|d_n\|^2_{L^\infty(0,T;L^2(\Omega))}\tag{65}$$

with $k_T^{(1)}=TC_2\varepsilon/\alpha$.

Therefore, for a suitable $T>0$, the convergence is quadratic even if ε is not necessarily small. We then have the following result:

THEOREM 6. *Problem* (1)-(2) *with assumptions* (3)(ii), (3') *and* (38) *has one and only one solution* $u\in L^\infty(0,T;H_0^1(\Omega))$. *The sequence* $\{u_n\}$ *defined by* (58) *converges quadratically to it.*

4. An application to the numerical solution of nonlinear partial differential equations. The Tau Method is a perturbation technique based on the ideas of best uniform approximation by polynomials. Given a linear differential equation with polynomial coefficients, or with coefficients approximated by polynomials to a sufficient degree of accuracy,

$$L(u)=F(x,t),\tag{66}$$

we attempt to solve a slightly perturbed form of the original problem, defined by the so-called *Tau problem*

$$L(u_{rs})=F(x,t)+\tau H_{rs}(x,t),\tag{67}$$

where $H_{rs}(x,t)$ is the product (or linear combination of products) of best uniform approximations of zero, of degrees r and s respectively, on a given domain D. The parameter (or vector parameter) τ is chosen for $u_{rs}(x,t)$ to be a bivariate polynomial which satisfies the initial or boundary conditions given for u.

The theory of the Tau Method, originally proposed by Lanczos in the late thirties, has been developed by Ortiz [8] and computational procedures for the numerical treatment of linear and nonlinear partial differential equations have been discussed by Ortiz and Samara [10] and Ortiz and Pun [11], [12] in very recent papers and in references given there.

Ortiz and Pham Ngoc Dinh have discussed aspects of the error analysis of the Tau Method in connection with nonlinear ordinary differential equations in [9]. Their approach is, essentially, to reduce the nonlinear problem to a sequence of linear problems for equations with variable coefficients.

In this section we shall discuss the numerical solution of two types of nonlinear partial differential equations by making use of the linear recursion schemes proposed in the early parts of this paper. As we shall see, results of a very remarkable accuracy are obtained.

Problem 1. Let us consider the nonlinear hyperbolic problem

$$\begin{aligned}&\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x^2}=u^2(x,t)+F(x,t), \qquad (x,t)\in D,\\ &u(x,0)=\left.\frac{\partial u(x,t)}{\partial t}\right|_{t=0}=g(x), \qquad 0\leqq x\leqq 1,\\ &u(0,t)=u(1,t)=h(t), \qquad 0\leqq t\leqq 1,\end{aligned} \tag{68}$$

where

$$F(x,t)=\exp(t)(1+\pi^2-\exp(t)\sin\pi x)\sin\pi x;$$

$$g(x)=\sin\pi x, \qquad h(t)\equiv 0,$$

and the domain $D=\{(x,t)\in\mathbb{R}^2\colon 0\leqq t, x\leqq 1\}$.

The exact solution of (68) is $u(x,t)=\exp(t)\sin\pi x$. We use the linear recursive scheme defined by

$$\frac{\partial^2 u_{n+1}}{\partial t^2}-\frac{\partial^2 u_{n+1}}{\partial x^2}-2u_n u_{n+1}=-u_n^2+F, \tag{69}$$

with $u_0=(1+t)\sin\pi x$, and u_{n+1} satisfying the conditions given in (68), but for the fact that functions F and g have been replaced by tight polynomial approximations. After 3 iterations the uniform norm of the error of approximation over the domain D becomes stable for $r=s=6$, 8 and 10. It is equal to 0.1×10^{-3}, 0.9×10^{-6} and 0.4×10^{-8} respectively.

Problem 2. Let us consider the nonlinear parabolic problem

$$\begin{aligned}&\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=u^2(x,t)+F(x,t), \qquad (x,t)\in D,\\ &u(x,0)=g(x), \quad 0\leqq x\leqq 1, \qquad u(0,t)=u(1,t)=h(t), \quad 0\leqq t\leqq 1,\end{aligned} \tag{70}$$

where F, g, and h are the same functions as in Problem 1. The domain D is also the same and the exact solution of (70) is identical to that of (68). We asssociate with (70) the linear recursive scheme defined by

$$\frac{\partial u_{n+1}}{\partial t}-\frac{\partial^2 u_{n+1}}{\partial x^2}-2u_n u_{n+1}=-u_n^2+F, \tag{71}$$

with $u_0=\sin\pi x$, and u_{n+1} satisfying the same conditions as (70), but with F and g approximated by polynomials. Applying the Tau Method to the numerical solution of the problems defined by the linear recursive scheme (71), we find that for $r=s=6$ and 8 the uniform norm of the error of approximation over D becomes stable after 3 iterations, the error being equal to 0.7×10^{-4} and 0.8×10^{-6} respectively. For $r=s=10$ the error becomes stable after four iterations, when it reaches 0.4×10^{-8}.

Acknowledgment. The authors wish to thank the referee of this paper for his constructive and useful remarks.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] R. BELLMAN, *Functional equations in the theory of dynamic programming*, (V), *part* V: *Positivity and quasilinearity*, Proc. Nat. Acad. Sci. U.S.A., 41 (1955), pp. 743-746.

[3] J. BOUJOT, A. PHAM NGOC DINH AND J. P. VEYRIER, *Oscillateurs harmoniques faiblement perturbés: l'algorithme des "pas de géant"*, RAIRO Analyse Numérique, 14 (1980), pp. 3-23.

[4] R. KALABA, *On nonlinear differential equations, the maximum operation and monotone convergence*, J. Math. Mech., 6 (1959), pp. 519-574.

[5] V. LAKSHMITKANTHAM AND S. LEELA, *Differential and Integral Inequalities*, Vol. I, Academic Press, New York, 1969.

[6] J. L. LIONS, *Équations différentielles opérationelles et problèmes aux limites*, Springer-Verlag, Berlin, 1961.

[7] ———, *Quelques méthodes de résolution des problèmes aux limites nonlinéaires*, Dunod, Gauthier-Villars, 1969.

[8] E. L. ORTIZ, *The Tau Method*, SIAM J. Numer. Anal., 6 (1969), pp. 480-492.

[9] E. L. ORTIZ AND A. PHAM NGOC DINH, *On the convergence of the Tau Method for nonlinear differential equations of Riccati's type*, Nonlinear Anal., 9 (1985), pp. 53-60.

[10] E. L. ORTIZ AND H. SAMARA, *Numerical solution of partial differential equations with variable coefficients with an operational approach to the Tau Method*, Comp. Math. Appls., 10 (1984), pp. 5-13.

[11] E. L. ORTIZ AND K.-S. PUN, *Numerical solution of nonlinear partial differential equations with the Tau Method*, J. Comput. Appl. Math., 12 & 13 (1985), pp. 511-516.

[12] ———, *Numerical solution of Burger's nonlinear partial differential equation with a bidimensional segmented formulation of the Tau Method*, in Computers and Mathematics with Applications, 1986, in press.

[13] A. PHAM NGOC DINH, *Sur un problème hyperbolique faiblement non linéaire en dimension* 1, Demonstratio Mathematicae, 16 (1983), pp. 269-289.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

015

RECURSION RELATIONS FOR SOLUTIONS TO THE SCHRÖDINGER EQUATION*

MICHAEL REACH†

Abstract. We will consider the general eigenfunction for a second-order differential operator in one variable. For many well-known elementary functions, we can also find a three-term recursion relation in the eigenvalue parameter. For practical computation, this is a very desirable property. Examples include Legendre functions, Bessel functions, etc.

In [Math. Z., 29 (1929), pp. 730-736], Bochner showed that the only polynomial solutions to this problem were the well-known ones. This paper will look for solutions that may not be polynomials.

It will be shown that, unfortunately, for the simple recursion relation considered here, no really new examples exist.

Key words. recursion relations, Bochner, elementary functions

AMS(MOS) subject classifications. 34, 39

1. Notation. Let L be a second-order differential operator in one variable x,

$$L\phi(x) \equiv g_1(x)\frac{d^2\phi}{dx^2} + g_2(x)\frac{d\phi}{dx} + g_3(x)\phi, \tag{1}$$

and let B be a second-order difference operator in k, acting on a sequence $\{l_k\}$ by

$$(Bl)_k \equiv a_k l_{k+1} + b_k l_k + c_k l_{k-1}. \tag{2}$$

For later convenience, subscripts like k will range over $\mathbf{Z}+c$ for some fixed $c \in \mathbf{C}$; that is, over all complex numbers differing from a given one by an integer. Constants will normally be complex; functions will be: $\mathbf{C} \to \mathbf{C}$ and as smooth as necessary.

2. Basic problem. We seek a sequence of functions $\{\phi_k(x)\}$ s.t.

$$L\phi_k(x) = \lambda_k \phi_k(x) \tag*{(3)$_k$}$$

and

$$(B\phi(x))_k = a_k\phi_{k+1}(x) + b_k\phi_k(x) + c_k\phi_{k-1}(x) = \theta(x)\phi_k(x) \tag*{(4)$_k$}$$

for all $k \in \mathbf{Z}+c$ and some L, $\{a_k, b_k, c_k\}$, $\{\lambda_k\}$, $\theta(x)$. Of course, we really seek the whole sextuple $(L, B, \lambda, \theta, \phi, c)$.

We can allow some (not all) of the ϕ_k's to be zero. This will enable us to include the classical orthogonal polynomials. We will see in § 5, though, that all ϕ_k's must be nonzero, either for each k big enough, or else for each k small enough. $\theta(x)$ should be nonconstant in x, and λ_k nonconstant in k. Note that L and θ are independent of k, and that B and λ are independent of x. This will allow us to commute them in several formulas.

3. Symmetries. Given a solution to our Basic Problem, many obvious changes yield new sextuples which will also be solutions. Nine of these changes will be listed,

* Received by the editors December 17, 1984, and in revised form August 30, 1985. This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy under contract DE-AC03-76SF00098.

† Department of Mathematics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720.

to be used extensively later. Assume, then, that $c' \in \mathbf{C}$ is an arbitrary constant, and that $(L, B, \lambda, \theta, \phi, c)$ solves the Basic Problem. The following is a list of new solution sextuples:

(a)

(S1) $$(c'L, B, c'\lambda, \theta, \phi, c).$$

This scales L by a constant. Since $(3)_k$ implies $(c'L)\phi_k = (c'\lambda_k)\phi_k$, (S1) is also a solution.

(b)

(S2) $$(L, B, \lambda, \theta, c'\phi, c).$$

ϕ is scaled by a constant here. (S2) clearly follows from multiplying $(3)_k$ and $(4)_k$ by c'.

(c)

(S3) $$(L+c', B, \lambda+c', \theta, \phi, c).$$

(S3) transfers a constant between L and λ_k. We see it by modifying $(3)_k$ to be $(L+c')\phi_k = (\lambda_k + c')\phi_k$.

(d)

(S4) $$(L, B, \lambda, \theta, \phi, c-c').$$

This new solution just has a shifted subscript. The $(3)_k$ corresponding to it is now $L\phi_{k+c'} = \lambda_{k+c'}\phi_{k+c'}$ for $k \in \mathbf{Z} + c - c'$ and the new $(4)_k$ is similar.

(e)

(S5) $$(L, c'B, \lambda, c'\theta, \phi, c).$$

(S5) scales the recursion relation by a constant. The new $(4)_k$ will be $((c'B)\phi)_k = (c'\theta(x))\phi_k$.

(f)

(S6) $$(L, B+c', \lambda, \theta+c', \phi, c).$$

In (S6), a constant is transferred between B and θ, so that $(4)_k$ is now $(B\phi)_k + c'\phi_k = (\theta(x)+c')\phi_k$.

(g) Symmetry seven is under the change of independent variable x, say by $x = x(x')$. If L' is the new differential operator resulting from the variable change, we derive the new solution (S7):

(S7) $$(L', B, \lambda, \theta(x(x')), \phi(x(x')), c).$$

Equations $(3)_k$ and $(4)_k$ change in the obvious ways. We will use this transformation most commonly for the simple shift $x = x' + c'$.

(h) Let $\phi_k(x) = f(x)\psi_k(x)$, for some function $f(x)$ and all k. Then

(S8) $$(f^{-1}Lf, B, \lambda, \theta, \psi(x), c)$$

is a solution allowing us to multiply ϕ by an arbitrary function of x. Equation $(4)_k$ will look similar, with ψ in place of ϕ, and $(3)_k$ will be $(f^{-1}Lf)\psi_k(x) = \lambda_k\psi_k(x)$.

(i) Let $\phi_k(x) = f_k\psi_k(x)$, for all k and for f_k independent of x. This multiplication of ϕ_k by an arbitrary constant depending only on k gives the new solution

(S9) $$(L, f^{-1}Bf, \lambda, \theta, \psi(x), c).$$

Here $(3)_k$ will be similar, and $(4)_k$ becomes $f_k^{-1}(B(f\psi))_k = \theta(x)\psi_k(x)$.

We will be using these to transfer to simpler-looking solutions, and eventually to reduce to a few fundamental ones. Clearly, the space of possible variations of these fundamental solutions will be extremely large.

4. Commutation relations. L is second order, and we can think of the operator of multiplying by $\theta(x)$ as zeroth order. Then

$$\begin{aligned} &[\theta, L] \equiv \theta L - L\theta && \text{is first order,}\\ &[\theta, [\theta, L]] && \text{is zeroth order,}\\ &[\theta, [\theta, [\theta, L]]] && = 0. \end{aligned}$$

Thus

$$[\theta, [\theta, [\theta, L]]]\phi_k = 0.$$

A typical term of this last expression is

$$\begin{aligned} -3\theta\theta L\theta\phi_k &= -3\theta\theta L(B\phi)_k && \text{by } (4)_k,\\ &= -3\theta\theta(B(L\phi))_k && \text{as } L \text{ and } B \text{ commute,}\\ &= -3\theta\theta(B(\lambda\phi))_k && \text{by } (3)_k,\\ &= -3(B\lambda\theta\theta\phi)_k && \text{since } B \text{ and } \theta \text{ also commute,}\\ &= -3(B\lambda BB\phi)_k && \text{once again using } (4)_k. \end{aligned}$$

Continuing in this way, we soon see that

$$0 = [\theta, [\theta, [\theta, L]]]\phi_k = ([B, [B, [B, \lambda]]]\phi)_k.$$

This last expression is a linear combination of seven terms, multiplying $\phi_{k-3}, \phi_{k-2}, \cdots, \phi_{k+3}$, respectively. If we assume the ϕ_k's linearly independent for different k's (as, for example, if $\lambda_k \neq \lambda_{k'}$, for $k \neq k'$), each of the seven terms must vanish. In particular, let us examine the terms containing ϕ_{k+3} and ϕ_{k-3}, which each must be zero. The term with ϕ_{k+3} is $a_{k+2}a_{k+1}a_k(\lambda_{k+3} - 3\lambda_{k+2} + 3\lambda_{k+1} - \lambda_k)\phi_{k+3}$. The term with ϕ_{k-3} is $c_kc_{k-1}c_{k-2}(\lambda_k - 3\lambda_{k-1} + 3\lambda_{k-2} - \lambda_{k-3})\phi_{k-3}$. These must be zero for all $k \in \mathbf{Z} + c$. Thus, for a given k,

$$(5) \qquad \begin{aligned} &\text{either} && (a_ka_{k+1}a_{k+2}\phi_{k+3} = c_{k+1}c_{k+2}c_{k+3}\phi_k = 0),\\ &\text{or} && \lambda_k - 3\lambda_{k+1} + 3\lambda_{k+2} - \lambda_{k+3} = 0. \end{aligned}$$

5. Pinning down λ.

THEOREM. *At least one of the following statements is true:*

(6)
(1) $c_k \neq 0$ *and* $\phi_k \neq 0$ *for all k sufficiently large negative, or*
(2) $a_k \neq 0$ *and* $\phi_k \neq 0$ *for all k sufficiently large positive.*

(*Recall that the a's and c's are the coefficients in* (4).)

Proof. Pick some $\phi_l \neq 0$. If, for all $k < l$, c_k and $\phi_k \neq 0$, the theorem is proven. Otherwise, there is some largest $m \leqq l$ such that

$$c_m \cdot \phi_{m-1} = 0.$$

Since this m is maximal, $\phi_m \neq 0$. Then $(4)_m$ becomes

$$0 + (b_m - \theta(x))\phi_m + a_m\phi_{m+1} = 0.$$

Thus $a_m\phi_{m+1} \neq 0$ and $\phi_{m+1} = \phi_m \cdot$ (polynomial in $\theta(x)$, degree 1). Now $(4)_{m+1}$ implies

$$\phi_m \cdot (\text{polynomial in } \theta(x), \text{degree } 2) + a_{m+1}\phi_{m+2} = 0.$$

Therefore $a_{m+1}\phi_{m+2} \neq 0$ and $\phi_{m+2} = \phi_m \cdot$ (polynomial in θ, degree 2). We continue this way to generate the ϕ_k's for $k > m$. Since θ is not constant, the nth degree polynomial in θ at the nth step cannot be zero. At the nth step we'll get $a_{m+n}\phi_{m+n+1} \neq 0$, so that we've shown $a_n \neq 0$ and $\phi_{n+1} \neq 0$ for all $n \geqq k$.

We will write $k \in K$ as a synonym for "k sufficiently large negative" or "k sufficiently large positive," respectively. It will turn out in § 7 that all we shall need is that the theorem be true for many contiguous k's. This explains why we can be so ambiguous about the finite end-point of K.

From this theorem and (5) we get that

$$\lambda_k - 3\lambda_{k+1} + 3\lambda_{k+2} - \lambda_{k+3} = 0 \quad \text{for } k \in K.$$

Therefore, $\lambda_k = r_1 k^2 + r_2 k + r_3$ for $k \in K$; r_1, r_2, r_3 fixed. $\lambda_k \neq$ constant, so $r_1 \neq 0$ or $r_2 \neq 0$.

If $r_1 \neq 0$, use (S1) to scale λ_k so that the leading coefficient is 1. Then (S4) can shift k and kill the linear term. Finally, (S3) transfers the constant term into L, yielding $\lambda_k = k^2$.

If $r_1 = 0$, $r_2 \neq 0$, (S1) can be used to scale the leading coefficient of λ_k to be 1. (S4) will kill the constant term by shifting k, giving $\lambda_k = k$. We can keep (S3) in reserve this time for possible need later.

We have reduced to either of two cases

$$\lambda_k = k \quad \text{or} \quad \lambda_k = k^2 \quad \text{for } k \in K. \tag{7}$$

6. Some convenient formulas.

LEMMA.

$$(L - \lambda_k)^n(\theta\phi_k) = \underbrace{[L, [L, \cdots, [L, \theta]] \cdots]}_{\leftarrow n\ L'\ s \rightarrow}\phi_k, \qquad n \geqq 0. \tag{8}$$

Proof. For $n = 1$, the proof is

$$(L - \lambda_k)\theta\phi_k = (L\theta - \theta\lambda_k)\phi_k = (L\theta - \theta L)\phi_k.$$

Larger n's can be shown similarly by induction.

LEMMA.

$$(L - \lambda_{k+1})(L - \lambda_k)(L - \lambda_{k-1})(\theta\phi_k) = 0.$$

Proof. $\theta\phi_k = (B\phi)_k =$ a linear combination of ϕ_{k-1}, ϕ_k, ϕ_{k+1}. Call $\Delta_k \equiv \lambda_k - \lambda_{k+1}$, $\nabla_k \equiv \lambda_k - \lambda_{k-1}$. Then from the previous lemma,

$$\begin{aligned} 0 &= (L - \lambda_{k+1})(L - \lambda_k)(L - \lambda_{k-1})(\theta\phi_k) \\ &= (L - \lambda_k + \Delta_k)(L - \lambda_k)(L - \lambda_k + \nabla_k)(\theta\phi_k) \\ &= ((L - \lambda_k)^3 + (\Delta_k + \nabla_k)(L - \lambda_k)^2 + \Delta_k\nabla_k(L - \lambda_k))(\theta\phi_k). \\ \text{By (8),} \quad &= ([L, [L, [L, \theta]]] + (\Delta_k + \nabla_k)[L, [L, \theta]] + \Delta_k\nabla_k[L, \theta])\phi_k = 0. \end{aligned}$$

Call

$$A_n \equiv \text{the operator } \underbrace{[L, [L, \cdots [L, \theta]] \cdots]}_{\longleftarrow n\ L'\ s \longrightarrow}.$$

Then

$$A_3\phi_k + (\Delta_k + \nabla_k)A_2\phi_k + (\Delta_k\nabla_k)A_1\phi_k = 0 \tag{9}$$

which is often a convenient form.

In [3], Grünbaum and Duistermaat attack the problem of finding a differential equation (instead of a recursion relation) in the spectral parameter, and arrive at a simpler but similar formula.

7. Finding possible potentials. We need not work with the general second-order operator for L. Using (S7) to set the leading coefficient to 1, and (S8) to kill the d/dx term, we can use

$$L=\frac{d^2}{dx^2}+V(x), \quad \text{without loss of generality.}$$

Hand calculations show (use $D=d/dx$)

$$A_1=2\theta'D+\theta'', \qquad A_2=4\theta''D^2+4\theta'''D+\theta''''-2\theta'V',$$

$$A_3=8\theta'''D^3+12\theta''''D^2+(6\theta^{(5)}-4\theta'V''-12\theta''V')D+(\theta^{(6)}-6\theta'''V'-8\theta''V''-2\theta'V''').$$

Case I. $\lambda_k=k$, $k\in K$ (from (7)). Here $\Delta_k=-1$, $\nabla_k=+1$. Then by (9)

$$(A_3-A_1)\phi_k=0 \quad \text{for } k\in K.$$

Since the finite order operator A_3-A_1 has here infinitely many independent solutions (by the theorem, $\phi_k\neq 0$ for all $k\in K$), it must be identically zero:

$$A_3-A_1\equiv 0. \tag{10}$$

We will use this argument again. Since we only need an infinite number of solutions, the sloppy definition of K in § 5 is sufficient. Take (10), and equate to zero each coefficient of a power of D.

D^3: $8\theta'''=0\Rightarrow\theta=r_1x^2+r_2x+r_3$ for some fixed r_1, r_2, r_3.

D^2: $12\theta''''=0\Rightarrow$nothing new.

D^1: $6\theta^{(5)}-4\theta'V''-12\theta''V'=2\theta'$. Using the D^1 equation$\Rightarrow -4(2r_1x+r_2)V''-24r_1V'=2(2r_1x+r_2)$.

D^0: nothing new.

There are now two subcases.

(a) $r_1=0$, so $-4r_2V''=2r_2$. $\theta\neq$ constant, so r_2 cannot be zero. Then

$$V=-\tfrac{1}{4}x^2+q_1x+q_2 \quad \text{for some fixed } q_1, q_2,$$

$$\theta=r_2x+r_3.$$

Use (S7) to shift $x\to x+2q_1$, and (S3) + (S4) to move a constant into λ_k, giving $V=-\frac{1}{4}x^2-\frac{1}{2}$ ($-\frac{1}{2}$ for later convenience). Use (S5), (S6) giving $\theta=x$.

(b) $r_1\neq 0$. Using (S7), (S5), (S6) yields $\theta=x^2$ and $-8r_1V''-24r_1V'=4r_1x$ for our two equations. The second one can be easily solved:

$$2xV''+6V'=-x, \quad (2x^3V')'=-x^3, \quad 2x^3V'=-\tfrac{1}{4}x^4-s_1,$$

$$V'=-\tfrac{1}{8}x-(s_1/2x^3), \qquad V=-\tfrac{1}{16}x^2+(s_1/x^2)+s_2.$$

Here s_1, s_2 are arbitrary constants. Thus Case I, $\lambda_k=k$, has two possible kinds of solution:

(a) $\theta=x$, $V=-\frac{1}{4}x^2-\frac{1}{2}$,

(b) $\theta=x^2$, $V=-\frac{1}{16}x^2+s_1/x^2+s_2$.

Solution to Case I. If $R_n(x)$ solves $y''-2xy'+2ny=0$ (Hermite equation), then $R_n((1/\sqrt{2})ix)\,e^{x^2/4}$ is an eigenfunction for (a).

For (b), pick an m s.t. $\frac{1}{4}-m^2=s_1$. Using (S3) and (S4) to move a constant into λ_k, we can get $V=-\frac{1}{16}x^2+(1-4m^2)/4x^2-(m+1)/2$. Then if $L_n^{(m)}(x)$ solves

$$xy''+(m+1-x)y'+ny=0 \tag{11}$$

(the Generalized Laguerre equation), one can check that $L_n^{(m)}(-x^2/4)\,e^{x^2/8}x^{m+1/2}$ is an eigenfunction for (b), eigenvalue n.

For the sake of completeness, the recursion relations for Hermite and Laguerre solutions are included (see [1, pp. 252, 241]): If R_n is a solution to the Hermite equation, then

$$R_{n+1}-2xR_n+2nR_{n-1}=0.$$

If $L_n^{(m)}$ solves the Generalized Laguerre equation, then

$$(n+1)L_{n+1}^{(m)}=(2n+m+1-x)L_n^{(m)}-(n+m)L_{n-1}^{(m)}.$$

These relations are well known. They do not work only for the orthogonal polynomials. Instead, for any element of the solution space for a given n, one can find elements of the solutions spaces for $n-1$ and $n+1$ such that the recursion relation will hold.

Case II. $\lambda_k=k^2$ for $k\in K$ (from (7)). In this case $\Delta_k=-2k-1$, $\nabla_k=2k-1$. Then by (9)

$$(A_3-2A_2+(1-4k^2)A_1)\phi_k=0 \quad \text{for } k\in K.$$

Here $k^2\phi_k=\lambda_k\phi_k=L\phi_k$, so $(A_3-2A_2+A_1(1-4L))\phi_k=0$ for $k\in K$. Again, since this operator has too many solutions,

$$A_3-2A_2+A_1(1-4L)=0.$$

We shall use $L=D^2+V(x)$ again, and set coefficients to zero. Only the D^3 and D^1 coefficients give new data.

D^3: $8\theta'''-8\theta'=0\Rightarrow\theta=r_1e^x+r_2e^{-x}+r_3$, r_1, r_2, r_3 constant.

Using (S7) to shift x, and (S6), (S5), we get to

$$\theta=e^x \quad \text{or} \quad \theta=\cosh x. \tag{12}$$

D^1: $6\theta^{(5)}-12\theta''V'-4\theta'V''-8\theta'''+2\theta'-8\theta'V=0$.

From (12), we can replace θ'' by θ, so

$$-12\theta V'-4\theta'V''-8\theta'V'=0.$$

Now a little algebra:

$$(\theta'V'+2\theta V)'=0,$$

$$(\theta'V'+2\theta V)=s_1,$$

$$((\theta')^2V)'=s_1\theta',$$

$$(\theta')^2V=s_1\theta+s_2,$$

$$V=\frac{s_1\theta+s_2}{(\theta')^2}. \tag{13}$$

We can list the possible solutions to (12) and (13):

(a) $\theta=e^x$, $V=e^{-2x}$ ($s_1=0$, use (S7) to shift x),

(b) $\theta=e^x$, $V=e^{-x}$ ($s_2=0$, use (S7) to shift x),

(c) $\theta = e^{x}$, $V = -\nu\omega\, e^{-x} - \omega^{2}\, e^{-2x}$ $(s_1 = -\nu\omega,\ s_2 = -\omega^{2})$,

(d) $\theta = \cosh x$, $V = \dfrac{(\beta^{2} - \alpha^{2}) \cosh x + (1 - 4\alpha^{2})/16}{\sinh^{2} x}$

$$(s_1 = \beta^{2} - \alpha^{2},\ s_2 = \tfrac{1}{16}(1 - 4\alpha^{2}))$$

where we have set the constants for later convenience.

Solutions to Case II.

(a) For $V = e^{-2x}$: If $C_n(x)$ solves

$$x^{2}y'' + y' + (x^{2} - n^{2})y = 0 \quad \text{(Bessel equation)}$$

then $C_n(e^{-x})$ is an eigenfunction for (a), eigenvalue n^{2}.

(b) $V = e^{-x}$: $C_{2n}(2e^{-x/2})$ is eigenfunction, eigenvalue n^{2}.

(c) $V = -\nu\omega\, e^{-x} - \omega^{2}\, e^{-2x}$. If $L_n^{(m)}(x)$ is a solution of (11), the generalized Laguerre equation, for given n and m, then

$$z^{n}\, e^{-z/2} L^{(2n)}_{-(v+1)/2-n}(z) \qquad (\text{with } z = 2\omega\, e^{-x})$$

is an eigenfunction to (c) with eigenvalue n^{2}.

(d) $$V = \frac{(\beta^{2} - \alpha^{2}) \cosh x + (1 - 4\alpha^{2})/16}{\sinh^{2} x}, \qquad \alpha, \beta \text{ constants.}$$

If $P_n^{\alpha\beta}(x)$ solves

$$(1 - x^{2})(P_n^{\alpha\beta})'' + [\beta - \alpha - (\alpha + \beta + 2)x](P_n^{\alpha\beta})' + n(\alpha + \beta + n + 1)P_n^{\alpha\beta} = 0$$

(Jacobi equation) then

$$\left(\sinh \frac{x}{2}\right)^{\alpha+1/2} \left(\cosh \frac{x}{2}\right)^{\beta+1/2} P^{\alpha\beta}_{n-(\alpha+\beta+1)/2}(\cosh x)$$

is an eigenfunction to (d) with eigenvalue n^{2}. (See [1, p. 214].)

For Bessel functions $C_n(x)$, the recursion relation is (see [1, p. 67]):

$$C_{n-1} + C_{n+1} = \frac{2n}{x} C_n.$$

This relation can be used to derive a recursion relation also for the Bessel functions that solve (b). For the recursion relation for the Jacobi functions P, see [1, p. 213]. The relation for the Generalized Laguerre function in (c) is not obvious. However, the following relation

$$\frac{(n+1)(n-m)}{(m+1)(m+2)} L^{(m)}_{n-m}(y) + \left(\frac{2n - m + 1}{(m+1)(m+3)} - \frac{1}{y}\right) y L^{(m+2)}_{n-m-1}(y)$$
$$+ \frac{1}{(m+2)(m+3)} y^{2} L^{(m+4)}_{n-m-2}(y) = 0$$

is sufficient, and can be derived by repeated use of

$$L_n^{(m)} = L_n^{(m+1)} - L_{n-1}^{(m+1)}$$

and

$$yL_{n-2}^{(m+2)} - (m + 1 - y)L_{n-1}^{(m+1)} + nL_n^{(m)} = 0$$

found in [1, pp. 241-2]. Here we use the standard normalizations for the $L_n^{(m)}$'s.

8. Conclusion. All the possible solutions found correspond to shifts, using the symmetries from § 3, of well-known functions; i.e., solutions to the Bessel, generalized Laguerre, Hermite and Jacobi differential equations. Note again that those solutions need not be the standard orthogonal polynomials, but can come from the full solution spaces. Since these functions are already well known, no attempt was made to catalogue them precisely: Some are special cases of others.

9. Final note. Though the results given here are hardly surprising, hardly anything is known about cases with higher-order L's or B's. For instance, the problem of finding the functions satisfying both a Schrödinger equation and also a *five* term recurrence, is, I think, completely uncharted territory. Some of the only nontrivial examples I have seen of this kind are given by Grünbaum in [4]. The examples he gives are of soliton-like functions. Though certain steps may be harder, many of these techniques should work for higher-order L's or B's.

REFERENCES

[1] W. MAGNUS, F. OBERHETTINGER AND R. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[2] S. BOCHNER, *Über Sturm-Louivillesche Polynomosysteme*, Math. Z., 29 (1929), pp. 730–736.

[3] J. J. DUISTERMAAT AND F. A. GRÜNBAUM, *Differential equations in the spectral parameter*, Comm. Math. Phys., to appear.

[4] F. A. GRÜNBAUM, *Recursion relations and a class of isospectral manifolds for Schrödinger's equation*, in Advances in Nonlinear Waves, Vol. I, E. Debnath, ed., Pitman, London, 1984, pp. 226–229.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

016

INFINITE SUMS IN THE THEORY OF DISPERSION OF CHEMICALLY REACTIVE SOLUTE*

A. E. DEGANCE† AND L. E. JOHNS‡

Abstract. We use the Cauchy partial fraction expansion to derive formulae for infinite sums arising in dispersion theory; the method is general. If $\{\lambda_j\}$ is the sequence of zeros of $f(\lambda)=0$ and if for some $g(\lambda)$

$$\frac{g(\lambda)}{f(\lambda)}=\sum_j \frac{1}{(\lambda-\lambda_j)}\frac{g(\lambda_j)}{f'(\lambda_j)},$$

then

$$\left.\frac{d^n}{d\lambda^n}\frac{\lambda-\lambda_k}{f(\lambda)}g(\lambda)\right|_{\lambda=\lambda_k}=(-1)^n n!\sum_{j\neq k}\frac{1}{(\lambda_k-\lambda_j)^n}\frac{g(\lambda_j)}{f'(\lambda_j)}.$$

Key words. infinite sums, Cauchy partial fraction expansion, dispersion theory

AMS(MOS) subject classifications. Primary 40C15; secondary 30B50

1. The problem. In our study of the dispersion of a chemically reactive solute in a cylinder of circular cross section we turn up infinite sums. The physical process is this: a cloud of solute is released into a solvent, which is in rectilinear flow inside a cylinder on which a chemical rearrangement of the solute takes place. The flow, being nonuniform over the cross section of the cylinder, distorts the cloud in the longitudinal direction; transverse diffusion opposes this; surface reaction moderates the influence of the streamlines near the cylinder. As time passes, the dispersion process is increasingly accurately represented by a constant coefficient dispersion equation in which the dispersion coefficients $X_{2\infty}$, $X_{3\infty}, \cdots$ turn out to be infinite sums, viz.

$$X_{2\infty}=D+\sum_{j\neq 1}\frac{\langle\psi_1, v\psi_j\rangle\langle\psi_j, v\psi_1\rangle}{\lambda_j^2-\lambda_1^2},$$

$$X_{3\infty}=\sum_{j\neq 1}\sum_{k\neq 1}\frac{\langle\psi_1, v\psi_j\rangle\langle\psi_j, v\psi_k\rangle\langle\psi_k, v\psi_1\rangle}{(\lambda_j^2-\lambda_1^2)(\lambda_k^2-\lambda_1^2)}$$

$$\cdots$$

(cf. DeGance and Johns [5]). The eigenfunctions, ψ_j, and the eigenvalues, $-\lambda_j^2$, satisfy

$$D\nabla^2\psi_j=-\lambda_j^2\psi_j,\quad (x,y)\in A,\quad D>0,$$

$$-D\mathbf{n}\cdot\nabla\psi_j+K\psi_j=0,\quad (x,y)\in\partial A,\quad K\leqq 0$$

where A is the cross section of the cylinder, ∂A is its boundary and $\langle,\rangle$ denotes the plain vanilla inner product. Aris [2] reviews the literature of this problem.

For a cylinder of circular cross section, $X_{2\infty}$ simplifies to

$$X_{2\infty}=1+16N_{pe}^2\frac{\lambda_1^2}{\lambda_1^2+\beta^2}\sum_{j\neq 1}\frac{\lambda_j^2}{\lambda_j^2+\beta^2}\frac{(\lambda_j^2+\lambda_1^2+2\beta^2)^2}{(\lambda_j^2-\lambda_1^2)^5}$$

where $\beta=-K/D\geqq 0$. The sum on the right-hand side can be written

$$S_0^{3,1}(\beta)+4(\lambda_1^2+\beta^2)S_0^{4,1}(\beta)+4(\lambda_1^2+\beta^2)^2S_0^{5,1}(\beta)$$

* Received by the editors October 16, 1984; accepted for publication (in revised form) February 25, 1986.

† Amoco Production Company, P.O. Box 3385, Tulsa, Oklahoma 74102.

‡ Department of Chemical Engineering, University of Florida, Gainesville, Florida 32611.

and the problem then is to sum

$$S_0^{m,k}(\beta)=\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\frac{\lambda_j^2}{\lambda_j^2+\beta^2}\qquad \forall\beta\geqq 0,\quad m=1,2,\cdots$$

where $0\leqq\lambda_1<\lambda_2<\cdots$ satisfy

$$\lambda_jJ_1(\lambda_j)-\beta J_0(\lambda_j)=0.$$

We derive formulae that express $S_0^{m,k}(\beta)$ as a rational function of β and λ_k^2.

Both Euler and Rayleigh worked on the problem of establishing formulae for certain sums (cf. Watson [8, pp. 500-502]). Indeed, Rayleigh [7] evaluated the sums[1]

$$S_0^{m,1}(0)=\sum_{j\neq 1}\frac{1}{\lambda_j^{2m}},\qquad \beta=0,\quad m=1,2,\cdots$$

where $0=\lambda_1<\lambda_2<\cdots$ satisfy $J_1(\lambda_j)=0$. Much later, Ahmed and Muldoon [1] showed how to evaluate the sums

$$S^{m,k}(\beta)=\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\qquad \forall\beta\geqq 0,\quad m=1,2,\cdots$$

using the logarithmic derivative and its Cauchy partial fraction expansion. What we do is not unlike what Rayleigh does nor what Ahmed and Muldoon do; but it is not limited to the expansion of the logarithmic derivative. Indeed, various partial fraction expansions are required to do various classes of sums, but the justification of each partial fraction expansion requires some analysis and while it is general in form it is specific in detail. The general idea then is the use of the Cauchy partial fraction expansion to deduce formulae for certain infinite sums; we illustrate this in two concrete problems of interest to us: the dispersion of a single solute in circular tubes, cf. §§ 2 and 3, and in narrow slits, cf. § 5.

In § 3 we evaluate $S_0^{m,k}(\beta)$ and its companion $S_1^{m,k}(\beta)$ where

$$S_1^{m,k}(\beta)=\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\frac{1}{\lambda_j^2+\beta^2}.$$

Now because $\lambda_jJ_1(\lambda_j)-\beta J_0(\lambda_j)=0$ implies

$$\frac{d}{d\beta}\lambda_j^2=\frac{2\lambda_j^2}{\lambda_j^2+\beta^2},$$

$S_0^{m,k}(\beta)$ and $S_1^{m,k}(\beta)$ $m=2,\cdots$ can be deduced from the expansion of the logarithmic derivative, e.g.,

$$S_0^{m+1,k}(\beta)=\frac{\lambda_k^2}{\lambda_k^2+\beta^2}S^{m+1,k}(\beta)-\frac{1}{2m}\frac{d}{d\beta}S^{m,k}(\beta),\qquad m=1,2,\cdots.$$

Thus, to show that various partial fraction expansions are required to do various classes of sums, we investigate

$$\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\frac{\lambda_j^2}{\lambda_j^2+\beta^2}\frac{J_0(\alpha\lambda_j)}{J_0(\lambda_j)},\qquad 0\leqq\alpha\leqq 1,$$

[1] In fact Rayleigh evaluated $\sum(1/\lambda_j^{2m})$ where $\{\lambda_j\}$ is the sequence of positive roots of $J_\nu(\lambda)=0$, $\nu=0,1,\cdots$.

and

$$\sum_{j\neq k} \frac{1}{(\lambda_j^2-\lambda_k^2)^m} \frac{1}{\lambda_j^2+\beta^2} \frac{J_1(\alpha\lambda_j)}{J_1(\lambda_j)}, \qquad 0 \leqq \alpha \leqq 1,$$

which, like similar sums that arise in multisolute dispersion (cf. DeGance and Johns [6]) cannot be deduced from the expansion of the logarithmic derivative.

2. The Cauchy partial fraction expansion. For $\beta \geqq 0$, the zeros of $f(\lambda) = \lambda J_1(\lambda) - \beta J_0(\lambda)$ are real (cf. Lommel's theorem (Watson [8, p. 482])). We let λ_j, $j = 1, \cdots$, denote the nonnegative zeros ordered so that $\lambda_j < \lambda_{j+1}$; for $\beta = 0$, $\lambda_1 = 0$ is a double root and $\lambda_j > 0, j = 2, \cdots$ is a simple root; for $\beta > 0$, λ_j is a simple root. Then $f(\lambda)$ vanishes for $\lambda = \pm\lambda_j$ and nowhere else.

The residues of $g(\lambda)/f(\lambda)$ at $\pm\lambda_j$ for $g(\lambda) = f'(\lambda)$, $g(\lambda) = J_0(\lambda)$ and $g(\lambda) = J_1(\lambda)$ are 1, $\pm\lambda_j/(\lambda_j^2+\beta^2)$ and $\beta/(\lambda_j^2+\beta^2)$. Neither $J_0(\pm\lambda_j)$ nor $J_1(\pm\lambda_j)$ survives the calculation of the second and third residues. Their cancellation results because

$$\frac{d}{d\lambda}(\lambda J_1(\lambda) - \beta J_0(\lambda)) = \lambda J_0(\lambda) + \beta J_1(\lambda) \quad \text{and} \quad \pm\lambda_j J_1(\pm\lambda_j) = \beta J_0(\pm\lambda_j).$$

We find, using the Cauchy partial fraction expansion (cf. Copson [4]),

$$\frac{d/d\lambda(\lambda J_1(\lambda) - \beta J_0(\lambda))}{\lambda J_1(\lambda) - \beta J_0(\lambda)} = \sum_{j=1} \frac{2\lambda}{\lambda^2 - \lambda_j^2},$$

$$\frac{J_0(\lambda)}{\lambda J_1(\lambda) - \beta J_0(\lambda)} = \sum_{j=1} \frac{2\lambda_j^2}{(\lambda^2-\lambda_j^2)(\lambda_j^2+\beta^2)}$$

and

$$\frac{J_1(\lambda)}{\lambda J_1(\lambda) - \beta J_0(\lambda)} = \sum_{j=1} \frac{2\beta\lambda}{(\lambda^2-\lambda_j^2)(\lambda_j^2+\beta^2)}.$$

The first form of Cauchy's theorem implies the second and third expansions; the second form implies the first expansion (cf. Copson [4, pp. 144–148]).

We show that the four hypotheses of the first form of Cauchy's theorem are satisfied so that the second and third expansions obtain. Condition (i) is satisfied because $g(\lambda)/f(\lambda)$ is regular save for the zeros of $f(\lambda)$. If C_j denotes a square contour on the Argand diagram with vertices at $(\pm j\pi, \pm j\pi)$ then the sequence of contours, $\{C_j\}$, satisfies Condition (ii). Indeed Dixon's theorem (cf. Watson [8, p. 480]), implies that the zeros of $\lambda J_1(\lambda) - \beta J_0(\lambda)$ fall between the zeros of $J_1(\lambda)$ and $J_0(\lambda)$, i.e., $\lambda_j(0) < \lambda_j(\beta) < \lambda_j(\infty)$, $\forall \beta \in (0, \infty)$; this and Schafheitlin's result (cf. Watson [8, pp. 490–492]), viz.,

$$\lambda_j(0) \in (j\pi - \tfrac{7}{8}\pi, j\pi - \tfrac{3}{4}\pi), \qquad j = 2, \cdots$$

and

$$\lambda_j(\infty) \in (j\pi - \tfrac{1}{4}\pi, j\pi - \tfrac{1}{8}\pi), \qquad j = 1, \cdots$$

imply that $\lambda_j(\beta) \notin C_k$ $\forall j$, k and β. To see that Conditions (iii) and (iv) are also satisfied, we note the asymptotic formula $J_1(\lambda)/J_0(\lambda) \to \tan(\lambda - \frac{1}{4}\pi) + 1/(2\lambda)$ in the right half-plane (cf. Watson [8, p. 496]); this implies that in the left half-plane

$$\frac{J_1(\lambda)}{J_0(\lambda)} \to -\tan\left(-\lambda - \frac{1}{4}\pi\right) + \frac{1}{2\lambda} = -\cot\left(\lambda - \frac{1}{4}\pi\right) + \frac{1}{2\lambda}.$$

Then, because

$$\left|\tan\left(\lambda-\frac{1}{4}\pi\right)\right|^2=\frac{\cosh(2\operatorname{Im}\lambda)-\sin(2\operatorname{Re}\lambda)}{\cosh(2\operatorname{Im}\lambda)+\sin(2\operatorname{Re}\lambda)}$$

we find that on the horizontal sides of C_j, $\tanh(j\pi)\leqq|\tan(\lambda-\frac{1}{4}\pi)|\leqq\coth(j\pi)$, whereas on the vertical sides, $|\tan(\lambda-\frac{1}{4}\pi)|=1$. We conclude that $|J_1(\lambda)/J_0(\lambda)|\to 1$ on C_j as $j\to\infty$. This implies that

$$\left|\frac{J_0(\lambda)}{\lambda J_1(\lambda)-\beta J_0(\lambda)}\right| \quad\text{and}\quad \left|\frac{J_1(\lambda)}{\lambda J_1(\lambda)-\beta J_0(\lambda)}\right|$$

are bounded by $(|\lambda|-\beta)^{-1}\leqq(j\pi-\beta)^{-1}$ on C_j as $j\to\infty$ and hence vanish uniformly as $j\to\infty$. This establishes the second and third expansions. Likewise, we conclude that the first expansion obtains, inasmuch as $[(d/d\lambda)(\lambda J_1(\lambda)-\beta J_0(\lambda))]/(\lambda J_1(\lambda)-\beta J_0(\lambda))$ satisfies the hypotheses of the second form of Cauchy's theorem.

The first expansion leads to the infinite product expansion of $f(\lambda)$ and thence to $S^{m,k}(\beta)$, Ahmed and Muldoon's sum; the second and third lead to $S_0^{m,k}(\beta)$ and $S_1^{m,k}(\beta)$. Thus, if we let $F(z;\beta)=\sqrt{z}J_1(\sqrt{z})-\beta J_0(\sqrt{z})$ and write $F(z;\beta)=(z-\lambda_k^2)F_k(z;\beta)$, then the second expansion implies that

$$R_{0k}(z;\beta)\equiv\frac{J_0(\sqrt{z})}{F_k(z,\beta)}=\frac{2\lambda_k^2}{\lambda_k^2+\beta^2}+(z-\lambda_k^2)\sum_{j\neq k}\frac{2\lambda_j^2}{(z-\lambda_j^2)(\lambda_j^2+\beta^2)}$$

and hence that

$$\frac{d^n}{dz^n}R_{0k}=(-1)^{n-1}n!\sum_{j\neq k}\frac{2\lambda_j^2}{(z-\lambda_j^2)^n(\lambda_j^2+\beta^2)}$$

$$+(-1)^n n!(z-\lambda_k^2)\sum_{j\neq k}\frac{2\lambda_j^2}{(z-\lambda_j^2)^{n+1}(\lambda_j^2+\beta^2)},\qquad n=1,2,\cdots.$$

On setting $z=\lambda_k^2$ we find that

$$R_{0k}^{(n)}(\lambda_k^2;\beta)=(-1)^{n-1}n!\sum_{j\neq k}\frac{2\lambda_j^2}{(\lambda_k^2-\lambda_j^2)^n(\lambda_j^2+\beta^2)}$$

and hence that $S_0^{m,k}(\beta)=-R_{0k}^{(m)}(\lambda_k^2;\beta)/(2m!)$ so that the evaluation of $S_0^{m,k}(\beta)$ reduces to the evaluation of $R_{0k}^{(m)}(\lambda_k^2;\beta)$.

Likewise, the third expansion implies that

$$R_{1k}(z;\beta)\equiv\frac{J_1(\sqrt{z})}{\sqrt{z}F_k(z;\beta)}=\frac{2\beta}{\lambda_k^2+\beta^2}+(z-\lambda_k^2)\sum_{j\neq k}\frac{2\beta}{(z-\lambda_j^2)(\lambda_j^2+\beta^2)};$$

hence a calculation shows that $S_1^{m,k}(\beta)=-R_{1k}^{(m)}(\lambda_k^2;\beta)/(2\beta m!)$ so that the evaluation of $S_1^{m,k}(\beta)$ reduces to the evaluation of $R_{1k}^{(m)}(\lambda_k^2;\beta)$.

3. The evaluation of $R_{0k}^{(m)}(\lambda_k^2;\beta)$. We observe that

$$R_{0k}^{(1)}(z;\beta)=\frac{(d/dz)J_0(\sqrt{z})}{F_k(z;\beta)}-\frac{J_0(\sqrt{z})(d/dz)F_k(z;\beta)}{F_k^2(z;\beta)},$$

$$R_{0k}^{(2)}(z;\beta)=\frac{(d^2/dz^2)J_0(\sqrt{z})}{F_k(z;\beta)}-2\frac{[(d/dz)J_0(\sqrt{z})][(d/dz)F_k(z;\beta)]}{F_k^2(z;\beta)}$$

$$+2\frac{J_0(\sqrt{z})[(d/dz)F_k(z;\beta)]^2}{F_k^3(z;\beta)}-\frac{J_0(\sqrt{z})(d^2/dz^2)F_k(z;\beta)}{F_k^2(z;\beta)},$$

etc., so that the evaluation of $R_{0k}^{(m)}(z;\beta)$ requires the evaluation of $(d^n/dz^n)J_0(\sqrt{z})$ and $(d^n/dz^n)F_k(z;\beta)$, $n=0,1,\cdots,m$. Because $F(z;\beta)=(z-\lambda_k^2)F_k(z;\beta)$ we find that

$$\frac{d^n}{dz^n}F(z;\beta)=n\frac{d^{n-1}}{dz^{n-1}}F_k(z;\beta)+(z-\lambda_k^2)\frac{d^n}{dz^n}F_k(z;\beta),$$

and hence in $R_{0k}^{(m)}(\lambda_k^2;\beta)$ we can replace $F_k^{(n)}(\lambda_k^2;\beta)$ by $F^{(n+1)}(\lambda_k^2;\beta)/(n+1)$, $n=0,1,\cdots,m$.

Now the formulaе

$$\frac{d}{dz}J_0(\sqrt{z})=-\frac{1}{2z}\sqrt{z}J_1(\sqrt{z})$$

and

$$\frac{d}{dz}\sqrt{z}J_1(\sqrt{z})=\frac{1}{2}J_0(\sqrt{z})$$

imply that $(d^n/dz^n)F(z;\beta)$, as well as $(d^n/dz^n)J_0(\sqrt{z})$, is a sum of two terms, each a rational polynomial in z and β times $J_0(\sqrt{z})$ or $\sqrt{z}J_1(\sqrt{z})$. But $F(\lambda_k^2;\beta)=\lambda_kJ_1(\lambda_k)-\beta J_0(\lambda_k)=0$ hence neither $J_0(\lambda_k)$ nor $J_1(\lambda_k)$ appears in $R_{0k}^{(m)}(\lambda_k^2;\beta)$. It follows that $R_{0k}^{(m)}(\lambda_k^2;\beta)$ is a rational polynomial in λ_k^2 and β and so also therefore is $S_0^{m,k}(\beta)$. The calculation of $S_0^{1,k}(\beta)$ and $S_0^{2,k}(\beta)$ illustrates this:

$$\frac{d}{dz}J_0(\sqrt{z})=-\frac{1}{2z}\sqrt{z}J_1(\sqrt{z})$$

and

$$\frac{d^2}{dz^2}J_0(\sqrt{z})=-\frac{1}{4z}J_0(\sqrt{z})+\frac{1}{2z^2}\sqrt{z}J_1(\sqrt{z})$$

imply

$$\left.\frac{d}{dz}J_0(\sqrt{z})\right|_{z=\lambda_k^2}=-\frac{\beta}{2\lambda_k^2}J_0(\lambda_k)$$

and

$$\left.\frac{d^2}{dz^2}J_0(\sqrt{z})\right|_{z=\lambda_k^2}=\left[-\frac{1}{4\lambda_k^2}+\frac{\beta}{2\lambda_k^4}\right]J_0(\lambda_k);$$

whereas

$$\frac{d}{dz}F(z;\beta)=\frac{1}{2}J_0(\sqrt{z})+\frac{\beta}{2z}\sqrt{z}J_1(\sqrt{z}),$$

$$\frac{d^2}{dz^2}F(z;\beta)=\left[-\frac{1}{4z}-\frac{\beta}{2z^2}\right]\sqrt{z}J_1(\sqrt{z})+\frac{\beta}{4z}J_0(\sqrt{z})$$

and

$$\frac{d^3}{dz^3}F(z;\beta)=\left[-\frac{1}{8z}-\frac{\beta}{2z^2}\right]J_0(\sqrt{z})+\left[\frac{1}{4z^2}+\frac{\beta}{z^3}-\frac{\beta}{8z^2}\right]\sqrt{z}J_1(\sqrt{z})$$

imply

$$F_k(\lambda_k^2;\beta)=\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]J_0(\lambda_k),$$

$$2F_k^{(1)}(\lambda_k^2;\beta)=\left[-\frac{1}{4\lambda_k^2}-\frac{\beta}{2\lambda_k^4}\right]\beta J_0(\lambda_k)+\frac{\beta}{4\lambda_k^2}J_0(\lambda_k)$$

and

$$3F_k^{(2)}(\lambda_k^2;\beta)=\left[-\frac{1}{8\lambda_k^2}-\frac{\beta}{2\lambda_k^4}\right]J_0(\lambda_k)+\left[\frac{1}{4\lambda_k^4}+\frac{\beta}{\lambda_k^6}-\frac{\beta}{8\lambda_k^4}\right]\beta J_0(\lambda_k).$$

Thus we have

$$R_{0k}^{(1)}(\lambda_k^2;\beta)=\frac{-\beta}{2\lambda_k^2\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]}-\frac{\frac{\beta}{8\lambda_k^2}+\frac{\beta}{2}\left[\frac{-1}{4\lambda_k^2}-\frac{\beta}{2\lambda_k^4}\right]}{\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]^2},$$

$$\begin{aligned}R_{0k}^{(2)}(\lambda_k^2;\beta)=&\frac{\frac{-1}{4\lambda_k^2}+\frac{\beta}{2\lambda_k^4}}{\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}}+\frac{\frac{\beta}{\lambda_k^2}\left[\frac{\beta}{8\lambda_k^2}+\frac{\beta}{2}\left[\frac{-1}{4\lambda_k^2}-\frac{\beta}{2\lambda_k^4}\right]\right]}{\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]^2}\\&+\frac{\frac{1}{24\lambda_k^2}+\frac{\beta}{6\lambda_k^4}-\frac{\beta}{3}\left[\frac{1}{4\lambda_k^4}-\frac{\beta}{8\lambda_k^4}+\frac{\beta}{\lambda_k^6}\right]}{\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]^2}+\frac{2\left[\frac{\beta}{8\lambda_k^2}+\frac{\beta}{2}\left[\frac{-1}{4\lambda_k^2}-\frac{\beta}{2\lambda_k^4}\right]\right]^2}{\left[\frac{1}{2}+\frac{\beta^2}{2\lambda_k^2}\right]^3}.\end{aligned}$$

The formulae for $S_0^{m,k}(\beta)$, $m=1,2,\cdots,5$, are recorded in Table 1; these are sufficient for $X_{2\infty}$. For $m=1,2,3$ the work can be done by hand; for higher values of m it

TABLE 1
The sums $S_0^{m,k}(\beta)$.

m	$2(\lambda_k^2+\beta^2)^m S_0^{m,k}(\beta)$
1	$\beta-\frac{\beta^2}{\lambda_k^2+\beta^2}$
2	$\frac{1}{3!}\left[\lambda_k^2-4\beta+\beta^2+\frac{4\beta^2}{\lambda_k^2}\right]-\frac{\beta^4}{2\lambda_k^2(\lambda_k^2+\beta^2)}$
3	$\frac{1}{4!}\left[-3\lambda_k^2+12\beta-2\beta^2+\frac{-12\beta^2+2\beta^3+\beta^4}{\lambda_k^2}+\frac{4\beta^4}{\lambda_k^4}\right]-\frac{\beta^6}{4\lambda_k^4(\lambda_k^2+\beta^2)}$
4	$\frac{1}{3\cdot 5!}\left[\lambda_k^4+(36+4\beta+3\beta^2)\lambda_k^2-144\beta+15\beta^2+8\beta^3+3\beta^4+\frac{144\beta^2-68\beta^3-28\beta^4+4\beta^5+\beta^6}{\lambda_k^2}+\frac{28\beta^4-14\beta^5-7\beta^6}{\lambda_k^4}+\frac{64\beta^6}{\lambda_k^6}\right]-\frac{\beta^8}{8\lambda_k^6(\lambda_k^2+\beta^2)}$
5	$\frac{1}{2\cdot 6!}\left[-5\lambda_k^4-(120+30\beta+16\beta^2)\lambda_k^2+480\beta-36\beta^2-64\beta^3-18\beta^4+\frac{-480\beta^2+424\beta^3+155\beta^4-38\beta^5-8\beta^6}{\lambda_k^2}+\frac{-384\beta^4+188\beta^5+88\beta^6-4\beta^7-\beta^8}{\lambda_k^4}+\frac{-348\beta^6+34\beta^7+17\beta^8}{\lambda_k^6}+\frac{36\beta^8}{\lambda_k^8}\right]-\frac{\beta^{10}}{16\lambda_k^8(\lambda_k^2+\beta^2)}$

cannot. The Symbolic Manipulation Program[2] (SMP) is indispensable for $m > 3$. The formulae for $S_1^{m,k}(\beta)$, $m = 1, 2, \cdots, 5$, are recorded in Table 2.

The dispersion coefficient $X_{2\infty}$ is then

$$X_{2\infty} = 1 + N_{pe}^2\left[\frac{-2\beta^6(1+\beta)^2}{9\lambda_1^6(\lambda_1^2+\beta^2)^3} + \frac{2\beta^4(1+\beta)^2(4+\beta)}{9\lambda_1^6(\lambda_1^2+\beta^2)^2} - \frac{2\beta^2(42+35\beta+25\beta^2+5\beta^3)}{45\lambda_1^6(\lambda_1^2+\beta^2)} \right.$$
$$\left. - \frac{(21+14\beta+\beta^2)}{45\lambda_1^2(\lambda_1^2+\beta^2)} + \frac{\beta(84+57\beta+10\beta^2)}{45\lambda_1^4(\lambda_1^2+\beta^2)} - \frac{1}{45(\lambda_1^2+\beta^2)}\right]$$

where it remains only to establish λ_1^2 as a function of β.[3] For $\beta \cong 0$, the following is useful:

$$\lambda_1^2 = 2\beta - \frac{1}{2}\beta^2 + \frac{1}{12}\beta^3 - \frac{1}{192}\beta^4 - \frac{1}{640}\beta^5 + \frac{31}{69120}\beta^6 + \frac{1}{387072}\beta^7 - \frac{3221}{123863040}\beta^8 + \cdots,$$

which was derived from $(d/d\beta)\lambda_1^2 = 2\lambda_1^2/(\lambda_1^2+\beta^2)$ by the method of Frobenius.

What we achieve, then, in the calculation of $X_{2\infty}$ is this: we need only estimate λ_1^2 vs. β; we need not estimate the summand, viz., λ_j^2 vs. β, $j = 2, 3, \cdots$, nor the sum itself. More generally we note that only λ_k^2 vs. β is required in order to get $S_0^{m,k}(\beta)$

TABLE 2
The sums $S_1^{m,k}(\beta)$.

m	$2\beta^2(\lambda_k^2+\beta^2)^m S_1^{m,k}(\beta)$
1	$-\beta + \dfrac{2\beta^2}{\lambda_k^2} - \dfrac{\beta^4}{\lambda_k^2(\lambda_k^2+\beta^2)}$
2	$\dfrac{1}{3!}\left[6\beta + \beta^2 + \dfrac{-12\beta^2+2\beta^3+\beta^4}{\lambda_k^2} - \dfrac{2\beta^4}{\lambda_k^4}\right] - \dfrac{\beta^6}{2\lambda_k^4(\lambda_k^2+\beta^2)}$
3	$\dfrac{1}{4!}\left[-24\beta - 7\beta^2 + \dfrac{48\beta^2-20\beta^3-10\beta^4}{\lambda_k^2} + \dfrac{44\beta^4-6\beta^5-3\beta^6}{\lambda_k^4} + \dfrac{24\beta^6}{\lambda_k^6}\right] - \dfrac{\beta^8}{4\lambda_k^6(\lambda_k^2+\beta^2)}$
4	$\dfrac{1}{3\cdot 5!}\left[\beta^2\lambda_k^2 + 360\beta + 141\beta^2 + 4\beta^3 + 3\beta^4 + \dfrac{-720\beta^2+516\beta^3+270\beta^4+8\beta^5+3\beta^6}{\lambda_k^2}\right.$ $\left. + \dfrac{-1236\beta^4+322\beta^5+167\beta^6+4\beta^7+\beta^8}{\lambda_k^4} + \dfrac{-992\beta^6+76\beta^7+38\beta^8}{\lambda_k^6} - \dfrac{206\beta^8}{\lambda_k^8}\right]$ $- \dfrac{\beta^{10}}{8\lambda_k^8(\lambda_k^2+\beta^2)}$
5	$\dfrac{1}{2\cdot 6!}\left[-9\beta^2\lambda_k^2 - 1440\beta - 684\beta^2 - 46\beta^3 - 32\beta^4\right.$ $+ \dfrac{2880\beta^2-3024\beta^3-1680\beta^4-112\beta^5-42\beta^6}{\lambda_k^2} + \dfrac{7344\beta^4-2928\beta^5-1593\beta^6-86\beta^7-24\beta^8}{\lambda_k^4}$ $\left. + \dfrac{8528\beta^6-1404\beta^7-732\beta^8-20\beta^9-5\beta^{10}}{\lambda_k^6} + \dfrac{4444\beta^8-270\beta^9-135\beta^{10}}{\lambda_k^8} + 1040\dfrac{\beta^{10}}{\lambda_k^{10}}\right]$ $- \dfrac{\beta^{12}}{16\lambda_k^{10}(\lambda_k^2+\beta^2)}$

[2] SMP Reference Manual, Inference Corporation, Los Angeles, California, 1983.

[3] SMP is useful in combining the formulae of Table 1 into a simple formula for $X_{2\infty}$. The results are correct; in particular they agree with all that we can do by hand, viz., small m, small β, etc., and when evaluated they agree with earlier estimates of $X_{2\infty}$.

for the formula we derive is a rational polynomial in λ_k^2 and β; yet we offer this caveat: to get[4] $\lim_{\beta\to 0} S_0^{m,1}(\beta)$ requires that λ_1^2 be accurate to order m, at least, in β. This is what is necessary to reproduce Rayleigh's results. Thus, approximations to λ_1^2 lead to lower order approximations to $S_0^{m,1}(\beta)$ so that crude approximations to λ_1^2 lead to nonsense.

4. Other sums. The partial fraction expansions

$$\frac{J_0(\alpha\lambda)}{\lambda J_1(\lambda)-\beta J_0(\lambda)}=\sum_{j=1}\frac{1}{\lambda^2-\lambda_j^2}\frac{2\lambda_j^2}{\lambda_j^2+\beta^2}\frac{J_0(\alpha\lambda_j)}{J_0(\lambda_j)},\qquad 0\leqq\alpha\leqq 1,$$

and

$$\frac{J_1(\alpha\lambda)}{\lambda J_1(\lambda)-\beta J_0(\lambda)}=\sum_{j=1}\frac{\lambda\beta}{\lambda^2-\lambda_j^2}\frac{2}{\lambda_j^2+\beta^2}\frac{J_1(\alpha\lambda_j)}{J_1(\lambda_j)},\qquad 0\leqq\alpha\leqq 1,$$

are of interest. Their justification rests on obtaining useful bounds for $J_0(\alpha\lambda)/J_0(\lambda)$ and $J_1(\alpha\lambda)/J_1(\lambda)$ on C_j.

Retaining the sequence of contours C_j in § 2 we find that Conditions (i) and (ii) are satisfied; to show that Conditions (iii) and (iv) are satisfied we use

$$J_0(\lambda)\sim\frac{\sqrt{2}}{\sqrt{\pi\lambda}}\left[\cos\left(\lambda-\frac{\pi}{4}\right)\left(1+O\left(\frac{1}{\lambda^2}\right)\right)-\sin\left(\lambda-\frac{\pi}{4}\right)\left(\frac{-1}{8\lambda}+O\left(\frac{1}{\lambda^3}\right)\right)\right],\qquad \operatorname{Re}\lambda\geqq 0$$

(cf. Watson [8, p. 199]) so that

$$\frac{J_0(\alpha\lambda)}{J_0(\lambda)}\to\frac{1}{\sqrt{\alpha}}\frac{\cos(\alpha\lambda)(1-1/(8\alpha\lambda))+\sin(\alpha\lambda)(1+1/(8\alpha\lambda))}{\cos(\lambda)(1-1(8\lambda))+\sin(\lambda)(1-1/(8\lambda))},\qquad \alpha>0.$$

We conclude for large λ that $|J_0(\alpha\lambda)/J_0(\lambda)|^2$ is bounded by $(1/\alpha)(\cosh(2\alpha\operatorname{Im}\lambda)+\sin(2\alpha\operatorname{Re}\lambda))/(\cosh(2\operatorname{Im}\lambda)+\sin(2\operatorname{Re}\lambda))$; but this is less than $2/\alpha$ on the vertical sides of C_j and is less than $(1/\alpha)\coth^2(\pi j)$ on the horizontal sides of C_j for $\alpha\in(0,1)$. It follows that $|J_0(\alpha\lambda)/J_0(\lambda)|$ is bounded by $2^{1/2}/\alpha^{1/2}$ on C_j as $j\to\infty$ and hence that the first expansion is justified for $\alpha\in(0,1]$. For $\alpha=0$, we find that $|J_0(\lambda)|^{-2}$ is bounded by $\pi|\lambda|/(\cosh(2\operatorname{Im}\lambda)+\sin(2\operatorname{Re}\lambda))$, and thus by $\pi\cdot\pi j$ on the vertical sides of C_j and by $\pi\cdot\sqrt{2}\pi j/\sinh(2\pi j)$ on the horizontal sides of C_j. It follows that $|J_0(\lambda)|^{-1}$ is bounded by $\pi\sqrt{j}$ on C_j as $j\to\infty$, and hence that the first expansion is justified for $\alpha=0$ inasmuch as $\pi\sqrt{j}/(\pi j-\beta)$ vanishes as $j\to\infty$. In a similar way we can show that $|J_1(\alpha\lambda)/J_1(\lambda)|$ is uniformly bounded on C_j by $2^{1/2}/\alpha^{1/2}$ so that Conditions (iii) and (iv) obtain therein, and hence that the second expansion is justified for $\alpha\in(0,1)$. The first expansion implies that

$$R_{0k}(z;\alpha,\beta)\equiv\frac{J_0(\alpha\sqrt{z})}{F_k(z:\beta)}=\frac{2\lambda_k^2}{\lambda_k^2+\beta^2}\frac{J_0(\alpha\lambda_k)}{J_0(\lambda_k)}+(z-\lambda_k^2)\sum_{j\neq k}\frac{2\lambda_j^2}{(z-\lambda_j^2)(\lambda_j^2+\beta^2)}\frac{J_0(\alpha\lambda_j)}{J_0(\lambda_j)}$$

and hence that

$$R_{0k}^{(n)}(\lambda_k^2;\alpha,\beta)=(-1)^{n-1}n!\sum_{j\neq k}\frac{2\lambda_j^2}{(\lambda_k^2-\lambda_j^2)^n(\lambda_j^2+\beta^2)}\frac{J_0(\alpha\lambda_j)}{J_0(\lambda_j)},$$

so that the evaluation of the infinite sum reduces to the evaluation of $R_{0k}^{(n)}(\lambda_k^2;\alpha,\beta)$.

[4] We became interested in this not only to check the results but because constant coefficient dispersion models are most useful for small values of β.

For example, because

$$R_{0k}^{(1)}(z;\alpha=0,\beta)=-\frac{(d/dz)F_k(z;\beta)}{F_k^2(z;\beta)}$$

and

$$R_{0k}^{(2)}(z;\alpha=0,\beta)=2\frac{[(d/dz)F_k(z;\beta)]^2}{F_k^3(z;\beta)}-\frac{(d^2/dz^2)F_k(z;\beta)}{F_k^2(z;\beta)},$$

we find at $\beta=0$ that

$$\frac{J_1(\lambda_k)}{2\lambda_k J_0^2(\lambda_k)}=\sum_{j\neq k}\frac{2}{\lambda_k^2-\lambda_j^2}\frac{1}{J_0(\lambda_k)}$$

and

$$\frac{1}{4\lambda_k^2}\frac{J_1^2(\lambda_k)}{J_0^3(\lambda_k)}+\frac{1}{3\lambda_k^2}\frac{1}{J_0(\lambda_k)}\left[\frac{1}{2}-\frac{J_1(\lambda_k)}{\lambda_k J_0(\lambda_k)}\right]=-2\sum_{j\neq k}\frac{2}{(\lambda_k^2-\lambda_j^2)^2}\frac{1}{J_0(\lambda_j)}.$$

Because

$$J_0(\lambda)=1-\frac{1}{4}\lambda^2+\cdots \quad\text{and}\quad \frac{J_1(\lambda)}{\lambda J_0(\lambda)}=\frac{1}{2}+\frac{1}{16}\lambda^2+\frac{1}{96}\lambda^4+\cdots$$

as $\lambda\to 0$ and $\lambda_1=0$, it follows for $k=1$ that

$$-\frac{1}{8}=\sum_{j=2}\frac{1}{\lambda_j^2}\frac{1}{J_0(\lambda_j)},\qquad -\frac{1}{96}=\sum_{j=2}\frac{1}{\lambda_j^4}\frac{1}{J_0(\lambda_j)};$$

because λ_k satisfies $J_1(\lambda_k)=0$, it follows for $k>1$ that

$$0=\sum_{j\neq k}\frac{1}{\lambda_k^2-\lambda_j^2}\frac{1}{J_0(\lambda_j)},\qquad -\frac{1}{24}\frac{1}{\lambda_k^2 J_0(\lambda_k)}=\sum_{j\neq k}\frac{1}{(\lambda_k^2-\lambda_j^2)^2}\frac{1}{J_0(\lambda_j)}.$$

The sum, $\sum_{j=2}1/(\lambda_j^4 J_0(\lambda_j))$, appears in the work of Bhattacharya and Gupta [3]; its summation, $-1/96$, was conjectured but not rigorously justified.

There are values of z other than $z=\lambda_k^2$ at which the expansion of $R_{0k}(z;\alpha,\beta)$, or more simply

$$R_0(z;\alpha=1,\beta)=\frac{J_0(\sqrt{z})}{\sqrt{z}J_1(\sqrt{z})-\beta J_0(\sqrt{z})}=\sum_{j=1}\frac{2\lambda_j^2}{(z-\lambda_j^2)(\lambda_j^2+\beta^2)},$$

yields useful information. For instance, on setting $z=0$ we find a generalization of Rayleigh's sum, viz., for $m=0,1,\cdots$

$$R_0(0;\alpha=1,\beta)=-m!\sum_{j=1}\frac{2}{\lambda_j^{2m}(\lambda_j^2+\beta^2)}$$

implies

$$\frac{1}{\beta}=\sum_{j=1}\frac{2}{\lambda_j^2+\beta^2},$$

$$\frac{1}{2\beta^2}=\sum_{j=1}\frac{2}{\lambda_j^2(\lambda_j^2+\beta^2)}.$$

More generally $z=\lambda_k^2(0)$ and $z=\lambda_k^2(\infty)$ are interesting: J_1 vanishes at the former, J_0 at the latter.

5. Sums for the parallel plane geometry. There are other applications of the Cauchy partial fraction expansion in dispersion calculations. For instance in the parallel plane geometry (a limiting cylinder) the coefficient $X_{2\infty}$ becomes

$$X_{2\infty}=1+16N_{pe}^2\frac{\lambda_1^2}{\lambda_1^2+\beta+\beta^2}\sum_{j\neq 1}\frac{\lambda_j^2}{(\lambda_j^2+\beta+\beta^2)}\frac{(\lambda_j^2+\lambda_1^2+2\beta+2\beta^2)^2}{(\lambda_j^2-\lambda_1^2)^5}$$

where $\lambda_1<\lambda_2<\cdots$ denotes the increasing sequence of nonnegative zeros of

$$\lambda_j\sin(\lambda_j)-\beta\cos(\lambda_j)=0, \qquad \beta\geqq 0.$$

The infinite sum can be rewritten

$$S_c^{3,1}(\beta)+4(\lambda_1^2+\beta+\beta^2)\ S_c^{4,1}(\beta)+4(\lambda_1^2+\beta+\beta^2)^2S_c^{5,1}(\beta)$$

where

$$S_c^{m,k}(\beta)=\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\frac{\lambda_j^2}{(\lambda_j^2+\beta+\beta^2)}.$$

In Tables 3 and 4 we record the SMP derived formulae for $S_c^{m,k}(\beta)$ and $S_s^{m,k}(\beta)$, $m=1,2,\cdots,5$, where

$$S_s^{m,k}(\beta)=\sum_{j\neq k}\frac{1}{(\lambda_j^2-\lambda_k^2)^m}\frac{1}{(\lambda_j^2+\beta+\beta^2)}.$$

TABLE 3

The sums $S_c^{m,k}(\beta)$.

m	$2(\lambda_k^2+\beta+\beta^2)^mS_c^{m,k}(\beta)$
1	$\frac{1}{2}+\beta^2-\frac{\beta(1+\beta)}{\lambda_k^2+\beta+\beta^2}$
2	$\frac{1}{3!}\left[\frac{1}{4}(-9-12\beta+4\beta^2+4\lambda_k^2)+\frac{15\beta(1+\beta)}{4\lambda_k^2}\right]-\frac{\beta^2(1+\beta)^2}{2\lambda_k^2(\lambda_k^2+\beta+\beta^2)}$
3	$\frac{1}{4!}\left[\frac{1}{2}(15+20\beta-6\lambda_k^2-4\beta^2)+\frac{\beta(1+\beta)(-9+3\beta+\beta^2)}{\lambda_k^2}+\frac{9}{2\lambda_k^4}\beta^2(1+\beta)^2\right]-\frac{\beta^3(1+\beta)^3}{4\lambda_k^4(\lambda_k^2+\beta+\beta^2)}$
4	$\frac{1}{5!}\left[\frac{1}{48}(-1575-2040\beta+456\beta^2+224\beta^3+48\beta^4+\lambda_k^2(600+112\beta+48\beta^2)+16\lambda_k^4)+\frac{\beta(1+\beta)(1035-1560\beta-360\beta^2+96\beta^3+16\beta^4)}{48\lambda_k^2}-\frac{5\beta^2(1+\beta)^2(3+24\beta+8\beta^2)}{16\lambda_k^4}\right]+\frac{21\beta^3(1+\beta)^3}{128\lambda_k^6}-\frac{\beta^4(1+\beta)^4}{8\lambda_k^6(\lambda_k^2+\beta+\beta^2)}$
5	$\frac{1}{6!}\left[\frac{-1}{16}(-2835-3480\beta+1096\beta^2+800\beta^3+144\beta^4+\lambda_k^2(1050+368\beta+128\beta^2)+40\lambda_k^4)-\frac{\beta(1+\beta)(45-1200\beta-213\beta^2+108\beta^3+16\beta^4)}{4\lambda_k^2}-\frac{\beta^2(1+\beta)^2(90-300\beta-90\beta^2+6\beta^3+\beta^4)}{2\lambda_k^4}+\frac{15\beta^3(1+\beta)^2(-66+15\beta+5\beta^2)}{8\lambda_k^6}\right]+\frac{9\beta^4(1+\beta)^4}{256\lambda_k^8}-\frac{\beta^5(1+\beta)^5}{16\lambda_k^8(\lambda_k^2+\beta+\beta^2)}$

TABLE 4
The sums $S_s^{m,k}(\beta)$.

m	$2\beta(\lambda_k^2+\beta+\beta^2)^m S_s^{m,k}(\beta)$
1	$\frac{3}{2}\frac{\beta}{\lambda_k^2}-1-\frac{\beta^2(1+\beta)}{\lambda_k^2(\lambda_k^2+\beta+\beta^2)}$
2	$\frac{1}{3!}\left[(6+\beta)+\frac{\beta(-21+12\beta+4\beta^2)}{4\lambda_k^2}+\frac{3\beta^2(1+\beta)}{4\lambda_k^4}\right]-\frac{\beta^3(1+\beta)^2}{2\lambda_k^4(\lambda_k^2+\beta+\beta^2)}$
3	$\frac{1}{4!}\left[-(24+7\beta)-\frac{\beta(-9+60\beta+20\beta^2)}{2\lambda_k^2}-\frac{3\beta^2(1+\beta)(-3+3\beta+\beta^2)}{\lambda_k^4}+\frac{27\beta^3(1+\beta)^2}{2\lambda_k^6}\right]-\frac{\beta^4(1+\beta)^3}{4\lambda_k^6(\lambda_k^2+\beta+\beta^2)}$
4	$\frac{1}{5!}\left[\frac{1}{6}(720+285\beta+2\beta\lambda_k^2+14\beta^2+6\beta^3)+\frac{\beta(3105+12600\beta+4536\beta^2+224\beta^3+48\beta^4)}{48\lambda_k^2}+\frac{\beta^2(1+\beta)(-2205+7800\beta+2760\beta^2+96\beta^3+16\beta^4)}{48\lambda_k^4}+\frac{5\beta^3(1+\beta)^2(-363+120\beta+40\beta^2)}{16\lambda_k^6}\right]-\frac{19\beta^4(1+\beta)^3}{128\lambda_k^8}-\frac{\beta^5(1+\beta)^4}{8\lambda_k^8(\lambda_k^2+\beta+\beta^2)}$
5	$\frac{1}{6!}\left[\frac{-1}{8}(5760+2805\beta+312\beta^2+128\beta^3+36\beta\lambda_k^2)-\frac{\beta(14895+37800\beta+14952\beta^2+1568\beta^3+336\beta^4)}{16\lambda_k^2}-\frac{\beta^2(1+\beta)(495+9000\beta+3435\beta^2+268\beta^3+48\beta^4)}{4\lambda_k^4}-\frac{5\beta^3(1+\beta)^2(-729+840\beta+300\beta^2+12\beta^3+2\beta^4)}{4\lambda_k^6}-\frac{15\beta^4(1+\beta)^3(-354+105\beta+35\beta^2)}{8\lambda_k^8}\right]+\frac{79\beta^5(1+\beta)^4}{256\lambda_k^{10}}-\frac{\beta^6(1+\beta)^5}{16\lambda_k^{10}(\lambda_k^2+\beta+\beta^2)}$

Here the Cauchy partial fraction expansions

$$\frac{\cos(\lambda)}{\lambda\sin(\lambda)-\beta\cos(\lambda)}=\sum_{j=1}\frac{2\lambda_j^2}{(\lambda^2-\lambda_j^2)(\lambda_j^2+\beta+\beta^2)}$$

and

$$\frac{\sin(\lambda)}{\lambda\sin(\lambda)-\beta\cos(\lambda)}=\sum_{j=1}\frac{2\lambda\beta}{(\lambda^2-\lambda_j^2)(\lambda_j^2+\beta+\beta^2)}$$

lead to useful representations of

$$(z-\lambda_k^2)\frac{\cos(\sqrt{z})}{\sqrt{z}(\sqrt{z}\sin(\sqrt{z})-\beta\cos(\sqrt{z}))}$$

and

$$(z-\lambda_k^2)\frac{\sin(\sqrt{z})}{\sqrt{z}(\sqrt{z}\sin(\sqrt{z})-\beta\cos(\sqrt{z}))}$$

and thence of $S_c^{m,k}(\beta)$ and $S_s^{m,k}(\beta)$.

The four conditions of the first form of Cauchy's theorem are satisfied for $f(\lambda)=\lambda\sin(\lambda)-\beta\cos(\lambda)$, $g(\lambda)=\cos(\lambda)$ and $g(\lambda)=\sin(\lambda)$. In particular, Condition (i) is satisfied because $g(\lambda)/f(\lambda)$ is regular save for the zeros of $f(\lambda)$. If C_j denotes the square contour on the Argand diagram with vertices at $(\pm(j+\frac{3}{4})\pi, \pm(j+\frac{3}{4})\pi)$ then the sequence $\{C_j\}$ satisfies Condition (ii). Indeed the zeros of $f(\lambda)$ are real; the proof parallels Lommel's proof for $\lambda J_1(\lambda)-\beta J_0(\lambda)$. And the zeros of $f(\lambda)$ lie on the interval $(j\pi, j\pi+\frac{1}{2}\pi)$; viz., $(d/d\lambda)\lambda\tan(\lambda)\geqq 0$ on the interval $(j\pi-\frac{1}{2}\pi,\ j\pi+\frac{1}{2}\pi)$ and $j\pi\tan(j\pi)=0$. To see that Conditions (iii) and (iv) are satisfied we note that $|\tan(\lambda)|^2=(\cosh(2\operatorname{Im}\lambda)-\cos(2\operatorname{Re}\lambda))/(\cosh(2\operatorname{Im}\lambda)+\cos(2\operatorname{Re}\lambda))$ implies that on the horizontal sides of C_j $\tanh(j+\frac{3}{4})\pi\leqq|\tan(\lambda)|\leqq\coth(j+\frac{3}{4})\pi$, and that on the vertical sides of C_j $|\tanh(\lambda)|=1$. Thus $|\tan(\lambda)|\to 1$ on C_j as $j\to\infty$. It follows that $|g(\lambda)/f(\lambda)|$ is bounded by $(|\lambda|-\beta)^{-1}\leqq((j+\frac{3}{4})\pi-\beta)^{-1}$ on C_j as $j\to\infty$ and hence vanishes uniformly as $j\to\infty$.

The dispersion coefficient $X_{2\infty}$ is then

$$X_{2\infty}=1+N_{pe}^2\frac{1}{(\lambda_1^2+\beta+\beta^2)^3}$$
$$\cdot\left[-\frac{7}{8}\frac{\beta^3(1+\beta)^3}{\lambda_1^6}-\frac{\beta(1+\beta)(-225-1230\beta-330\beta^2+48\beta^3+8\beta^4)}{360\lambda_1^2}\right.$$
$$+\frac{\beta^2(1+\beta)^2(-33+30\beta+10\beta^2)}{24\lambda_1^4}+\frac{585+810\beta-78\beta^2-192\beta^3-24\beta^4-8\lambda_1^4}{360}$$
$$\left.-\lambda_1^2\left[\frac{7}{12}+\frac{17\beta}{45}+\frac{\beta^2}{15}\right]\right];$$

for $\beta\cong 0$, the following is useful:

$$\lambda_1^2=\beta-\frac{1}{3}\beta^2+\frac{4}{45}\beta^3-\frac{16}{945}\beta^4+\frac{16}{14175}\beta^5+\frac{64}{93555}\beta^6-\frac{69248}{212837625}\beta^7+\frac{512}{8292375}\beta^8+\cdots$$

which was derived from $(d/d\beta)\lambda_1^2=2\lambda_1^2/(\lambda_1^2+\beta+\beta^2)$ by the method of Frobenius.

6. Conclusion. The foregoing analysis turns on two items. Firstly, the existence of the Cauchy partial fraction expansion for $g(\lambda)/f(\lambda)$. This is justified if Conditions (i), (ii), (iii) and (iv) of Cauchy's theorem (cf. Copson [4, pp. 144-148]), are satisfied. Secondly, the evaluation of

$$\frac{d^n}{d\lambda^n}\frac{\lambda-\lambda_k}{f(\lambda)}g(\lambda)|_{\lambda=\lambda_k}.$$

This can be done because $\lambda-\lambda_k$ is an unrepeated factor of $f(\lambda)$ so that if $f(\lambda)$ is regular then $f_k(\lambda)=f(\lambda)/(\lambda-\lambda_k)$ is regular; hence $f_k^{(n)}(\lambda_k)=f^{(n+1)}(\lambda_k)/(n+1)$.

Acknowledgment. The help of Dale Prouty of Inference Corporation in using SMP is gratefully acknowledged.

REFERENCES

[1] S. Ahmed and M. E. Muldoon, *Reciprocal power sums of differences of zeros of special functions*, this Journal, 14 (1983), pp. 372-382.

[2] R. Aris, *Hierarchies of models in reactive systems*, in Dynamics and Modeling of Reactive Systems, W. E. Stewart, W. H. Ray and C. C. Conley, eds., Academic Press, New York, 1980, pp. 1-35.

[3] R. N. Bhattacharya and V. K. Gupta, *On the Taylor-Aris theory of solute transport in a capillary*, SIAM J. Appl. Math., 44 (1984), pp. 33-39.

[4] E. T. Copson, *An Introduction to the Theory of Functions of a Complex Variable*, Oxford Univ. Press, London, 1962.

[5] A. E. DeGance and L. E. Johns, *The theory of dispersion of chemically active solutes in a rectilinear flow field*, Appl. Sci. Res., 34 (1978), pp. 189–225.

[6] ———, *The theory of dispersion of chemically active solutes in a rectilinear flow field. The vector problem*, Appl. Sci. Res., 42 (1985), pp. 55–88.

[7] Lord Rayleigh (J. W. Strutt), *Note on the numerical calculation of the roots of fluctuating functions*, Proc. London Math. Soc., V (1874), pp. 119–124.

[8] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, London, 1944.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

017

AN ELEMENTARY INEQUALITY FOR WHICH EQUALITY HOLDS IN AN INFINITE-DIMENSIONAL SET*

RAY REDHEFFER† AND ALEXANDER VOIGT‡

Abstract. An inequality involving an infinite series of geometric means has the curious property that the set for which equality holds is of infinite dimension.

Key words. inequality, infinite set

AMS(MOS) subject classifications. 26-01, 26A86, 26D15

1. Introduction. Throughout this note $n \in N = \{1, 2, 3, \cdots\}$ and $\{a_n\}$ for $n \in N$ is a sequence of positive real numbers. The geometric mean of the first n of these numbers is denoted by

$$G_n = (a_1 a_2 \cdots a_n)^{1/n}.$$

Our principal objective is to establish the following.

THEOREM 1. *Let* $a_1 + a_3 + a_5 + \cdots < \infty$. *Then*

$$G_1 - 2G_2 + 3G_3 - 4G_4 + \cdots \leqq a_1 + a_3 + a_5 + \cdots \tag{1}$$

whenever the left-hand side converges. If $a_{2j-1} \neq o(1/j)$ *the inequality is strict, but if* $a_{2j-1} = o(1/j)$ *there is exactly one choice of* $\{a_{2j}\}$ *for which equality holds.*

The condition $a_{2j-1} = o(1/j)$ allows an infinite-dimensional subset of the Hilbert space l_2, and if we had $O(1/j)$ instead of $o(1/j)$, the subset would be isomorphic to the Hilbert cube. We know of no other inequality that exhibits behavior such as this.

In the course of proving Theorem 1 we review some significant classical inequalities from our own point of view. We also supplement the theorem by an inequality in the opposite direction and discuss the difficult problem of convergence.

2. Inequalities of Bernoulli, Rado and Maclaurin. Let us start from Bernoulli's inequality

$$(1+y)^n > 1 + ny, \quad n \geqq 2, \quad y > -1, \quad y \neq 0 \tag{2}$$

for which an inductive proof can be given with ease. Setting $x = 1 + y$ we get the equivalent inequality

$$x^n > nx - (n-1), \quad n \geqq 2, \quad x > 0, \quad x \neq 1. \tag{3}$$

If we choose $x = G_n / G_{n-1}$ in (3) and multiply by G_{n-1} the result is

$$a_n \geqq nG_n - (n-1)G_{n-1}, \qquad n \geqq 2, \tag{4}$$

with strict inequality unless $G_n = G_{n-1}$. The arithmetic mean

$$A_n = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

satisfies $a_n = nA_n - (n-1)A_{n-1}$ and (4) becomes

$$n(A_n - G_n) \geqq (n-1)(A_{n-1} - G_{n-1}), \tag{5}$$

* Received by the editors September 16, 1985; accepted for publication January 7, 1986.

† Department of Mathematics, University of California, Los Angeles, California 90024. This author was a guest professor at the Mathematisches Institut I, Universität Karlsruhe, Karlsruhe, West Germany, under auspices of the U.S. Special Program, Alexander von Humboldt-Stiftung.

‡ Department of Mathematics, Universität Karlsruhe, Karlsruhe, West Germany.

which is one form of Rado's inequality. As indicated in [2], the latter provides the basis for an inductive proof of Maclaurin's inequality $G_n \leqq A_n$, with full control of the cases of equality. Addition of the results (4) gives the following, which is also known as Rado's inequality:

$$\sum_{m+1}^{n} a_k \geqq nG_n - mG_m, \qquad n > m \geqq 1.$$

It is seen below that (4) leads to Theorem 1 and that is the reason why these familiar results have been reviewed above. Of course, (3) can be obtained by noting that the function $f(x) = x^n - nx$ satisfies $f''(x) \geqq 0, f'(1) = 0$, and hence $f(x) \leqq f(1)$. The proof based on Bernoulli's inequality has been preferred here because of its historical interest.

3. Proof of Theorem 1. When the series converge we set

$$E = a_2 + a_4 + a_6 + \cdots, \quad U = a_1 + a_3 + a_5 + \cdots, \quad G = G_1 - 2G_2 + 3G_3 - 4G_4 + \cdots,$$

and we denote the partial sums by $E(m)$, $U(m)$ and $G(m)$, respectively, where the summation stops at index m. Hence m is even, odd, and indifferent for E, U, G, respectively. By (4)

$$a_3 \geqq 3G_3 - 2G_2, \quad a_5 \geqq 5G_5 - 4G_4, \quad a_7 \geqq 7G_7 - 6G_6,$$

and so on. Upon recalling that $a_1 = G_1$ and adding, we get $U(m) \geqq G(m)$ for all odd integers m. By § 2 we have equality if, and only if,

$$G_3 = G_2, \quad G_5 = G_4, \quad G_7 = G_6,$$

and so on. The equation $G_{2k+1} = G_{2k}$ is equivalent to

$$(6) \qquad (a_{2k+1})^{2k} = a_1 a_2 \cdots a_{2k}.$$

Hence, we can prescribe the a_{2k} arbitrarily and the a_{2k+1} are then determined uniquely.

Equation (6) gives $G_{2k} = a_{2k+1}$ and the series G when $G_{2k} = G_{2k+1}$ reduces to the following *without parentheses*:

$$a_1 + (-2a_3 + 3a_3) + (-4a_5 + 5a_5) + \cdots.$$

The series *with parentheses* converges by hypothesis, and the parentheses can be dropped if, and only if, $(2k+1)a_{2k+1} \to 0$. This gives Theorem 1.

4. A two-sided inequality. Adding (4) for even values of m gives $-E(m) \leqq G(m)$ for m even, and equality is equivalent to

$$(a_{2k})^{2k-1} = a_1 a_2 \cdots a_{2k-1}, \qquad k = 1, 2, \cdots, m/2.$$

Here we can prescribe a_{2k-1} arbitrarily and the a_{2k} are then uniquely determined. The equality $-E = G$ is possible if, and only if, $a_{2j} = o(1/j)$. Letting $m \to \infty$ and ignoring the cases of equality, we are led to the following theorem:

THEOREM 2. *Let $a_1 + a_2 + a_3 + \cdots < \infty$. Then the series G is convergent and satisfies $-E \leqq G \leqq U$.*

The only remaining problem is to establish the convergence. To this end let $\varepsilon > 0$ be given and choose m so large that

$$a_m + a_{m+1} + a_{m+2} + \cdots < \varepsilon.$$

With m fixed, let

$$P = P_n = a_{m+1} a_{m+2} \cdots a_n, \qquad n > m.$$

By Maclaurin's inequality between arithmetic and geometric means

$$P<\left(\frac{\varepsilon}{n-m}\right)^{n-m}.$$

Raising both sides to power $1/n$ we get an inequality which implies

$$\limsup_{n\to\infty} nP^{1/n} \leqq \varepsilon.$$

This gives $\limsup nG_n \leqq \varepsilon$ and hence $nG_n \to 0$ as $n\to\infty$. If we use this fact, and apply the method of § 3 to a segment

$$a_{m+1}+a_{m+2}+\cdots+a_n,$$

we find that the series G satisfies the Cauchy criterion, hence converges.

The proof of Theorem 2 shows that the simultaneous equalities $G(2n)=-E(2n)$, $G(2n+1)=U(2n+1)$ are possible only if all a_j are equal.

5. The problem of convergence. That G may converge without convergence of E or U is shown by the following remark, whose proof is left to the reader: Let S be an infinite subset of N and let $a_j>0$ be arbitrarily prescribed for $j\in N-S$. Then we can determine $a_j>0$ with $j\in S$ in such a way that the series G converges. Although convergence of both E and U ensures convergence of G, as seen in Theorem 2, difficulties remain when convergence of E or U alone is postulated. Here we give a sufficient condition which is related to Theorem 1:

THEOREM 3. *Let the series U be convergent, and suppose there are positive constants m, M such that*

$$ma_{2k+1}\leqq G_{2k}\leqq Ma_{2k+1}, \qquad k\in N.$$

Then G converges if, and only if, $a_{2k+1}=o(1/k)$.

The special case $m=M=1$ reduces to the case of equality in Theorem 1.

If G converges then $kG_{2k+1}\to 0$ and the condition $ka_{2k+1}\to 0$ follows from

$$ka_{2k+1}=kG_{2k+1}(a_{2k+1}/G_{2k})^{2k/(2k+1)}\leqq kG_{2k+1}m^{-2k/(2k+1)}.$$

The main difficulty is in the converse.

Suppose, then, that $a_{2k+1}=o(1/k)$. We define b_{2k+1} by

$$a_{2k+1}b_{2k+1}=-2kG_{2k}+(2k+1)G_{2k+1} \tag{7}$$

and note that (4) gives $b_{2k+1}\leqq 1$. A lower bound can be obtained from the identity

$$b_{2k+1}=(G_{2k}/a_{2k+1})^{2k/(2k+1)}(2k+1-2k(G_{2k}/a_{2k+1})^{1/(2k+1)})$$

together with $0<G_{2k}/a_{2k+1}\leqq M$ and the limit

$$\lim_{k\to\infty}(2k+1-2kM^{1/(2k+1)})=1-\log M,$$

whose verification is left to the reader. Hence, there is a constant C such that $|b_{2k+1}|\leqq C$. Referring to (7) we see that

$$\sum_{k=1}^{\infty}|-2kG_{2k}+(2k+1)G_{2k+1}|$$

converges by the comparison test. This establishes existence of $\lim G(2n-1)$ as $n\to\infty$. The convergence of G now follows from

$$nG_{2n}\leqq Mna_{2n+1}\to 0, \qquad n\to\infty.$$

In a similar fashion, if E converges, and

$$ma_{2k} \leqq G_{2k-1} \leqq Ma_{2k}$$

for positive constants m, M, then G converges if, and only if, $a_{2k} = o(1/k)$.

6. Historical note. The inequality $G_n \leqq A_n$ is attributed by Hardy, Littlewood and Pólya [2] to Maclaurin and we have followed this attribution here. By a somewhat different procedure Rado's inequality (5) is also derived in [2] and is used to give an inductive proof of $G_n \leqq A_n$. Another derivation of Rado's inequality was given by Jacobsthal [3] in 1951, which is to say, 17 years after the appearance of the first edition [2]. Here the starting point is the identity

$$A_n = \frac{G_{n-1}}{n}\left((n-1)\frac{A_{n-1}}{G_{n-1}} + \left(\frac{G_n}{G_{n-1}}\right)^n\right)$$

and (5) follows by the choice $x = G_n/G_{n-1}$ in (3). This is the choice we made to get the equivalent equation (4), though the latter is not mentioned in [1], [2], [3]. We believe that the particular sequence of events in § 2 may have some advantage over other treatments but, because of overlap with [1], [2], [3], this part of our paper should be regarded as expository.

Theorem 2 is deduced in [5] from a general inequality, of which $-E \leqq G$ and $G \leqq U$ are extremely special cases. The proof of convergence following Theorem 2 was outlined in [4], but so briefly that the argument caused difficulty for some readers. The missing details have been supplied here. We have not come across any prior statement of Theorems 1 or 3.

REFERENCES

[1] E. F. Beckenbach and Richard Belleman, *Inequalities*, Ergebnisse der Math., 30 (1983), pp. 11-12.

[2] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press, London, 1952.

[3] Ernst Jacobsthal, *Über das arithmetische und geometrische Mittel*, Norske vid. Selsk. Forh., 25 (1951), p. 122.

[4] R. M. Redheffer, *Recurrent inequalities*, Proc. London Math. Soc., 17 (1967), pp. 683-699.

[5] ———, *Easy proofs of hard inequalities*, in General Inequalities 3, E. F. Beckenbach and W. Walter, eds., Birkhäuser, Basel, Boston, Stuttgart, 1983, pp. 123-140.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

018

NEW INEQUALITIES OF MARKOV TYPE*

P. DÖRFLER†

Abstract. For any polynomial f with complex coefficients we define

$$\|f\|:=\left\{\int_a^b |f(t)|^2 w(t)\,dt\right\}^{1/2},$$

where $w:(a,b)\to R$ is a positive and integrable function with all moments finite. It is well known that there exists a constant γ_n, not depending on f, such that $\|f'\|\leqq \gamma_n\|f\|$ for all f, $\deg f\leqq n$. In the present paper we consider the analogous inequality for derivatives of higher order and compute the best possible γ_n. This constant turns out to be the largest singular value of a certain matrix. Two examples are given.

Key words. Markov inequality, orthogonal polynomials, matrix norm

AMS(MOS) subject classifications. 33A65, 41A17, 41A44

1. Let $-\infty\leqq a<b\leqq\infty$ and denote by $w:(a,b)\to R$ a positive and integrable function with all moments

$$\int_a^b t^n w(t)\,dt, \qquad n\geqq 0$$

finite. For any polynomial f with complex coefficients we define

$$\|f\|:=\left\{\int_a^b |f(t)|^2 w(t)\,dt\right\}^{1/2}. \tag{1}$$

In [3] Mirsky showed that there exists a constant γ_n, not depending on f, such that

$$\|f'\|\leqq \gamma_n\|f\| \tag{2}$$

for every polynomial f, $\deg f\leqq n$.

In this paper we show that the best possible value for γ_n is the largest singular value of a certain matrix. Moreover, (2) will be generalized to derivatives of higher order. Finally, the method which yields the best possible γ_n will be illustrated by two examples.

2. First of all we introduce orthogonal polynomials with reference to the papers of Bellman [1] and Mirsky [3]. Under the above-mentioned assumptions there exists, according to [6], a sequence of real orthonormal polynomials p_n, $n\geqq 0$, associated with the weight function $w(t)$, i.e.,

$$\int_a^b p_n p_m w\,dt=\delta_{nm}, \qquad n,m\geqq 0. \tag{3}$$

If (2) is valid for real polynomials, the same is true for complex ones since $f=g+ih$, g, h real, implies $\|f\|^2=\|g\|^2+\|h\|^2$. Therefore, let f be a real polynomial of degree n. The (unique) representations

$$f(t)=\sum_{k=0}^{n} c_k p_k(t), \quad f^{(r)}(t)=\sum_{m=0}^{n-r} d_m p_m(t), \quad 0\leqq r\leqq n,$$

imply

$$\|f\|^2=\sum_{k=0}^{n} c_k^2, \qquad \|f^{(r)}\|^2=\sum_{m=0}^{n-r} d_m^2 \tag{4}$$

* Received by the editors February 5, 1985, and in revised form November 25, 1985.

† Institut für Mathematik und Angewandte Geometrie, Montanuniversität Leoben, A-8700 Leoben, Austria.

because of (3). On the other hand,

$$\sum_{k=0}^{n} c_k p_k^{(r)}(t) = \sum_{m=0}^{n-r} d_m p_m(t)$$

yields

$$(5) \qquad \sum_{k=0}^{n} c_k e_{kj}^{(r)} = d_j, \qquad 0 \leqq j \leqq n-r,$$

where

$$e_{kj}^{(r)} := \int_a^b p_k^{(r)} p_j w \, dt.$$

Let $c^t := (c_0, \cdots, c_n) \in R^{n+1}$, $d^t := (d_0, \cdots, d_{n-r}) \in R^{n-r+1}$ and

$$(6) \qquad A_n^{(r)} := \begin{pmatrix} e_{00}^{(r)} & \cdots & e_{n0}^{(r)} \\ \vdots & & \vdots \\ e_{0n-r}^{(r)} & \cdots & e_{nn-r}^{(r)} \end{pmatrix}.$$

According to these definitions, (5) can be interpreted as the linear transformation

$$A_n^{(r)} c = d.$$

Since $\|f\| = |c|$ and $\|f^{(r)}\| = |d|$, where $|c|$, $|d|$ are the Euclidean vector norms in R^{n+1} and R^{n-r+1}, respectively (cf. (4)), we have the following problem:

Which is the best possible constant $\gamma_n^{(r)}$ such that

$$|A_n^{(r)} c| \leqq \gamma_n^{(r)} |c| \quad \forall c \in R^{n+1} \, ?$$

Obviously, this constant is the best possible in

$$\|f^{(r)}\| \leqq \gamma_n^{(r)} \|f\|.$$

The answer to this question is well known [5]:

$\gamma_n^{(r)}$ is the 2-norm of $A_n^{(r)}$. This norm is the largest singular value of $A_n^{(r)}$, which is the square root of the largest eigenvalue of $(A_n^{(r)})^t A_n^{(r)}$.

Let $A = (a_{ik}) \in R^{m \times n}$ be any real matrix. Then the Frobenius norm

$$(7) \qquad \|A\|_F := \left\{ \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik}^2 \right\}^{1/2}$$

provides an upper bound for the 2-norm of A [5]. Consequently we obtain an upper bound for $\gamma_n^{(r)}$ in this way which, moreover, can be expressed easily in terms of the p_j. Since

$$p_n^{(r)} = \sum_{m=0}^{n-r} e_{nm}^{(r)} p_m, \qquad n \geqq 0,$$

it follows that

$$\|A_n^{(r)}\|_F^2 = \sum_{j=0}^{n} \|p_j^{(r)}\|^2.$$

Together with the trivial lower bound for $\gamma_n^{(r)}$,

$$\max_{0\leqq j\leqq n} \|p_j^{(r)}\| \leqq \gamma_n^{(r)},$$

this gives the estimation

$$\max_{0\leqq j\leqq n} \|p_j^{(r)}\| \leqq \gamma_n^{(r)} \leqq \left\{ \sum_{j=0}^{n} \|p_j^{(r)}\|^2 \right\}^{1/2} \tag{8}$$

which, restricted to $r=1$, is better than that given in [3]. Let us sum up now the results.

THEOREM 1. *Let* $\|f\|$, p_j, $A_n^{(r)}$ *be defined as in* (1), (3) *and* (6), *and let* f *be any polynomial with complex coefficients,* $\deg f \leqq n$, $n\geqq 0$. *Then the best possible constant* $\gamma_n^{(r)}$ *such that*

$$\|f^{(r)}\| \leqq \gamma_n^{(r)}\|f\|, \tag{9}$$

is the largest singular value of $A_n^{(r)}$. *Moreover, estimation* (8) *holds.*

Example 1. We consider the case $a=-\infty$, $b=\infty$, $w(t)=e^{-t^2}$. A corresponding system of orthonormal polynomials is

$$p_n(t)=\{\pi^{1/2}2^n n!\}^{-1/2}H_n(t), \qquad n\geqq 0,$$

where $H_n(t)$ denotes the nth Hermite polynomial [6]. Since $H_n'=2nH_{n-1}$, $n\geqq 1$, it is easy to prove by induction that

$$p_n^{(r)}=\left\{2^r r!\binom{n}{r}\right\}^{1/2} p_{n-r} \tag{10}$$

holds for $0\leqq r\leqq n$. Consequently,

$$e_{kj}^{(r)}=\begin{cases} 0 & \text{if } k\neq r+j,\\ \left\{2^r r!\binom{k}{r}\right\}^{1/2} & \text{if } k=r+j. \end{cases}$$

Obviously, the largest eigenvalue of $(A_n^{(r)})^t A_n^{(r)}$ is $2^r r!\binom{n}{r}$ and therefore, by Theorem 1

$$\gamma_n^{(r)}=\left\{2^r r!\binom{n}{r}\right\}^{1/2}.$$

By (10) we have $\|p_n^{(r)}\|=\gamma_n^{(r)}$. Thus, in (8) the equality sign holds on the left-hand side. Moreover, (9) becomes an equality for the polynomials $p_n(t)$ themselves.

These results, as far as they concern the case $r=1$, were also proved by Mirsky [3] by a straightforward computation.

3. We want to give a further application of Theorem 1 which, however, requires some additional considerations. We need the following lemma.

LEMMA. *Let* $A=(a_{ik})\in R^{m\times n}$ *be any real matrix with column sums* $s_k := \sum_{i=1}^m a_{ik}$ *and row sums* $z_i := \sum_{k=1}^n a_{ik}$. *Then*

$$\sigma_A^2 \geqq \max\left\{\frac{1}{n}\sum_{i=1}^m z_i^2, \frac{1}{m}\sum_{k=1}^n s_k^2\right\},$$

where σ_A *denotes the largest singular value of* A.

Proof. σ_A^2 is the largest eigenvalue of $M := A^tA$. Since M is normal the following estimation holds [4]:

$$\sigma_A^2 \geqq \frac{1}{l}\max|S_l|,$$

where S_l is the sum of the elements in any $l\times l$ principal submatrix of M. The representation $m_{ik}=\sum_{j=1}^{m} a_{ji}a_{jk}$ of the elements of M yields

$$S_n=\sum_{i,k=1}^{n} m_{ik}=\sum_{i,k=1}^{n}\sum_{j=1}^{m} a_{ji}a_{jk}=\sum_{j=1}^{m}\sum_{i,k=1}^{n} a_{ji}a_{jk}=\sum_{j=1}^{m} z_j^2.$$

Thus, for the special choice $l=n$, we have obtained the estimation

$$\sigma_A^2\geqq\frac{1}{n}S_n=\frac{1}{n}\sum_{j=1}^{m} z_j^2.$$

Since σ_A is the 2-norm of A, which is invariant under transposition, the assertion is proved. □

Example 2. Put $w(t)=e^{-t}t^{\alpha}$, $\alpha\in R$, $\alpha>-1$, $a=0$, $b=\infty$. A corresponding system of orthonormal polynomials is

$$p_n(t)=\left\{\frac{n!}{\Gamma(n+\alpha+1)}\right\}^{1/2} L_n^{\alpha}(t),\qquad n\geqq 0,$$

where $L_n^{\alpha}(t)$ denotes the nth Laguerre polynomial [6]. As one can prove easily, the recurrence formula [2, p. 109]

$$\frac{d}{dt}L_n^{\alpha}-\frac{d}{dt}L_{n-1}^{\alpha}+L_{n-1}^{\alpha}=0,\qquad n\geqq 1$$

leads to

$$p_n'=c(p_{n-1}'-p_{n-1}),\quad n\geqq 1,\quad c:=(1+\alpha/n)^{-1/2}. \tag{11}$$

Hence, for $k\geqq 1$, $j\geqq 0$,

$$e_{kj}^{(1)}=\begin{cases}-c & \text{if } k=j+1,\\ ce_{k-1j}^{(1)} & \text{if } k\neq j+1,\end{cases}$$

and so

$$A_n:=-A_n^{(1)}=\begin{pmatrix} 0 & c & c^2 & \cdots & c^n\\ \vdots & \ddots & c & \cdots & c^{n-1}\\ \vdots & & \ddots & \ddots & \vdots\\ \vdots & & & \ddots & \vdots\\ 0 & \cdots & \cdots & 0 & c\end{pmatrix}.$$

An upper bound for $\gamma_n^{(1)}$ is $\|A_n^{(1)}\|_F=\|A_n\|_F$ (cf. (7)). A short calculation yields

$$(\gamma_n^{(1)})^2\leqq\|A_n\|_F^2=\frac{c^2}{c^2-1}\left(c^2\frac{c^{2n}-1}{c^2-1}-n\right). \tag{12}$$

Substituting for c in (12) and observing that $(1+\alpha/n)^n\geqq 1+\alpha$, we obtain the "nicer" estimation

$$\gamma_n^{(1)}\leqq\|A_n\|_F\leqq(1+\alpha)^{-1/2}n,$$

which, moreover, implies that $\gamma_n^{(1)}=O(n)$.

Next we present a lower bound for $\gamma_n^{(1)}$ by applying the above lemma to A_n. Let s_j, $0\leqq j\leqq n$, denote the column sums of A_n. Then, by a short calculation,

$$(\gamma_n^{(1)})^2\geqq\frac{1}{n}\sum_{j=1}^{n} s_j^2=\frac{c^2}{n(c-1)^3}\left[c^2\frac{c^{2n}-1}{c+1}-2c(c^n-1)+n(c-1)\right]. \tag{13}$$

Finally we show that $\gamma_n^{(1)}/n > K$ for some positive constant K as $n \to \infty$. We divide the right-hand side of (13) by n^2, substitute for c and take the limit as $n \to \infty$. This yields

$$g(\alpha) := 4(4e^{-\alpha/2} - e^{-\alpha} + \alpha - 3)/\alpha^3,$$

which obviously is positive for all $\alpha > -1$. (If $\alpha = 0$, then $g(0) = \lim_{\alpha \to 0} g(\alpha) = \frac{1}{3}$.) With respect to Theorem 1 we thus have proved the following.

THEOREM 2. *Let f be any polynomial with complex coefficients,* $\deg f \leqq n$, $n \geqq 0$. *Let* $\|f\|$ *be defined by* (1) *with* $a = 0$, $b = \infty$, $w(t) = e^{-t}t^\alpha$, $\alpha \in R$, $\alpha > -1$. *Then we have*

$$\|f'\| \leqq (1+\alpha)^{-1/2} n \|f\|.$$

The exponent of n in this inequality cannot be replaced by a smaller one for arbitrary n.

In Example 2 it is difficult to compute the exact value of $\gamma_n^{(r)}$ for arbitrary n and r. The estimates (12) and (13) for $\gamma_n^{(1)}$, however, are sharp enough to give the right order as $n \to \infty$. In the case $\alpha = 0$ the bounds obtained by taking the limits in (12) and (13) as $\alpha \to 0$ have the form

$$\left\{\frac{(2n+1)(n+1)}{6}\right\}^{1/2} \leqq \gamma_n^{(1)} \leqq \left\{\frac{n(n+1)}{2}\right\}^{1/2}.$$

REFERENCES

[1] R. BELLMAN, *A note on an inequality of E. Schmidt*, Bull. Amer. Math. Soc., 50 (1944), pp. 734-736.

[2] N. N. LEBEDEW, *Spezielle Funktionen und ihre Anwendungen*, Bibliographisches Institut, Mannheim, 1973.

[3] L. MIRSKY, *An inequality of the Markov-Bernstein type for polynomials*, this Journal, 14 (1983), pp. 1004-1008.

[4] W. V. PARKER, *Sets of complex numbers associated with a matrix*, Duke Math. J., 15 (1948), pp. 711-715.

[5] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[6] G. SZEGÖ, *Orthogonal polynomials*, AMS. Coll. Pub. Vol. XXIII, Amer. Math. Soc., Providence, RI, 1939.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

019

A WHIPPLE'S TRANSFORMATION FOR HYPERGEOMETRIC SERIES IN $U(n)$ AND MULTIVARIABLE HYPERGEOMETRIC ORTHOGONAL POLYNOMIALS*

R. A. GUSTAFSON†

Abstract. A generalization of the terminating form of Whipple's transformation is given for hypergeometric series in $U(n)$. This is proved by explicit evaluation of a generalized Biedenharn-Elliott identity for multiplicity-free $U(n)$ Racah coefficients. As a corollary, a Dougall's theorem is obtained for hypergeometric series in $U(n)$. A family of summation theorems for classical ordinary and basic hypergeometric series is also proved. Finally, a family of orthogonal polynomials on the discrete set $x_1+\cdots+x_n=N$, where the x_i and N are nonnegative integers, is defined which generalizes the (discrete) Racah polynomials of Wilson. We prove a recurrence relation, duality theorem and an identity similar to an addition theorem for these generalized Racah polynomials.

Key words. hypergeometric series in $U(n)$, multivariable orthogonal polynomials, representation theory of $U(n)$, Whipple's transformation, Dougall's transformation

AMS(MOS) subject classification. 33A75

Introduction. In 1976, Holman, Biedenharn and Louck [15] defined a multivariable generalization of classical well-poised hypergeometric series which they called "well-poised in $SU(n)$", $n\geqq 2$. These new special functions were closely connected to the multiplicity-free Racah and Wigner coefficients for $SU(n)$ (see §§ 1 and 2 below). They showed that the hypergeometric series well-poised in $SU(n)$ satisfied a generalization of the terminating ${}_4F_3(-1)$ summation theorem and gave an $SU(3)$ analogue of Whipple's theorem [4, eq. (4.3.4)]. In 1980, Holman [14] defined a general "hypergeometric series in $U(n)$" and proved generalizations of the Saalschütz summation theorem, a Vandermonde (or Gauss) summation theorem and the terminating form of the ${}_5F_4(1)$ summation theorem. These results were proved by explicitly computing identities involving multiplicity-free Racah and Wigner coefficients for $U(n)$ (or $SU(n)$). However, it remained an outstanding problem to find a true generalization of Whipple's theorem to hypergeometric series in $U(n)$.

The importance of finding such a generalized Whipple's theorem had been noted by Andrews [3] and Milne [20]. They hoped that if there were such a generalized Whipple's theorem, then there would also be a q-analogue of it. In the classical case, Watson's q-analogue of Whipple's theorem [24] had been used to prove several partition identities including the Rogers-Ramanujan identities (Watson [24]). We remark that in a different direction Andrews [3] has given a generalization of Watson's q-analogue of Whipple's theorem.

In this paper we prove a generalization of Whipple's theorem for hypergeometric series in $U(n)$. We first prove in § 1 (Prop. 1.50) a generalization of the Biedenharn-Elliottt (B-E) identity for certain multiplicity-free $U(n)$ Racah coefficients. Generalizations of the B-E identity have been previously stated and proved by more than one author [10], [19]. In § 2 we explicitly compute a degenerate case of this B-E identity and obtain a generalization of Whipple's theorem (Thm. 2.24). For the classical case $n=2$ this proof is already given in [15].

* Received by the editors April 22, 1985, and in final form September 13, 1985.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843, and School of Mathematics, Tata Institute of Fundamental Research, Bombay 400 005, India. The work of this author was supported by an Indo-American Fellowship.

As corollaries of Theorem 2.24, we obtain in § 3 several other summation and transformation theorems for hypergeometric series in $U(n)$. In particular we obtain generalizations of Dougall's theorem (Cor. 3.1), an analogue of Dougall's theorem (Cor. 3.4) and a terminating form of Whipple's ${}_6F_5(-1)$ transformation (Cor. 3.8 and 3.10). Finally, we prove a family of summation theorems for classical terminating ${}_{n+1}F_n(1)$ (Thm 3.13) and their q-analogues, ${}_{n+1}\varphi_n(1)$ (Thm. 3.18), for $n \geqq 1$.

In § 4 we define a family of orthogonal polynomials in several variables (Prop. 4.5) which are orthogonal on the discrete set $x_1 + \cdots + x_n = N$, where the x_i, $1 \leqq i \leqq n$, and N are nonnegative integers and $n \geqq 2$. In §§ 5, 6 and 7 a recurrence relation, duality theorem and an identity similar to an addition theorem are proved for these generalized Racah polynomials.

Finally, in the appendix we prove the key technical Proposition A.16 which is needed in the explicit computation of the special case of the B-E identity used in the proof of Theorem 2.24.

The approach we have taken in this paper relies on the algebraic properties of the Racah coefficients or the "Racah-Wigner algebra of tensor operators" (see [8]). In fact, the key B-E identity is a consequence of the associativity law in the Racah-Wigner algebra [8], [19].

Another complementary approach to the transformation identities for hypergeometric series in $U(n)$ is by means of difference equations. By this method Milne [20] gave an elementary proof of Holman's ${}_5F_4(1)$ summation theorem.

The results in this paper involve only the multiplicity-free Racah coefficients for $U(n)$. A better understanding of the nonmultiplicity-free Racah coefficients should lead to further generalizations of the transformation identities discussed here and to new orthogonality relations.

1. Vector coupling coefficients, recoupling coefficients and the Biedenharn–Elliott identity. We shall recall some facts about Wigner coefficients for $U(n)$, $n \geqq 1$, which are expounded in greater length in [6]. The irreducible representations of $U(n)$ are in one-to-one correspondence with the set of n-tuples of integers $m = [m] = [m]_n = [m_{1n}, m_{2n}, \cdots, m_{nn}]$ satisfying

$$m_{1n} \geqq m_{2n} \geqq \cdots \geqq m_{nn}. \tag{1.1}$$

m is the highest weight of the irreducible representation π_m of $U(n)$ acting on the representation space V_m. The Gelfand-Zetlin basis (over $\mathbb{C}$) of V_m is defined in [17] or [6]. A Gelfand pattern $(m) = (m)_n$ is an array of integers $m_{k,l}$, $1 \leqq k \leqq l \leqq n$, satisfying the "betweenness" conditions

$$m_{k,l+1} \leqq m_{k,l} \leqq m_{k+1,l+1} \tag{1.2}$$

for $1 \leqq k \leqq l \leqq n-1$. The first row of the Gelfand pattern (m) is $m = [m_{1n}, \cdots, m_{nn}]$. To each Gelfand pattern (m) is associated an orthonormal basis vector in V_m, which is also denoted by (m). The vector (m) is called a "Gelfand state" and the set of all Gelfand states (m) with first row m is an orthonormal basis of V_m.

The weight $\Delta(m) = [\Delta_1, \cdots, \Delta_n] = [\Delta]$ of a Gelfand state (m) is given by

$$\Delta_l = \sum_{k=1}^{l} m_{kl} - \sum_{k=1}^{l-1} m_{kl-1} \tag{1.3}$$

for $2 \leqq l \leqq n$ and $\Delta_1 = m_{11}$. The set of weights satisfies a natural partial ordering:

$$\Delta \geqq \Delta' \quad \text{if and only if } \sum_{i=1}^{l} \Delta_i \geqq \sum_{i=1}^{l} \Delta'_i \tag{1.4}$$

for all l, $1 \leqq l \leqq n$. There are unique states in V_m of highest weight $[m_{1n}, \cdots, m_{nn}]$ and lowest weight $[m_{nn}, m_{nn-1}, \cdots, m_{1n}]$.

We are interested in the explicit decomposition of tensor products of the form $V_m \otimes V_{m'}$ or $V_{m''} \otimes V_m \otimes V_{m'}$ where m is an arbitrary n-tuple of integers satisfying (1.1) and m', m'' are of the form $[p', 0, \cdots, 0]$ and $[p'', 0, \cdots, 0]$ respectively, where p' and p'' are nonnegative integers.

We have

$$V_m \otimes V_{m'} \approx \oplus V_\mu \tag{1.5}$$

where the sum is over all n-tuples of integers $\mu = [\mu_{1n}, \cdots, \mu_{nn}]$ satisfying

$$\mu_{1n} \geqq m_{1n} \geqq \mu_{2n} \geqq m_{2n} \geqq \cdots \geqq \mu_{nn} \geqq m_{nn} \tag{1.6a}$$

and

$$\sum_{i=1}^{n} (\mu_{in} - m_{in}) = p'. \tag{1.6b}$$

A space V_μ on the left-hand side of (1.5) is said to *occur* in $V_m \otimes V_{m'}$.

The set of all pairs $(m) \otimes (m')$ where (m), (m') are Gelfand states of V_m and $V_{m'}$ respectively forms an orthonormal basis, called the Gelfand-Zetlin basis, of $V_m \otimes V_{m'}$. Similarly the set of all triples $(m'') \otimes (m) \otimes (m')$ forms an orthonormal basis (Gelfand-Zetlin basis) for $V_{m''} \otimes V_m \otimes V_{m'}$ where (m''), (m), (m') are Gelfand states of $V_{m''}$, V_m and $V_{m'}$. We shall call these vectors Gelfand states of $V_m \otimes V_{m'}$ and $V_{m''} \otimes V_m \otimes V_{m'}$ respectively.

For each highest weight μ satisfying (1.6a and b) let T_μ be a nontrivial intertwining map

$$T_\mu : V_m \otimes V_{m'} \to V_\mu. \tag{1.7}$$

The map T_μ is determined up to scalar multiples. We shall assume that the restriction of T_μ to the orthogonal complement of the kernel is unitary and also that T_μ has real matrix coefficients with respect to the Gelfand-Zetlin bases of $V_m \otimes V_{m'}$ and V_μ (see [6, § 2]). Under these conditions T_μ is determined up to a scalar factor (phase factor) of ±1. This phase factor is then fixed by convention.

We choose the following phase convention. Let

$$(m) = \begin{pmatrix} [m]_n \\ (m)_{n-1} \end{pmatrix}$$

be a Gelfand state of V_m of highest weight $\Delta(m) = [m_{1n}, \cdots, m_{nn}] = m$. The first row of (m) is $[m]_n = m$ and the last $n-1$ rows of (m) are $(m)_{n-1}$, which is a Gelfand pattern for $U(n-1)$. Similarly let

$$(\mu) = \begin{pmatrix} [\mu]_n \\ (m)_{n-1} \end{pmatrix}$$

be a state of V_μ whose first row is $[\mu]_n = \mu$ and last $n-1$ rows are identical to the last $n-1$ rows of (m). Since the first row of $(m)_{n-1}$ is $[m_{1n}, m_{2n}, \cdots, m_{n-1n}]$, then it follows from (1.6a) that (μ) satisfies the betweenness conditions (1.2). Finally let

$$(m') = \begin{pmatrix} [p', 0, \cdots, 0] \\ (0)_{n-1} \end{pmatrix}$$

be the state of lowest weight in $V_{m'}$. Then we require that

$$\langle(\mu), T_\mu(m \otimes (m'))\rangle > 0, \tag{1.8}$$

where $\langle , \rangle$ is the Hermitian inner product on V_μ for which the Gelfand-Zetlin basis is orthonormal. That the inner product (1.8) is nonzero is a consequence of [14, eq. (6)]. This phase convention agrees with that of Chacón, Ciftan and Biedenharn [9] and also Arišaukas, Jucys and Jucys [1] (see also [14]).

Remark 1.9. For the sake of simplicity the dependence on the highest weights m and m' in T_μ will not be denoted, but understood. Thus two distinct operators whose domains are different spaces $V_m \otimes V_{m'}$ but whose images are the same space V_μ will both be denoted by T_μ. Also under essentially the same conditions as above and with the same notation, we construct the $U(n)$ intertwining map:

$$T_\mu : V_{m'} \otimes V_m \to V_\mu. \tag{1.10}$$

Definition 1.11. Let notation be as in (1.5). For arbitrary (m), (m') and (μ) in the representation spaces V_m, $V_{m'}$ and V_μ respectively, the inner product

$$\langle (\mu), T_\mu((m) \otimes (m')) \rangle \tag{1.12}$$

is called a *vector-coupling, Wigner* or *Clebsch-Gordan* coefficient of $U(n)$. These coefficients are simply the matrix coefficients of the intertwining map T_μ.

As a consequence of the definition of the maps T_μ and as discussed in [6, eqs. (2.8a and b)] (with different notation), we have the following orthogonality relations satisfied by the Wigner coefficients:

Proposition 1.13. *With notation as above, let (m), $(\bar{m})$ be fixed states of V_m and (m'), $(\bar{m}')$ be fixed states of $V_{m'}$; then*

$$\sum_{(\mu)} \langle (\mu), T_\mu((m) \otimes (m')) \rangle \langle (\mu), T_\mu((\bar{m}) \otimes (\bar{m}')) \rangle = \begin{cases} 1 & \text{if } (m) = (\bar{m}) \quad \text{and} \quad (m') = (\bar{m}'), \\ 0 & \text{otherwise,} \end{cases} \tag{1.14a}$$

where the sum is over all states (μ) of all distinct V_μ occurring in $V_m \otimes V_{m'}$ (as in (1.5)). Similarly if (μ), (μ') are fixed states of V_μ and $V_{\mu'}$, respectively, occurring in $V_m \otimes V_{m'}$, then

$$\sum_{(m) \otimes (m')} \langle (\mu), T_\mu((m) \otimes (m')) \rangle \langle (\mu'), T_{\mu'}((m) \otimes (m')) \rangle = \begin{cases} 1 & \text{if } \mu = \mu' \quad \text{and} \quad (\mu) = (\mu'), \\ 0 & \text{otherwise,} \end{cases} \tag{1.14b}$$

where the sum is over all Gelfand states $(m) \otimes (m')$ of $V_m \otimes V_{m'}$.

We now consider decompositions of the triple tensor product $V_{m''} \otimes V_m \otimes V_{m'}$ with m an arbitrary $U(n)$ highest weight and m', m'' of the form $[p', 0, \cdots, 0]$ and $[p'', 0, \cdots, 0]$ respectively, with p' and p'' nonnegative integers. We first construct $U(n)$ intertwining maps of the form

$$1 \otimes T_\mu : V_{m''} \otimes (V_m \otimes V_{m'}) \to V_{m''} \otimes V_\mu \tag{1.15a}$$

where V_μ occurs in $V_m \otimes V_{m'}$ (as in (1.5)) and T_μ is defined as in (1.7); and also

$$T_\nu : V_{m''} \otimes V_\mu \to V_\nu \tag{1.15b}$$

where V_ν occurs in $V_{m''} \otimes V_\mu$ and T_ν is defined as above.

Proposition 1.16. *For a fixed highest weight ν such that V_ν occurs in $V_{m''} \otimes V_m \otimes V_{m'}$ (i.e.* $\mathrm{Hom}_{U(n)}(V_{m''} \otimes V_m \otimes V_{m'}, V_\nu) \neq 0$), *then a basis over* $\mathbb{C}$ *for* $\mathrm{Hom}_{U(n)}(V_{m''} \otimes V_m \otimes V_{m'}, V_\nu)$ *is given by the set I of all $U(n)$ intertwining maps of the form*

$$T_\nu(1 \otimes T_\mu) : V_{m''} \otimes V_m \otimes V_{m'} \to V_\nu \tag{1.17}$$

for all highest weights μ such that V_μ occurs in $V_m \otimes V_{m'}$ and V_ν occurs in $V_{m''} \otimes V_m$.

Proof. One checks that the dimension over $\mathbb{C}$ of Hom $(V_{m''}\otimes V_m\otimes V_{m'}, V_\nu)$ equals the order of I. This can be done for example by a $U(n)$ character computation. Second, we show that the elements of I are linearly independent over $\mathbb{C}$. Let N and M be integers such that there exists states $(m'')_l\in V_{m''}$, $(m)_{k,l}\in V_m$ and $(m')_{k,l}\in V_{m'}$ for $k=1,\cdots,N$ and $l=1,\cdots,M$ such that

$$(1.18a)\qquad \sum_{k=1}^{N}(m)_{k,l}\otimes(m')_{k,l}\in V_\mu \quad \text{component of } V_m\otimes V_{m'}$$

for all l, $1\leqq l\leqq M$, and

$$(1.18b)\qquad v=\sum_{l=1}^{M}\sum_{k=1}^{M}(m'')_l\otimes(m)_{k,l}\otimes(m')_{k,l}\in V_\nu \quad \text{component of } V_{m''}\otimes V_m\otimes V_{m'}.$$

It follows that $T_\nu(1\otimes T_\mu)(v)\neq 0$ and $T(v)=0$ for all $T\in I$ such that $T\neq T_\nu(1\otimes T_\mu)$. This implies that the elements of I are linearly independent over $\mathbb{C}$.

Now with m, m', m'' as in (1.15a and b) and μ' and ν are $U(n)$ highest weights such that V_μ occurs in $V_{m''}\otimes V_m$ and V_ν occurs in $V_{\mu'}\otimes V_{m'}$, then we construct the $U(n)$ intertwining maps

$$(1.19a)\qquad T_{\mu'}\otimes 1:(V_{m''}\otimes V_m)\otimes V_{m'}\to V_{\mu'}\otimes V_{m'}$$

and

$$(1.19b)\qquad T_\nu: V_{\mu'}\otimes V_{m'}\to V_\nu.$$

As above, a basis over $\mathbb{C}$ for Hom $(V_{m''}\otimes V_m\otimes V_{m'}, V_\nu)$, if nonzero, is given by the set of all $U(n)$ intertwining maps of the form:

$$(1.20)\qquad T_\nu(T_{\mu'}\otimes 1): V_{m''}\otimes V_m\otimes V_{m'}\to V_\nu$$

for all highest weights μ' such that $V_{\mu'}$ occurs in $V_{m''}\otimes V_m$ and V_ν occurs in $V_{\mu'}\otimes V_{m'}$.

We thus obtain two different bases over $\mathbb{C}$ for Hom $(V_{m''}\otimes V_m\otimes V_{m'}, V_\nu)$: One basis consisting of maps of the form (1.17) and another basis consisting of maps of the form (1.20). The entries of the change of basis matrices between these two are called *recoupling* or (multiplicity-free) *Racah coefficients.*

DEFINITION 1.21. Let assumptions be as in (1.15a and b). Define the (multiplicity-free) Racah coefficients

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix}\in\mathbb{R}$$

by the following identity:

$$(1.22)\qquad T_\nu(1\otimes T_\mu)=\sum_{\mu'}\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix}T_\nu(T_{\mu'}\otimes 1)$$

where the sum is over all highest weights μ' such that $V_{\mu'}$ occurs in $V_{m''}\otimes V_m$ and V_ν occurs in $V_{\mu'}\otimes V_{m'}$.

Remark 1.23. In Definition 1.21 we can vary the assumptions on the highest weights m, m' and m''. Let m'' be an arbitrary $U(n)$ highest weight and m, m' be of the form $[p,0,\cdots,0]$, $[p',0,\cdots,0]$ respectively, where p, p' are nonnegative integers. If $\mu=m+m'=[p+p',0,\cdots,0]$ then the intertwining map $T_\nu(1\otimes T_\mu)$ is defined similarly to (1.15a and b). For all highest weights μ' such that $V_{\mu'}$ occurs in $V_{m''}\otimes V_m$, the maps $T_\nu(T_{\mu'}\otimes 1)$ are defined similarly to (1.19a and b). They still form a basis of Hom $(V_{m''}\otimes V_m\otimes V_{m'}, V_\nu)$. We therefore define $\left[\begin{smallmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{smallmatrix}\right]$ by the same expression (1.22).

Remark 1.24. The notation in Definition 1.21 for the multiplicity-free Racah coefficients above differs from the standard notation of Biedenharn et al. [17]. The standard notation for a general Racah coefficient of $U(n)$ is much more complicated and, in the case above, would have a number of redundancies. The notation of Definition 1.21 is similar to that of the 6-j symbols for $SU(2)$ [7], [12] or more general groups [10], [29]. However, the Racah coefficients are not identical to the 6-j symbols but differ by a product of dimension factors and a sign factor [29]. The 6-j symbols themselves for $U(n)$, $n>2$, are not completely satisfactory as different authors choose different sign factors (see [28]). Since we will not need these complicated sign factors in this paper, we avoid them.

We now state the orthogonality relations satisfied by the Racah coefficients above.

PROPOSITION 1.25. *With notation as in Definition* 1.21 *let the highest weights μ, ν, m, m', m'' satisfy the assumptions of* (1.15a *and* b) *and also with μ replaced by the highest weight $\bar{\mu}$. Similarly let μ', ν, m, m', m'' satisfy the corresponding assumptions for* (1.19a *and* b) *and also with $\bar{\mu}'$ in place of μ'. Then we have*

$$\sum_{\mu'} \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \bar{\mu} \end{bmatrix} = \begin{cases} 1 & \textit{if } \mu=\bar{\mu}, \\ 0 & \textit{otherwise}, \end{cases} \tag{1.26a}$$

where the sum is over all highest weights μ' such that $V_{\mu'}$ occurs in $V_{m''}\otimes V_m$ and V_ν occurs in $V_{\mu'}\otimes V_{m'}$. Also

$$\sum_{\mu} \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} \begin{bmatrix} m'' & m & \bar{\mu}' \\ m' & \nu & \mu \end{bmatrix} = \begin{cases} 1 & \textit{if } \mu'=\bar{\mu}', \\ 0 & \textit{otherwise}, \end{cases} \tag{1.26b}$$

where the sum is over all highest weights μ such that V_μ occurs in $V_m\otimes V_{m'}$ and V_ν occurs in $V_{m''}\otimes V_\mu$.

Proof. Assume that μ, $\bar{\mu}$ are highest weights of irreducible representations occurring in $V_m\otimes V_{m'}$ and also that ν and $\bar{\nu}$ are highest weights of irreducible representations occurring in $V_{m''}\otimes V_\mu$ and $V_{m''}\otimes V_{\bar{\mu}}$ respectively. Let (ν) be a fixed state of V_ν and $(\bar{\nu})$ of $V_{\bar{\nu}}$. We first prove the following identity:

$$\begin{aligned} &\sum_{(m'')\otimes(m)\otimes(m')} \langle(\nu), T_\nu(1\otimes T_\mu)((m'')\otimes(m)\otimes(m'))\rangle \\ &\qquad\cdot\langle(\bar{\nu}), T_{\bar{\nu}}(1\otimes T_{\bar{\mu}})((m'')\otimes(m)\otimes(m'))\rangle \\ &= \begin{cases} 1 & \text{if } \mu=\bar{\mu},\ \nu=\bar{\nu} \text{ and } (\nu)=(\bar{\nu}), \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{1.27}$$

where the sum is over all states $(m'')\otimes(m)\otimes(m')$ of $V_{m''}\otimes V_m\otimes V_{m'}$.

By the definition (1.7) of T_ν, $T_{\bar{\nu}}$, T_μ and $T_{\bar{\mu}}$ the left-hand side of (1.27) equals

$$\begin{aligned} &\sum_{(m'')\otimes(m)\otimes(m')} \Big\{\sum_{(\mu)} \langle(\nu), T_\nu((m'')\otimes(\mu))\rangle \\ &\qquad\cdot\langle(m'')\otimes(\mu), (1\otimes T_\mu)((m'')\otimes(m)\otimes(m'))\rangle\Big\} \\ &\quad\cdot\Big\{\sum_{(\bar{\mu})} \langle(\bar{\nu}), T_{\bar{\nu}}((m'')\otimes(\bar{\mu}))\rangle\langle(m'')\otimes(\bar{\mu}), (1\otimes T_{\bar{\mu}})((m'')\otimes(m)\times(m'))\rangle\Big\} \end{aligned} \tag{1.28a}$$

$$\begin{aligned} &= \sum_{(m'')\otimes(m)\otimes(m')} \Big\{\sum_{(\mu)} \langle(\nu), T_\nu((m'')\otimes(\mu))\rangle\langle(\mu), T_\mu((m)\otimes(m'))\rangle\Big\} \\ &\quad\cdot\Big\{\sum_{(\bar{\mu})} \langle(\bar{\nu}), T_{\bar{\nu}}((m'')\otimes(\bar{\mu}))\rangle\langle(\bar{\mu}), T_{\bar{\mu}}((m)\otimes(m'))\rangle\Big\} \end{aligned} \tag{1.28b}$$

where the outer sum in (1.28a and b) is the same as in (1.27) and the inner sums are over all states (μ), $(\bar{\mu})$ of V_μ and $V_{\bar{\mu}}$ respectively. By interchanging the order of summation in (1.28b) we find that (1.28b) and the left-hand side of (1.27) equal

$$\sum_{(m''),(\mu),(\bar{\mu})} \langle(\nu), T_\nu((m'')\otimes(\mu))\rangle\langle(\bar{\nu}), T_{\bar{\nu}}((m'')\otimes(\bar{\mu}))\rangle \tag{1.29}$$
$$\cdot\left\{\sum_{(m)\otimes(m')} \langle(\mu), T_\mu((m)\otimes(m'))\rangle\langle(\bar{\mu}), T_{\bar{\mu}}((m)\otimes(m'))\rangle\right\}$$

where the outer sum is over all states (m''), (μ), $(\bar{\mu})$ of $V_{m''}$, V_μ and $V_{\bar{\mu}}$ respectively and the inner sum is over all states $(m)\otimes(m')$ of $V_m\otimes V_{m'}$. By identity (1.14b) the inner sum in (1.29) equals 1 if $\mu=\bar{\mu}$ and $(\mu)=(\bar{\mu})$ and equals 0 otherwise. Hence if $\mu\neq\bar{\mu}$, then (1.29) and the left-hand side of (1.27) equals 0. If $\mu=\bar{\mu}$, then (1.29) equals

$$\sum_{(m'')\otimes(\mu)} \langle(\nu), T_\nu((m'')\otimes(\mu))\rangle\langle(\bar{\nu}), T_{\bar{\nu}}((m'')\otimes(\mu))\rangle \tag{1.30}$$

where the sum is over all states of $V_{m''}\otimes V_\mu$. By equation (1.14b) the expression (1.30) and hence the left-hand side of (1.27) equals 1 if $\nu=\bar{\nu}$ and $(\nu)=(\bar{\nu})$, otherwise it equals 0. This completes the proof of (1.27).

With assumptions similar to that in (1.27) and with a similar proof we have

$$\sum_{(m'')\otimes(m)\otimes(m')} \langle(\nu), T_\nu\otimes(T_{\mu'}\otimes 1)((m'')\otimes(m)\otimes(m'))\rangle \tag{1.31}$$
$$\cdot\langle(\nu), T_{\bar{\nu}}\otimes(T_{\bar{\mu}'}\otimes 1)((m'')\otimes(m)\otimes(m'))\rangle$$
$$=\begin{cases}1 & \text{if } \mu'=\bar{\mu}', \quad \nu=\bar{\nu} \quad \text{and } (\nu)=(\bar{\nu}),\\ 0 & \text{otherwise,}\end{cases}$$

where the sum is over all states in $V_{m''}\otimes V_m\otimes V_{m'}$.

Applying (1.22), we obtain the following:

$$\sum_{(m'')\otimes(m)\otimes(m')} \langle(\nu), T_\nu(1\otimes T_\mu)((m'')\otimes(m)\otimes(m'))\rangle \tag{1.32}$$
$$\cdot\langle(\nu), T_\nu(1\otimes T_{\bar{\mu}})((m'')\otimes(m)\otimes(m'))\rangle$$
$$=\sum_{\mu',\bar{\mu}'}\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu\end{bmatrix}\begin{bmatrix} m'' & m & \bar{\mu}' \\ m' & \nu & \bar{\mu}\end{bmatrix}$$
$$\cdot\left\{\sum_{(m'')\otimes(m)\otimes(m')} \langle(\nu), T_\nu\otimes(T_{\mu'}\otimes 1)((m'')\otimes(m)\otimes(m'))\rangle\right.$$
$$\left.\cdot\langle(\nu), T_\nu\otimes(T_{\bar{\mu}'}\otimes 1)((m'')\otimes(m)\otimes(m'))\rangle\right\}$$

where the sum on the left-hand side and the corresponding sum on the right-hand side is over all states of $V_{m''}\otimes V_m\otimes V_{m'}$ and the outer sum on the right-hand side is over all highest weights μ', $\bar{\mu}'$ of irreducible representations occurring in $V_{m''}\otimes V_m$ such that V_ν occurs in $V_{\mu'}\otimes V_{m'}$ and $V_{\bar{\mu}'}\otimes V_{m'}$. From (1.31) the inner sum on the right-hand side of (1.32) equals 1 if $\mu'=\bar{\mu}'$ and 0 otherwise. Thus the right-hand side of (1.32) reduces to the following expression:

$$\sum_{\mu'}\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu\end{bmatrix}\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \bar{\mu}\end{bmatrix}. \tag{1.33}$$

By identity (1.27) the left-hand side of (1.32) equals 1 if $\mu=\bar{\mu}$ and 0 otherwise. This completes the proof of equation (1.26a).

Fix highest weights m, m' and m'' satisfying the assumptions of (1.15a and b) and ν such that V_ν occurs in $V_{m''}\otimes V_m \otimes V_{m'}$. Observe that the number of highest weights μ and the number of μ' for which $\left[\begin{smallmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{smallmatrix}\right]$ is defined are both equal to $\dim_{\mathbb{C}} (\operatorname{Hom}(V_{m''}\otimes V_m \otimes V_{m'}, V_\nu))$. Thus equation (1.26a) implies that the Racah coefficients $\left[\begin{smallmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{smallmatrix}\right]$ are the entries of a real orthogonal matrix with rows indexed by the highest weights μ' and columns indexed by the μ. Equation (1.26b) is a consequence of the orthogonality of this matrix. This completes the proof of Proposition 1.25.

COROLLARY 1.34. *With assumptions as in Proposition* 1.25 *we have*

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} = \begin{bmatrix} m' & m & \mu \\ m'' & \nu & \mu' \end{bmatrix}. \tag{1.35}$$

Proof. Let $i: V_{m''}\otimes V_m \otimes V_{m'} \to V_{m'}\otimes V_m \otimes V_{m''}$ be the isomorphism mapping the state $(m'')\otimes(m)\otimes(m')$ of $V_{m''}\otimes V_m \otimes V_{m'}$ to the state $(m')\otimes(m)\otimes(m'')$ of $V_{m'}\otimes V_m \otimes V_{m''}$. By Remark 1.9 we see that $T_\nu(1\otimes T_\mu) = T_\nu(T_\mu \otimes 1)i$ and $T_\nu(T_{\mu'}\otimes 1) = T_\nu(1\otimes T_{\mu'})i$. It follows that

$$T_\nu(T_\mu \otimes 1) = \sum_{\mu'} \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} T_\nu(1\otimes T_{\mu'}), \tag{1.36}$$

where the sum is over all highest weights μ' such that $V_{\mu'}$ occurs in $V_m \otimes V_{m''}$ and V_ν occurs in $V_{m'}\otimes V_{\mu'}$. If we now interchange the labels m'' and m' and interchange μ and μ', we obtain

$$T_\nu(T_{\mu'} \otimes 1) = \sum_{\mu} \begin{bmatrix} m' & m & \mu \\ m'' & \nu & \mu' \end{bmatrix} T_\nu(1\otimes T_{\mu}) \tag{1.37}$$

where the sum is over all highest weights μ such that V_μ occurs in $V_m\otimes V_{m'}$ and V_ν occurs in $V_{m''}\otimes V_\mu$.

Multiplying both sides of (1.22) by $\left[\begin{smallmatrix} m'' & m & \bar{\mu}' \\ m' & \nu & \mu \end{smallmatrix}\right]$ and summing over μ, we obtain

$$\begin{aligned} \sum_\mu \begin{bmatrix} m'' & m & \bar{\mu}' \\ m' & \nu & \mu \end{bmatrix} T_\nu(1\otimes T_\mu) &= \sum_{\mu,\mu'} \begin{bmatrix} m'' & m & \bar{\mu}' \\ m' & \nu & \mu \end{bmatrix}\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} T_\nu(T_{\mu'}\otimes 1) \\ &= T_\nu(T_{\bar{\mu}'}\otimes 1) \end{aligned} \tag{1.38}$$

by (1.26a) with summations over μ and μ' as in (1.26a and b). Setting $\bar{\mu}' = \mu'$ and equating the coefficients of $T_\nu(1\otimes T_\mu)$ in identities (1.37) and (1.38), we obtain (1.35).

We now prove a generalization of the Biedenharn-Elliott identity for the multiplicity free Racah coefficients above. For $SU(2)$, the original statement and proofs are due to Biedenharn [5] and Elliott [13] independently. For a more complete exposition of the $SU(2)$ Biedenharn-Elliott identity see [12] and [7]. A generalization of this identity to arbitrary groups for general Racah coefficients was given by Derome and Sharp [10]. Louck and Biedenharn [19] give an alternative proof for the $U(n)$ generalization of this identity. All these proofs consider general (nonmultiplicity-free) Racah coefficients. The proof given below for multiplicity-free Racah coefficients is a simple generalization of the $SU(2)$ proof involving recoupling coefficients for four-fold tensor products (see [12]).

Let m_1 be the highest weight for an arbitrary irreducible representation of $U(n)$. Let m_2, m_3 and m_4 be $U(n)$ highest weights of the form $m_j = [p_j, 0, \cdots, 0]$ where p_j is a nonnegative integer for $j = 2, 3, 4$. Let V_m be an irreducible representation space occurring in $\bigotimes_{i=1}^4 V_{m_i}$. A basis for $\operatorname{Hom}_{U(n)}(\bigotimes_{i=1}^4 V_{m_i}, V_m)$ is given by the following intertwining maps:

$$T_m(T_{m_{123}}\otimes 1)(\cdot T_{m_{12}}\otimes 1\otimes 1):\bigotimes_{i=1}^{4} V_{m_i}\to V_m \tag{1.39}$$

where

$$T_{m_{12}}: V_{m_1}\otimes V_{m_2}\to V_{m_{12}},\quad T_{m_{123}}: V_{m_{12}}\otimes V_{m_3}\to V_{m_{123}},\quad T_m: V_{m_{123}}\otimes V_{m_4}\to V_m, \tag{1.40}$$

as m_{12}, m_{123} vary over all highest weights of irreducible representations occurring in $V_{m_1}\otimes V_{m_2}$, $V_{m_{12}}\otimes V_{m_3}$ respectively and such that V_m occurs in $V_{m_{123}}\otimes V_{m_4}$.

Let m_{23} be the highest weight $[p_2+p_3, 0, \cdots, 0]=m_2+m_3$ and m_{14} be the highest weight of some irreducible representation occurring in $V_{m_1}\otimes V_{m_4}$. If V_m occurs in $V_{m_{23}}\otimes V_{m_{14}}$, then consider the intertwining map:

$$T_m(T_{m_{23}}\otimes T_{m_{14}}):\bigotimes_{i=1}^{4} V_{m_i}\to V_m \tag{1.41}$$

where

$$T_{m_{23}}: V_{m_2}\otimes V_{m_3}\to V_{m_{23}},\quad T_{m_{14}}: V_{m_1}\otimes V_{m_4}\to V_{m_{14}},\quad T_m: V_{m_{23}}\otimes V_{m_{14}}\to V_m. \tag{1.42}$$

We now express the intertwining map (1.41) in terms of the basis (1.39):

$$T_m(T_{m_{23}}\otimes T_{m_{14}})=\sum_{m_{12},m_{123}} C_{m_{12},m_{123}}T_m(T_{m_{123}}\otimes 1)(T_{m_{12}}\otimes 1\otimes 1) \tag{1.43}$$

where $C_{m_{12},m_{123}}\in\mathbb{C}$ are constants. We will compute $C_{m_{12},m_{123}}$ in two different ways.

We have

$$\begin{aligned} T_m(T_{m_{23}}\otimes T_{m_{14}})&=T_m(1\otimes T_{m_{14}})(T_{n_{23}}\otimes 1\otimes 1)\\ &=\sum_{\bar m_{123}}\begin{bmatrix} m_{23} & m_1 & \bar m_{123}\\ m_4 & m & m_{14}\end{bmatrix}T_m(T_{\bar m_{123}}\otimes 1)(T_{m_{23}}\otimes 1\otimes 1),\end{aligned} \tag{1.44}$$

where the sum is over all highest weights $\bar m_{123}$ such that $V_{\bar m_{123}}$ occurs in $V_{m_1}\otimes V_{m_{23}}$ and V_m occurs in $V_{\bar m_{123}}\otimes V_{m_4}$. Also we have

$$T_m(T_{\bar m_{123}}\otimes 1)(T_{m_{23}}\otimes 1\otimes 1)=\sum_{\bar m_{12}}\begin{bmatrix} m_1 & m_2 & \bar m_{12}\\ m_2 & \bar m_{123} & m_{23}\end{bmatrix}T_m(T_{\bar m_{123}}\otimes 1)(T_{\bar m_{12}}\otimes 1\otimes 1), \tag{1.45}$$

where the sum is over all highest weights $\bar m_{12}$ such that $V_{\bar m_{12}}$ occurs in $V_{m_1}\otimes V_{m_2}$ and $V_{\bar m_{123}}$ occurs in $V_{\bar m_{12}}\otimes V_{m_3}$.

From (1.44) and (1.45) it follows that

$$C_{m_{12},m_{123}}=\begin{bmatrix} m_1 & m_2 & m_{12}\\ m_3 & m_{123} & m_{23}\end{bmatrix}\begin{bmatrix} m_{23} & m_1 & m_{123}\\ m_4 & m & m_{14}\end{bmatrix}. \tag{1.46}$$

On the other hand, we have

$$\begin{aligned} T_m(T_{m_{23}}\otimes T_{m_{14}})&=T_m(T_{m_{23}}\otimes 1)(1\otimes 1\otimes Tm_{14})\\ &=\sum_{m_{124}}\begin{bmatrix} m_{14} & m_2 & m_{124}\\ m_3 & m & m_{23}\end{bmatrix}T_m(1\otimes T_{m_{124}})(1\otimes 1\otimes T_{m_{14}})\end{aligned} \tag{1.47}$$

where the sum is over all highest weights m_{124} such that $V_{m_{124}}$ occurs in $V_{m_{14}}\otimes V_{m_2}$ and V_m occurs in $V_{m_{124}}\otimes V_{m_3}$. Similarly,

$$T_m(1\otimes T_{m_{124}})(1\otimes 1\otimes T_{m_{14}})=\sum_{\bar m_{12}}\begin{bmatrix} m_2 & m_1 & \bar m_{12}\\ m_4 & m_{124} & m_{14}\end{bmatrix}T_m(1\otimes T_{m_{124}})(1\otimes 1\otimes T_{\bar m_{12}}) \tag{1.48}$$

and

$$(1.49)\quad T_m(1\otimes T_{m_{124}})(1\otimes 1\otimes T_{\bar{m}_{12}}) = \sum_{\bar{m}_{123}} \begin{bmatrix} m_3 & \bar{m}_{12} & \bar{m}_{123} \\ m_4 & m & m_{124} \end{bmatrix} T_m(T_{\bar{m}_{123}}\otimes 1)(T_{\bar{m}_{12}}\otimes 1\otimes 1),$$

with the sums in (1.48) and (1.49) satisfying conditions similar to that in (1.45).

Equations (1.43) and (1.46)-(1.49) imply the following:

Proposition 1.50 (Generalization of Biedenharn-Elliott Identity). *Let m_1 be an arbitrary $U(n)$ highest weight and m_2, m_3, m_4 be of the form $[p_j, 0, \cdots, 0]$ with p_j a nonnegative integer for $j=2$, 3 4. Let the conditions in* (1.40)-(1.42) *be satisfied, then*

$$(1.51)\quad \begin{aligned} &\begin{bmatrix} m_1 & m_2 & m_{12} \\ m_3 & m_{123} & m_{23} \end{bmatrix}\begin{bmatrix} m_{23} & m_1 & m_{123} \\ m_4 & m & m_{14} \end{bmatrix} \\ &= \sum_{m_{124}} \begin{bmatrix} m_3 & m_{12} & m_{123} \\ m_4 & m & m_{124} \end{bmatrix}\begin{bmatrix} m_2 & m_1 & m_{12} \\ m_4 & m_{124} & m_{14} \end{bmatrix}\begin{bmatrix} m_{14} & m_2 & m_{124} \\ m_3 & m & m_{23} \end{bmatrix}, \end{aligned}$$

where the sum is over all highest weights m_{124} such that $V_{m_{124}}$ occurs in $V_{m_{14}}\otimes V_{m_2}$ and $V_{m_{12}}\otimes V_{m_4}$ and V_m occurs in $V_{m_{124}}\otimes V_{m_3}$.

2. A generalization of Whipple's transformation. In this section we will explicitly compute a special case of the generalized Biedenharn-Elliott identity (1.51). We will then obtain a transformation relating a homogeneous Holman hypergeometric series $W^{(n)}$, "well-poised in $SU(n)$," to a product of gamma functions times a nonhomogeneous Holman hypergeometric series $F^{(n-1)}$ in $U(n-1)$. For $n=2$ this result reduces to a terminating form of the classical Whipple's transformation between a balanced ${}_4F_3$ and a well-poised ${}_7F_6$ hypergeometric series [25, eq. (7.7)].

We begin by writing explicit expressions for the Racah coefficients in identity (1.51). These multiplicity-free Racah coefficients have been (essentially) computed by Ališaukas, Jucys and Jucys [1] (see also [2], [14]) and also by Wong [28], [29].

We start with the following definition.

Definition 2.1. If $\lambda=[\lambda_1, \cdots, \lambda_n]$ is a $U(n)$ highest weight with $\lambda_i \geqq 0$ for $i=1, \cdots, n$ then define

$$(2.2)\qquad \mathcal{M}(\lambda)=\frac{(\lambda_1+n-1)!(\lambda_2+n-2)!\cdots(\lambda_n)!}{\prod_{1\leqq i<j\leqq n}(\lambda_i-\lambda_j+j-i)}.$$

We now express the multiplicity-free Racah coefficient in Definition 1.21 as a product of $\mathcal{M}(\lambda)$ factors times a $U(n+1):U(n)$ reduced Wigner coefficient (see [6] and [9]). This result is given in Wong [29] and Holman [14] except for simple factors due to slightly different definitions. We note that there is a simple typographical error in formula (1.9) of Wong [28] which is corrected in formula (1.13) of [28] and also in formula (3.4) of [29] which is corrected in Holman [14]. In proving the (corrected) formula (1.9) of [28] one uses identity (2.25b) of [18].

Proposition 2.3. *Let $n \geqq 2$. Suppose that the $U(n)$ highest weights m, m', m'', ν, μ, μ' are partitions* (*i.e. all components are nonnegative integers*) *and satisfy the assumptions of Proposition* 1.25, *then*

$$(2.4)\quad \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} = \left[\frac{\mathcal{M}(\nu)\mathcal{M}(m)}{\mathcal{M}(\mu')\mathcal{M}(\mu)}\right]^{1/2}\left\langle\begin{pmatrix} [\nu,0]_{n+1} \\ \mu \end{pmatrix}\middle|\begin{bmatrix} [m',0]_{n+1} \\ m' \end{bmatrix}\middle|\begin{pmatrix} [\mu',0]_{n+1} \\ m \end{pmatrix}\right\rangle$$

where

$$\left\langle\begin{pmatrix} [\nu,0]_{n+1} \\ \mu \end{pmatrix}\middle|\begin{bmatrix} [m',0]_{n+1} \\ m' \end{bmatrix}\middle|\begin{pmatrix} [\mu',0]_{n+1} \\ m \end{pmatrix}\right\rangle$$

is a $U(n+1):U(n)$ *reduced Wigner coefficient* [6], [9] *and if* $\lambda=[\lambda_1,\cdots,\lambda_n]$ *is a partition, then* $[\lambda,0]_{n+1}$ *is the partition* $[\lambda_1,\cdots,\lambda_n,0]$.

Similarly we obtain from [29, formula (3.3)] the following:

PROPOSITION 2.5. *Let* $n\geqq 2$. *Suppose that the* $U(n)$ *highest weights* m, m', m'', ν, μ, μ' *are partitions and satisfy the assumptions of Remark* 1.23; *then we have*

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix}=\left[\frac{\mathcal{M}(\nu)\mathcal{M}(m)}{\mathcal{M}(\mu')\mathcal{M}(\mu)}\right]^{1/2}\left\langle\begin{pmatrix}[\nu,0]_{n+1}\\ m''\end{pmatrix}\middle|\begin{bmatrix}[m',0]_{n+1}\\ [0,\cdots,0]_n\end{bmatrix}\middle|\begin{pmatrix}[\mu,0]_{n+1}\\ m''\end{pmatrix}\right\rangle \tag{2.6}$$

with notation as in Proposition 2.3.

We now describe the formula for the $U(n+1):U(n)$ reduced Wigner coefficients appearing in (2.4) and (2.6). We first establish our notation.

DEFINITION 2.7. If $h=[h_1,\cdots,h_l]$ is a $U(l)$ highest weight and $q=[q_1,\cdots,q_m]$ is a $U(m)$ highest weight where l and m are positive integers, $l\geqq m$ and

$$h_1\geqq q_1\geqq\cdots\geqq h_m\geqq q_m\geqq h_{m+1}\geqq\cdots\geqq h_l, \tag{2.8}$$

then

$$S_{lm}(h;q)=\left[\frac{\prod_{k=1}^{m}\prod_{s=1}^{k}\Gamma(h_s-q_k+k-s+1)}{\prod_{k=1}^{l-1}\prod_{s=k+1}^{l}\Gamma(q_k-h_s+s-k)}\right]^{1/2}. \tag{2.9a}$$

We also denote the omission of all factors containing q_i for some i, $1\leqq i\leqq m$, in the following manner:

$$S_{l\underset{i}{m}}(h;q)=\left[\frac{\prod_{s=i+1}^{l}\Gamma(q_i-h_s+s-i)}{\prod_{s=1}^{i}\Gamma(h_s-q_i+i-s+1)}\right]^{1/2}S_{lm}(h;q); \tag{2.9b}$$

and similarly for some i, $1\leqq i\leqq l$,

$$S_{\underset{i}{l}m}(h;q)=\left[\frac{\prod_{k=1}^{i-1}\Gamma(q_k-h_i+i-k)}{\prod_{k=1}^{m}\Gamma(h_i-q_k+k-i+1)}\right]^{1/2}S_{lm}(h;q), \tag{2.9c}$$

$$S_{\underset{i}{l}\underset{i}{m}}(h;q)=\left[\frac{\prod_{k=1}^{i-1}\Gamma(q_k-h_i+i-k)}{\prod_{k=i+1}^{m}\Gamma(h_i-q_k+k-i+1)}\right]^{1/2}S_{l\underset{i}{m}}(h;q). \tag{2.9d}$$

In terms of these quantities we can express the reduced $U(n+1):U(n)$ Wigner coefficients in two different ways. Let h, h' be $U(n+1)$ highest weights such that

$$h_1\geqq h_1'\geqq h_2\geqq h_2'\geqq\cdots\geqq h_{n+1}\geqq h_{n+1}' \tag{2.10a}$$

and

$$\sum_{i=1}^{n+1}(h_i-h_i')=p \tag{2.10b}$$

for some nonnegative integer p. Similarly let q, q' be $U(n)$ highest weights such that

$$q_1\geqq q_1'\geqq\cdots\geqq q_n\geqq q_n' \tag{2.11a}$$

and

$$\sum_{i=1}^{n}(q_i-q_i')=p' \tag{2.11b}$$

for some nonnegative integer p', $p'\leqq p$. We also assume h and q satisfy the betweenness condition:

$$h_1\geqq q_1\geqq h_2\geqq q_2\geqq\cdots\geqq q_n\geqq h_{n+1} \tag{2.12}$$

and similarly for h' and q'. From the calculation of Chacón, Ciftan and Biedenharn [9] (see also [14]) we have

$$\tag{2.13}\begin{aligned}\left\langle\begin{pmatrix}h\\q\end{pmatrix}\middle|\begin{bmatrix}[p,0,\cdots,0]_{n+1}\\ [p',0,\cdots,0]_n\end{bmatrix}\middle|\begin{pmatrix}h'\\q'\end{pmatrix}\right\rangle &= \sqrt{(p-p')!}\,\frac{S_{n+1\,n+1}(h;h)S_{n+1\,n}(h';q')S_{nn}(q;q')}{S_{n+1\,n+1}(h;h')S_{n+1\,n}(h;q)}\\ &\cdot S_{nn}(q';q')\sum_{\rho_1,\cdots,\rho_n}(-1)^{\rho_1+\cdots+\rho_n}\\ &\cdot\left[\frac{S_{nn}(\bar q;\bar q)}{S_{n+1\,n}(h';\bar q)S_{nn}(q;\bar q)}\right]^2\cdot\left[\frac{S_{n+1\,n}(h;\bar q)}{S_{nn}(\bar q;q')}\right]^2\end{aligned}$$

where $\bar q_i = q'_i+\rho_i$, ρ_i is a nonnegative integer for $i=1,\cdots,n$ and $\bar q$ is restricted so that all factors on the right-hand side of (2.13) are defined and finite.

From the calculation of Ališaukas, Jucys and Jucys [1] we find

$$\tag{2.14}\begin{aligned}&\left\langle\begin{pmatrix}h\\q\end{pmatrix}\middle|\begin{matrix}[p,0,\cdots,0]_{n+1}\\ [p',0,\cdots,0]_n\end{matrix}\middle|\begin{pmatrix}h'\\q'\end{pmatrix}\right\rangle\\ &\quad=\frac{1}{\sqrt{(p-p')!}}\,\frac{S_{n+1\,n+1}(h;h)S_{nn}(q';q')S_{n+1\,n+1}(h;h')}{S_{nn}(q;q')S_{n+1n}(h';q')}\\ &\quad\cdot S_{n+1\,n}(h;q)\sum_{[r_2,\cdots,r_{n+1}]}(-1)^{\varphi}\\ &\quad\cdot[S_{\substack{n+1\\1}\,\substack{n+1\\1}}(r;r)]^2\left[\frac{S_{\substack{n+1\\1}\,n}(r;q')}{S_{n+1\,\substack{n+1\\1}}(h;r)S_{\substack{n+1\\1}\,n+1}(r;h')S_{\substack{n+1\\1}n}(r;q)}\right]^2\end{aligned}$$

where $\varphi=\sum_{i=2}^{n}(h_i-r_i)$ and the sum is over all $U(n)$ highest weights $r=[r_2,\cdots,r_{n+1}]$ such that right-hand side of (2.14) is defined and finite. Note there is an "understood" r_1 component of r which does not appear in the expression (2.14).

Recall the generalized Biedenharn-Elliott identity (1.51). With notation as in (1.51) and $n\geqq 2$ we now compute a special case of identity (1.51) by setting

$$\tag{2.15} m_{23}=m_2+m_3,\quad m_{123}=m_{12}+m_3,\quad m_{12}=m_1-m_2^*,$$

where if $\lambda=[\lambda_1,\cdots,\lambda_n]$ is a highest weight of a $U(n)$ representation, then

$$\tag{2.16}\lambda^*=\bar\lambda=[-\lambda_n,\cdots,-\lambda_1]$$

is the highest weight of the contragredient $U(n)$ representation. For simplicity we shall further assume that m_1 is a partition, i.e. all the components are nonnegative integers.

Applying Propositions 2.3, 2.5 and Corollary 1.34, we have

$$\tag{2.17a}\begin{aligned}\begin{bmatrix}m_1 & m_2 & m_{12}\\ m_3 & m_{123} & m_{23}\end{bmatrix}&=\left[\frac{\mathcal{M}(m_{123})\mathcal{M}(m_2)}{\mathcal{M}(m_{23})\mathcal{M}(m_{12})}\right]^{1/2}\\ &\cdot\left\langle\begin{pmatrix}[m_{123},0]_{n+1}\\ m_1\end{pmatrix}\middle|\begin{bmatrix}[m_3,0]_{n+1}\\ [0,\cdots,0]_n\end{bmatrix}\middle|\begin{pmatrix}[m_{12},0]_{n+1}\\ m_1\end{pmatrix}\right\rangle,\end{aligned}$$

$$\tag{2.17b}\begin{aligned}\begin{bmatrix}m_{23} & m_1 & m_{123}\\ m_4 & m & m_{14}\end{bmatrix}&=\left[\frac{\mathcal{M}(m)\mathcal{M}(m_1)}{\mathcal{M}(m_{123})\mathcal{M}(m_{14})}\right]^{1/2}\\ &\cdot\left\langle\begin{pmatrix}[m,0]\\ m_{14}\end{pmatrix}\middle|\begin{bmatrix}[m_4,0]\\ m_4\end{bmatrix}\middle|\begin{pmatrix}[m_{123},0]\\ m_1\end{pmatrix}\right\rangle,\end{aligned}$$

$$\begin{bmatrix} m_3 & m_{12} & m_{123} \\ m_4 & m & m_{124} \end{bmatrix} = \left[\frac{\mathcal{M}(m_{12})\mathcal{M}(m)}{\mathcal{M}(m_{123})\mathcal{M}(m_{124})}\right]^{1/2} \cdot \left\langle \begin{pmatrix} [m,0] \\ m_4 \end{pmatrix} \middle| \begin{bmatrix} [m_3,0] \\ [0,\cdots,0]_n \end{bmatrix} \middle| \begin{pmatrix} [m_{124},0] \\ m_4 \end{pmatrix} \right\rangle, \tag{2.17c}$$

$$\begin{bmatrix} m_2 & m_1 & m_{12} \\ m_4 & m_{124} & m_{14} \end{bmatrix} = \left[\frac{\mathcal{M}(m_{124})\mathcal{M}(m_1)}{\mathcal{M}(m_{12})\mathcal{M}(m_{14})}\right]^{1/2} \cdot \left\langle \begin{pmatrix} [m_{124},0] \\ m_{12} \end{pmatrix} \middle| \begin{bmatrix} [m_2,0] \\ m_2 \end{bmatrix} \middle| \begin{pmatrix} [m_{14},0] \\ m_1 \end{pmatrix} \right\rangle, \tag{2.17d}$$

and

$$\begin{bmatrix} m_{14} & m_2 & m_{124} \\ m_3 & m & m_{23} \end{bmatrix} = \left[\frac{\mathcal{M}(m)\mathcal{M}(m_2)}{\mathcal{M}(m_{124})\mathcal{M}(m_{23})}\right]^{1/2} \cdot \left\langle \begin{pmatrix} [m,0] \\ m_{14} \end{pmatrix} \middle| \begin{bmatrix} [m_3,0] \\ [0,\cdots,0]_n \end{bmatrix} \middle| \begin{pmatrix} [m_{124},0] \\ m_{14} \end{pmatrix} \right\rangle, \tag{2.17e}$$

where $[m,0]=[m,0]_{n+1}$, etc. in the notation of Proposition 2.3.

For a $U(n)$ highest weight $\lambda=[\lambda_1,\cdots,\lambda_n]$ (similarly for a $U(n+1)$ highest weight) define

$$d(\lambda) = \text{dimension of } V_\lambda = \frac{\prod_{1\leqq i<j\leqq n}(p_{in}-p_{jn})}{\prod_{i=1}^{n-1} i!} \tag{2.18}$$

where $p_{in}=\lambda_i+n-i$ for $1\leqq i\leqq n$. Now recalling the definition (2.16) of the contragredient highest weight $\bar{\lambda}=\lambda^*$ and applying identity (A.17) from Appendix A, we have in place of (2.17d)

$$\begin{bmatrix} m_2 & m_1 & m_{12} \\ m_4 & m_{124} & m_{14} \end{bmatrix} = (-1)^{p_2}\left[\frac{\mathcal{M}(m_{124})\mathcal{M}(m_1)}{\mathcal{M}(m_{12})\mathcal{M}(m_{14})}\right]^{1/2} \cdot \left(\frac{d([m_{124},0]_{n+1})d(m_1)}{d(m_{12})d([m_{14},0]_{n+1})}\right)^{1/2} \cdot \left\langle \begin{pmatrix} \overline{[m_{14},0]_{n+1}} \\ \bar{m}_1 \end{pmatrix} \middle| \begin{bmatrix} [m_2,0] \\ m_2 \end{bmatrix} \middle| \begin{pmatrix} \overline{[m_{124},0]_{n+1}} \\ \overline{m_{12}} \end{pmatrix} \right\rangle \tag{2.19}$$

where $m_2=[p_2,0,\cdots,0]_n$ as in Proposition 1.50.

To compute the reduced Wigner coefficients appearing in (2.17a), (2.17c) and (2.17e), we use the Chacón, Ciftan and Biedenharn computation (2.13) and to compute the reduced Wigner coefficient in (2.17b) we use the Ališaukas, Jucys and Jucys computation (2.14). Finally to compute the reduced Wigner coefficient in (2.19), we use (2.13) and a summation theorem of Ališaukas, Jucys and Jucys given in identity (14) of [1] and described in [14, eq. (30)].

To describe the final result of this explicit calculation of the generalized Biedenharn-Elliott identity, we need to introduce Holman's "hypergeometric series in $U(n)$" [14], [15].

Definition 2.20. Let $n\geqq 2$. We define

$$W_q^{(n)}\left(\begin{array}{cccc|ccc|ccc|c} A_{12} & & & & a_{11} & \ldots & a_{1k} & b_{11} & \ldots & b_{1j} & z_1 \\ A_{13} & A_{23} & & & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & \ddots & & & & & & & & \\ A_{1n} & A_{2n} & \cdots & A_{n-1,\,n} & a_{n1} & \cdots & a_{nk} & b_{n1} & \cdots & b_{nj} & z_n \end{array}\right)$$

$$\text{(2.21)}\quad \equiv q! \sum_{y_1+\cdots+y_n=q} \left(\prod_{i=1}^{n-1}\prod_{j=i+1}^{n} \frac{(A_{ij}+y_i-y_j)}{A_{ij}}\right)\left(\prod_{i=1}^{k}\prod_{l=1}^{n}(a_{li})_{y_l}\right)$$
$$\cdot\left(\prod_{i=1}^{j}\prod_{l=1}^{n}(b_{li})_{y_l}\right)^{-1}\left(\prod_{i=1}^{n} z_i^{y_i}\right),$$

where q is a nonnegative integer and the sum is over all n-tuples $(y_1,\cdots,y_n)$ of nonnegative integers such that $y_1+\cdots+y_n=q$. The symbol $(a)_m$ denotes the rising factorial $(a)(a+1)\cdots(a+m-1)$ for $m\geqq 1$ and $a_0=1$.

The series (2.21) is called *well-poised in* $SU(n)$, $n\geqq 2$, if $j\geqq n$,

$$\text{(2.22)}\quad \begin{array}{ll} A_{ir}-A_{is}=A_{sr}, & \text{for } s<r, \\ a_{ir}-a_{sr}=A_{is}, & \text{for } i<s, \\ b_{ir}-b_{sr}=A_{is}, & \text{for } i<s, \\ b_{ii}=1, & 1\leqq i\leqq n. \end{array}$$

If we denote the series (2.21), *whether or not well-poised*, by $W_q^{(n)}((A)|(a)|(b)|(z))$, then for $n\geqq 2$ define

$$\text{(2.23a)}\quad F^{(n)}((A)|(a)|(b)|(z))=\sum_{q=0}^{\infty}\frac{1}{q!}W_q^{(n)}((A)|(a)|(b)|(z)).$$

For $n=1$ define

$$\text{(2.23b)}\quad F^{(1)}((A)|(a)|(b)|(z))=\sum_{q=0}^{\infty}\frac{(a_{11})_q\cdots(a_{1k})_q}{(b_{11})_q\cdots(b_{1j})_q}z_1^q,$$

which is a classical ${}_kF_{j-1}$ hypergeometric series.

We now state a generalization of Whipple's transformation (see [25] and [4, eq. (4.3.4)]).

THEOREM 2.24. *Let q be a nonnegative integer and let $z_1,\cdots,z_n$, $w_1,\cdots,w_n$, a, $b\in\mathbb{C}$ such that w_i-z_i is a nonnegative integer for $i=1,2,\cdots,n-1$. We also set $s=\sum_{k=1}^{n}(w_k-z_k)$. For $n\geqq 2$ we have*

$$W_q^{(n)}\left(\begin{array}{llll} z_1-z_2+1 & & & \\ z_1-z_3+2 & z_2-z_3+1 & & \\ \vdots & \vdots & & \\ z_1-z_n+n-1 & z_2-z_n+n-2 & \cdots & z_{n-1}-z_n+1 \end{array}\right.$$

$$\left|\begin{array}{ccccc} z_1-w_1 & z_1-w_2+1 & \cdots & z_1-w_n+n-1 & z_1+s-q-a \\ z_2-w_1-1 & z_2-w_2 & \cdots & z_2-w_n+n-2 & z_2+s-q-a-1 \\ \vdots & \vdots & & \vdots & \vdots \\ z_n-w_1-n+1 & z_n-w_2-n+2 & \cdots & z_n-w_n & z_n+s-q-a-n+1 \end{array}\right.$$

$$\left|\begin{array}{ccccc} 1 & z_1-z_2+2 & \cdots & z_1-z_n+n & z_1-b \\ z_2-z_1 & 1 & \cdots & z_2-z_n+n-1 & z_2-b-1 \\ \vdots & \vdots & & \vdots & \vdots \\ z_n-z_1-n+2 & z_n-z_2-n+3 & \cdots & 1 & z_n-b-n+1 \end{array}\right|\left.\begin{array}{c} -1 \\ -1 \\ \vdots \\ -1 \end{array}\right)$$

$$\text{(2.25)}\quad =\frac{\Gamma(w_n-z_n+1)}{\Gamma(1+s-q)}\prod_{l=1}^{n-1}\left[\frac{\Gamma(w_l-z_n+n-l+1)\Gamma(a-z_l+l)}{\Gamma(z_l-z_n+n-l+1)\Gamma(a-w_l+l)}\right]$$

$$\cdot\frac{\Gamma(a-s+q-z_n+n)\Gamma(b-z_n+n-q)}{\Gamma(a-w_n+n)\Gamma(b-z_n+n)}$$

$$\cdot F^{(n-1)}\left(\begin{array}{llll} z_1-z_2+1 & & & \\ z_1-z_3+2 & z_2-z_3+1 & & \\ \vdots & \vdots & & \\ z_1-z_{n-1}+n-2 & z_2-z_{n-1}+n-3 & \cdots & z_{n-2}-z_{n-1}+1 \end{array}\right.$$

$$\left|\begin{array}{lll} z_1-w_1 & z_1-w_2+1 & \cdots \\ z_2-w_1-1 & z_2-w_2 & \cdots \\ \vdots & \vdots & \\ z_{n-1}-w_1-n+2 & z_{n-1}-w_2-n+3 & \cdots \end{array}\right.$$

$$\begin{array}{lll} z_1-w_n+n-1 & z_1+s-q-a & z_1+q-b \\ z_2-w_n+n-2 & z_2+s-q-a-1 & z_2+q-b-1 \\ \vdots & \vdots & \vdots \\ z_{n-1}-w_n+1 & z_{n-1}+s-q-a-n+2 & z_{n-1}+q-b-n+2 \end{array}$$

$$\left|\begin{array}{lll} 1 & z_1-z_2+2 & z_1-z_n+n \\ z_2-z_1 & 1 & z_2-z_n+n-1 \\ \vdots & \vdots & \vdots \\ z_{n-1}-z_1-n+3 & z_{n-1}-z_2-n+4 & z_{n-1}-z_n+2 \end{array}\right.$$

$$\left.\begin{array}{ll|l} z_1-a & z_1-b & 1 \\ z_2-a-1 & z_2-b-1 & 1 \\ \vdots & \vdots & \vdots \\ z_{n-1}-a-n+2 & z_{n-1}-b-n+2 & 1 \end{array}\right).$$

Proof. If λ is a $U(n)$ highest weight, let $(\lambda)_k$ denote the kth component of λ, i.e. $\lambda=[(\lambda)_1,\cdots,(\lambda)_k,\cdots,(\lambda)_n]$. Together with the notation and assumptions of Proposition 1.50 and equations (2.15), we will also assume that

$$\begin{aligned}(m_1)_1>(m_1)_2+(p_2+p_3+p_4)>(m_1)_3\\ +2(p_2+p_3+p_4)>\cdots>(m_1)_n+(n-1)(p_2+p_3+p_4).\end{aligned}\tag{2.26}$$

After a lengthy explicit calculation of identity (1.51) using formulas (2.17a–e) and (2.19) as described above, we find

$$\begin{aligned}\sum_{y_1+y_2+\cdots+y_n=p_3}\Bigg\{\prod_{1\leqq r<s\leqq n}\left(\frac{A_{rs}+y_r-y_s}{A_{rs}}\right)\prod_{l=1}^{n}\Bigg[\frac{(p_2+(m_1)_n-(m)_l+l-n)_{y_l}}{((m_1)_1-(m)_l+l-1)_{y_l}}\\ \cdot\prod_{k=1}^{n}\frac{((m_{14})_k-(m)_l+l-k)_{y_l}}{((m)_k-(m)_l+l-k+1)_{y_l}}\Bigg]\Bigg\}\\ =\frac{(-1)^{p_3}}{(p_3)!}\prod_{l=2}^{n}\left[\frac{\Gamma((m)_1-(m_{14})_l+l)}{\Gamma((m)_1-(m)_l+l)}\frac{\Gamma((m)_l-(m_1)_n+n-l+1)}{\Gamma((m_{14})_l-(m_1)_n+n-l+1)}\right]\\ \cdot\frac{\Gamma((m)_1-(m_{14})_1+1)}{(p_2)!}\frac{\Gamma((m)_1-(m_1)_n-p_2+n)}{\Gamma((m_{14})_1-(m_1)_n+n)}\frac{\Gamma((m)_1-(m_1)_1-p_3+1)}{\Gamma((m)_1-(m_1)_1+1)}\\ \cdot\sum_{\substack{y_l\geqq 0\\ 2\leqq l\leqq n}}\prod_{2\leqq r<s\leqq n}\left(\frac{A_{rs}+y_r-y_s}{A_{rs}}\right)\end{aligned}\tag{2.27}$$

$$\cdot \prod_{l=2}^{n} \left[\frac{(p_3+(m_1)_1-(m)_l+l-1)_{y_l}}{((m_1)_1-(m)_l+l-1)_{y_l}} \right.$$

$$\left. \cdot \frac{(p_2+(m_1)_n-(m)_l+l-n)_{y_l}}{((m_1)_n-(m)_l+l-n)_{y_l}} \prod_{k=1}^{n} \frac{((m_{14})_k-(m)_l+l-k)_{y_l}}{((m)_k-(m)_l+l+k+1)_{y_l}} \right]$$

where on both sides $A_{rs}=(m)_s-(m)_r+r-s$ and the y_i are nonnegative integers ($1 \leqq i \leqq n$ on the left-hand side, $2 \leqq i \leqq n$ on the right-hand side). On the left-hand side $y_i=(m)_i-(m_{124})_i$, $1 \leqq i \leqq n$, and on the right-hand side $y_i=(m)_i-r_i$, $2 \leqq i \leqq n$, where the r_i are indices of summation in the Ališaukas, Jucys and Jucys form (2.14) for the reduced Wigner coefficient in (2.17b).

Two remarks should be made. The first is that condition (2.26) implies that the only restriction on the left-hand sum in (2.27) is that $\sum_{i=1}^{n} y_i = p_3$. Also the sum on the right-hand side of (2.27) terminates at $y_i=(m)_i-(m_{14})_i$ for $2 \leqq i \leqq n$ (or possibly smaller values of y_2 and y_n).

Now fix $(m)_i-(m_{14})_i$ for $2 \leqq i \leqq n$ and fix p_3 but let m_1, m_{14}, $(m)_1$, p_2 and p_4 vary subject to the condition (2.26) and also that

$$(2.28) \qquad \sum_{i=1}^{n} [(m_{14})_i-(m_1)_i]=p_4, \qquad \sum_{i=1}^{n} [(m)_i-(m_{14})_i]=p_2+p_3.$$

The identity (2.27) is valid for all nonnegative integers $(m)_1$, $(m_1)_i$, $(m_{14})_i$ for $1 \leqq i \leqq n$ satisfying some condition:

$$(2.29) \quad (m)_1>(m_{14})_1+C_1>(m_1)_1+C_2>(m_{14})_2+C_3>(m_1)_2+C_4>\cdots>(m_1)_n+C_{2n}$$

for some constants $C_1, \cdots, C_{2n}>0$. For simplicity we shall also assume $(m)_1-(m_{14})_1>p_2$.

Note that the factors in front of the summation sign on the right-hand side of (2.27) may be written as

$$(2.30) \qquad \frac{(-1)^{p_3}}{(p_3)!} \prod_{l=2}^{n} \left[\frac{((m)_1-(m)_l+l)_{((m)_l-(m_{14})_l)}}{((m_{14})_l-(m_1)_n+n-l+1)_{((m)_l-(m_{14})_l)}} \right] \cdot \frac{(p_2+1)_{((m)_1-(m_{14})_1-p_2)} \cdot ((m_{14})_1-(m_1)_n+n)_{((m)_1-(m_{14})_1-p_2)}}{((m)_1-(m_1)_1-p_3+1)_{p_3}}.$$

We would have a similar formula if we assume $(m)_1-(m_{14}) \leqq p_2$. Also note that $(m)_1-(m_{14})_1-p_2=p_3-\sum_{i=2}^{n}((m)_i-(m_{14})_i)$ by condition (2.28), so all the subscripts of the rising factorials in (2.30) are constants.

By substituting (2.30) and clearing denominators, the identity (2.27) becomes a polynomial identity of fixed degree in the variables $(m)_1$, $(m_1)_i$, $(m_{14})_i$ for $1 \leqq i \leqq n$. It follows from (2.29) that (2.27) is true for all values of $(m)_1, (m_1)_i, (m_{14})_i \in \mathbb{C}$, $1 \leqq i \leqq n$, such that both sides of (2.27) are defined.

Now substitute $q=p_3$, $a=-(m_1)_n$, $b=-(m_1)_1-(n-1)$, $z_l=-(m)_{n-l+1}$ and $w_l=-(m_{14})_{n-l+1}$ for $1 \leqq l \leqq n$. We then obtain identity (2.25). Q.E.D.

Remark 2.31. For $n=2$ one uses formula (2.5) of [15] to translate Theorem 2.24 above into the classical terminating form of Whipple's transformation [4, eq. (4.3.4)] between a well-poised ${}_7F_6$ and a balanced ${}_4F_3$ hypergeometric series. In this form the parameter q is no longer restricted to be a nonnegative integer, but may take on arbitrary complex values (subject to the usual condition that denominators do not vanish).

Remark 2.32. On the left-hand side of identity (2.25) the $W^{(n)}$ series satisfies Holman's "well-poised in $SU(n)$" conditions (2.22) (also [15, eqs. (3.2)]. On the right-hand side the $F^{(n-1)}$ series satisfies a generalization of the "balanced" or "Saalschützian" condition. With notation as in (2.21) and (2.23a and b) this balanced condition is

$$\sum_{l=1}^{n-1} a_{il}+n=\sum_{l=1}^{n-1} b_{il}, \tag{2.33}$$

for each i, $1 \leqq i \leqq n-1$. Note that the $F^{(n)}$ series in Holman's generalization of the Saalschütz summation theorem [14, Thm. 3] is "balanced" in this sense just as in the classical case.

3. Corollaries of Theorem 2.24. In this section we will discuss a generalization of Dougall's theorem [11], an analogue of Dougall's theorem, limiting cases of Theorem 2.24 and new summation theorems for classical hypergeometric series and basic hypergeometric series.

If we set $q=b-a$ in identity (2.25), then the right-hand side of (2.25) can be summed by means of Holman's generalization of the Saalschütz summation theorem [14]. By continuation we may drop the assumption that w_i-z_i are integers, $1 \leqq i \leqq n-1$. We obtain a generalization of Dougall's theorem [14, eq. (4.3.5)].

COROLLARY 3.1. *With notation as in Theorem* 2.24 *let q be a nonnegative integer and set $s=\sum_{k=1}^{n}(w_k-z_k)$. For $n \geqq 2$ we have*

$$W_q^{(n)}\left(\begin{array}{c}
\begin{array}{llll}
z_1-z_2+1 & & & \\
z_1-z_3+2 & z_2-z_3+1 & & \\
\vdots & \vdots & & \\
z_1-z_n+n-1 & z_2-z_n+n-2 & \cdots & z_{n-1}-z_n+1
\end{array}\\
\left|\begin{array}{ccccc}
z_1-w_1 & z_1-w_2+1 & \cdots & z_1-w_n+n-1 & z_1+s-b\\
z_2-w_1-1 & z_2-w_2 & \cdots & z_2-w_n+n-2 & z_2+s-b-1\\
\vdots & \vdots & & \vdots & \vdots\\
z_n-w_1-n+1 & z_n-w_2-n+2 & \cdots & z_n-w_n & z_n+s-b-n+1
\end{array}\right.\\
\left|\begin{array}{ccccc|c}
1 & z_1-z_2+2 & \cdots & z_1-z_n+n & z_1-b & -1\\
z_2-z_1 & 1 & \cdots & z_2-z_n+n-1 & z_2-b-1 & -1\\
\vdots & \vdots & & \vdots & \vdots & \vdots\\
z_n-z_1-n+2 & z_n-z_2-n+3 & \cdots & 1 & z_n-b-n+1 & -1
\end{array}\right.
\end{array}\right) \tag{3.2}$$

$$=\frac{\Gamma(1+s)\Gamma(b-z_n+n-q)\Gamma(z_n-n+1-b+s)\Gamma(b-z_n+n-s)}{\Gamma(1+s-q)\Gamma(b-z_n+n)\Gamma(w_n-n+1-b)\Gamma(b-w_n+n-q)}$$
$$\cdot\prod_{l=1}^{n-1}\left[\frac{\Gamma(w_l+q-b-l+1)\Gamma(b-w_l+l)}{\Gamma(z_l+q-b-l+1)\Gamma(b-z_l+l)}\right].$$

Proof. We have used

$$\begin{aligned}
&\prod_{l=1}^{n-1}\left[\frac{\Gamma(w_l-z_n+n-l+1)\Gamma(z_n-w_l+l-n)\Gamma(b-z_l+l-q)}{\Gamma(z_l-z_n+n-l+1)\Gamma(z_n-z_l+l-n)\Gamma(b-w_l+l-q)}\right]\\
&\quad=\prod_{l=1}^{n-1}\left[\frac{\Gamma(w_l+q-b-l+1)}{\Gamma(z_l+q-b-l+1)}\right],
\end{aligned} \tag{3.3}$$

where we assume z_i-z_j is not an integer for $1 \leqq i \neq j \leqq n-1$.

An analogue of Dougall's theorem is obtained if we set $q=s=\sum_{k=1}^{n}(w_k-z_k)$.

Corollary 3.4. *With notation as in Theorem* 2.24 *let q and $w_i - z_i$ for $i = 1, \cdots, n-1$ be nonnegative integers and let $q = s = \sum_{K=1}^{n} (w_k - z_k)$. For $n \geqq 2$ we have*

$$\begin{array}{l}
W_q^{(n)}\left(\begin{array}{llll}
z_1-z_2+1 & & & \\
z_1-z_3+2 & z_2-z_3+1 & & \\
\vdots & \vdots & \ddots & \\
z_1-z_n+n-1 & z_2-z_n+n-2 & \cdots & z_{n-1}-z_n+1
\end{array}\right. \\
\qquad\left|\begin{array}{ccccc}
z_1-w_1 & z_1-w_2+1 & \cdots & z_1-w_n+n-1 & z_1-a \\
z_2-w_1-1 & z_2-w_2 & \cdots & z_2-w_n+n-2 & z_2-a-1 \\
\vdots & \vdots & & \vdots & \vdots \\
z_n-w_1-n+1 & z_n-w_2-n+2 & \cdots & z_n-w_n & z_n-a-n+1
\end{array}\right. \\
\qquad\left|\begin{array}{ccccc|c}
1 & z_1-z_2+2 & \cdots & z_1-z_n+n & z_1-b & -1 \\
z_2-z_1 & 1 & \cdots & z_2-z_n+n-1 & z_2-b-1 & -1 \\
\vdots & \vdots & & \vdots & \vdots & \vdots \\
z_n-z_1-n+2 & z_n-z_2-n+3 & \cdots & 1 & z_n-b-n+1 & -1
\end{array}\right) \\
\quad = q! \prod_{l=1}^{n} \left[\dfrac{\Gamma(a-z_l+l)\Gamma(b-w_l+l)}{\Gamma(a-w_l+l)\Gamma(b-z_l+l)}\right].
\end{array} \tag{3.5}$$

Proof. With $q = s$ we can apply Holman's Saalschütz theorem to the right-hand side of (2.25). We then obtain (3.5) after using the following identity:

$$\begin{aligned}
&\frac{\Gamma(z_n-n+1+q-b)}{\Gamma(w_n-n+1-b)} \prod_{l=1}^{n-1} \left[\frac{\Gamma(w_l-z_n+n-l+1)\Gamma(z_n-w_l+l-n)}{\Gamma(z_l-z_n+n-l+1)\Gamma(z_n-z_l+l-n)}\right] \\
&\quad = \frac{\Gamma(b-w_n+n)}{\Gamma(b-z_n+n-q)},
\end{aligned} \tag{3.6}$$

where we assume $z_i - z_j$ is not an integer for $1 \leqq i \neq j \leqq n-1$.

Remark 3.7. For $n = 2$, Corollary 3.1 reduces to the classical Dougall's theorem when the $W_q^{(2)}$ series is replaced by the corresponding well-poised ${}_7F_6$ hypergeometric series by means of formula (2.5) of [15]. In this classical form q is no longer restricted to be a nonnegative integer, but may take on arbitrary complex values (if $z_1 - w_1$ is a negative integer).

In Corollary 3.4 for $n = 2$, the well-poised ${}_7F_6$ series corresponding to the $W_q^{(2)}$ series is in general not defined because the integer $z_1 - w_1 + 1 \leqq 1$ is a denominator parameter in the ${}_7F_6$ series. Hence Corollary 3.4 is not a generalization of a classical result, but rather an analogue of Dougall's theorem.

Limiting cases of identity (2.25) also yield transformation formulas for hypergeometric series in $U(n)$. For example, taking the limit $w_n \to \infty$ in (2.25) gives a form of Holman's generalization of the terminating ${}_5F_4(1)$ summation theorem ([14, Thm. 4] and see also [20]). By taking the limits $a \to \infty$ or $b \to \infty$ we obtain generalizations of the terminating form of Whipple's ${}_6F_5(-1)$ transformation theorem ([4, eq. (4.4.2)] and [25]).

Taking the limit $a \to \infty$ in (2.25), we find

Corollary 3.8. *With notation and assumptions as in Theorem* 2.24, *we have*

$$W_q^{(n)}\left(\begin{array}{llll}
z_1-z_2+1 & & & \\
z_1-z_3+2 & z_2-z_3+1 & & \\
\vdots & \vdots & \ddots & \\
z_1-z_n+n-1 & z_2-z_n+n-2 & \cdots & z_{n-1}-z_n+1
\end{array}\right.$$

$$\left|\begin{array}{cccc} z_1-w_1 & z_1-w_2+1 & \cdots & z_1-w_n+n-1 \\ z_2-w_1-1 & z_2-w_2 & \cdots & z_2-w_n+n-2 \\ \vdots & \vdots & & \vdots \\ z_n-w_1-n+1 & z_n-w_2-n+2 & \cdots & z_n-w_n \end{array}\right.$$

$$\left|\begin{array}{ccccc|c} 1 & z_1-z_2+2 & \cdots & z_1-z_n+n & z_1-b & 1 \\ z_2-z_1 & 1 & \cdots & z_2-z_n+n-1 & z_2-b-1 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ z_n-z_1-n+2 & z_n-z_2-n+3 & \cdots & 1 & z_n-b-n+1 & 1 \end{array}\right)$$

$$(3.9)\qquad =\frac{\Gamma(w_n-z_n+1)}{\Gamma(1+s-q)}\,\frac{\Gamma(b-z_n+n-q)}{\Gamma(b-z_n+n)}\prod_{l=1}^{n-1}\left[\frac{\Gamma(w_l-z_n+n-l+1)}{\Gamma(z_l-z_n+n-l+1)}\right]$$

$$\cdot F^{(n-1)}\left(\begin{array}{cccc} z_1-z_2+1 & & & \\ z_1-z_3+2 & z_2-z_3+1 & & \\ \vdots & \vdots & & \\ z_1-z_{n-1}+n-2 & z_2-z_{n-1}+n-3 & \cdots & z_{n-2}-z_{n-1}+1 \end{array}\right.$$

$$\left|\begin{array}{ccccc} z_1-w_1 & z_1-w_2+1 & \cdots & z_1-w_n+n-1 & z_1+q-b \\ z_2-w_1-1 & z_2-w_2 & \cdots & z_2-w_n+n-2 & z_2+q-b-1 \\ \vdots & \vdots & & \vdots & \vdots \\ z_{n-1}-w_1-n+2 & z_{n-1}-w_2-n+3 & \cdots & z_{n-1}-w_n+1 & z_{n-1}+q-b-n+2 \end{array}\right.$$

$$\left|\begin{array}{ccccc|c} 1 & z_1-z_2+2 & \cdots & z_1-z_n+n & z_1-b & 1 \\ z_2-z_1 & 1 & \cdots & z_2-z_n+n-1 & z_2-b-1 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ z_{n-1}-z_1-n+3 & z_{n-1}-z_2-n+4 & \cdots & z_{n-1}-z_n+2 & z_{n-1}-b-n+2 & 1 \end{array}\right).$$

Now taking the limit $b\to\infty$ in (2.25), we find

Corollary 3.10. *With notation and assumptions as in Theorem* 2.24, *we have*

$$W_q^{(n)}\left(\textit{as above}\left|\begin{array}{cccc} z_1-w_1 & \cdots & z_1-w_n+n-1 & z_1+s-q-a \\ \vdots & & \vdots & \vdots \\ z_n-w_1-n+1 & \cdots & z_n-w_n & z_n+s-q-a-n+1 \end{array}\right.\right.$$

$$\left|\begin{array}{ccc|c} 1 & \cdots & z_1-z_n+n & 1 \\ \vdots & & \vdots & \vdots \\ z_n-z_1-n+2 & \cdots & 1 & 1 \end{array}\right)$$

(3.11)

$$=\frac{\Gamma(a-s+q-z_n+n)\Gamma(w_n-z_n+1)}{\Gamma(a-w_n+n)\Gamma(1+s-q)}\prod_{l=1}^{n}\left[\frac{\Gamma(w_l-z_n+n-l+1)\Gamma(a-z_l+l)}{\Gamma(z_l-z_n+n-l+1)\Gamma(a-w_l+l)}\right]$$

$$\cdot F^{(n-1)}\left(\textit{as above}\left|\begin{array}{cccc} z_1-w_1 & \cdots & z_1-w_n+n-1 & z_1+s-q-a \\ \vdots & & \vdots & \vdots \\ z_{n-1}-w_1-n+2 & \cdots & z_{n-1}-w_n+1 & z_{n-1}+s-q-a-n-2 \end{array}\right.\right.$$

$$\left|\begin{array}{cccc|c} 1 & \cdots & z_1-z_n+n & z_1-a & 1 \\ \vdots & & \vdots & \vdots & \vdots \\ z_{n-1}-z_1-n+3 & \cdots & z_{n-1}-z_n+2 & z_{n-1}-a-n+2 & 1 \end{array}\right).$$

Remark 3.12. We also obtain interesting summation theorems on taking the limit $a \to \infty$ in identity (3.5) or the limit $b \to \infty$ in identities (3.5) and (3.2). After dividing by $q!$, if we sum with respect to q both sides of the limit of identity (3.2) we obtain a generalization of the binomial theorem (see also Milne [21]).

Now divide both sides of identity (3.2) by $q!$ and sum over $q \geqq 0$. The left-hand side becomes a terminating $F^{(n)}$ series which is summable by Holman's generalization of the terminating form of the Gauss summation theorem [14, Thm. 1]. The right-hand side becomes a classical ${}_{n+1}F_n(1)$ series such that the sum of the numerator parameters equals the sum of the denominator parameters. This is for $n \geqq 2$. For $n=1$ we obtain the same result by Vandermonde's theorem or by the classical Gauss summation theorem [4]. Thus we have

THEOREM 3.13. *For $n \geqq 1$, let $b_l \in \mathbb{C}$ and K_l be a nonnegative integer for $l=1, 2, \cdots, n$. We assume that b_l is not a negative integer or zero for $1 \leqq l \leqq n$. Setting $c=\sum_{l=1}^n K_l$, we have*

$$(3.14) \qquad {}_{n+1}F_n\left(\begin{matrix} -c, b_1+K_1, b_2+K_2, \cdots, b_n+K_n; \\ b_1, b_2, \cdots, b_n; \end{matrix} \; 1\right) = (-1)^c c! \prod_{l=1}^{n} \frac{1}{(b_l)_{K_l}}.$$

We also prove a "q-analogue" of Theorem 3.13. First, we give a definition.

DEFINITION 3.15. Define the basic hypergeometric series

$$(3.16) \qquad {}_{r+1}\varphi_r\left(\begin{matrix} a_1, \cdots, a_{r+1} \\ b_1, \cdots, b_r \end{matrix} ; q, x\right) = \sum_{n=0}^{\infty} \frac{(a_1; q)_n \cdots (a_{r+1}; q)_n x^n}{(b_1; q)_n \cdots (b_r; q)_n (q; q)_n}$$

with

$$(3.17) \qquad (a; q)_n = \begin{cases} 1, & n=0, \\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1, 2, \cdots. \end{cases}$$

THEOREM 3.18. *For $n \geqq 1$, let $\beta_l \in \mathbb{C}$ and $\gamma_l = q^{K_l}$ for $1 \leqq l \leqq n$, where K_l is a nonnegative integer and β_l is not a negative power of q or 1. Setting $\sigma = \prod_{l=1}^n \gamma_l = q^c$ where $c = \sum_{l=1}^n K_l$, we have*

$$(3.19) \qquad {}_{n+1}\phi_n\left(\begin{matrix} \sigma^{-1}, \beta_1\gamma_1, \beta_2\gamma_2, \cdots, \beta_n\gamma_n; \\ \beta_1, \beta_2, \cdots, \beta_n; \end{matrix} \; q, 1\right) = \frac{(\sigma^{-1}; q)_c}{\prod_{l=1}^n (\beta_l; q)_{K_l}}.$$

Proof. We first prove the following identity:

$$(3.20) \qquad {}_1\phi_0(\sigma^{-1}; q, q^r) = \sum_{j=0}^{c} \frac{(\sigma^{-1}; q)_j}{(q; q)_j} q^{rj} = \begin{cases} 0, & 0 < r \leqq c, \\ (\sigma^{-1}; q)_c, & r=0. \end{cases}$$

From the q-binomial theorem [4, eq. (8.2.4)], we have

$$(3.21) \qquad {}_1\phi_0(a; q, z) = \prod_{j=0}^{\infty} \frac{(1-aq^j z)}{(1-q^j z)}.$$

Hence ${}_1\phi_0(\sigma^{-1}; q, q^r) = 0$ if $0 < r \leqq c$. We also have [4, p. 66]

$$(3.22) \qquad {}_1\phi_0(a; q, z) - a\, {}_1\phi_0(a; q, qz) = (1-a)\, {}_1\phi_0(aq; q, z).$$

Setting $a=\sigma^{-1}$ and $z=1$ in (3.22), we find

$$(3.23)\qquad {}_1\phi_0(q^{-c};q,1)=(1-q^c)\,{}_1\phi_0(q^{-c+1};q,1)$$

and by induction

$$(3.24)\qquad {}_1\phi_0(q^{-c};q,1)=(1-q^{-c})(1-q^{-c+1})\cdots(1-q^{-1})\,{}_1\phi_0(1;q,1)$$

where ${}_1\phi_0(1;q,1)=1$. This completes the proof of identity (3.20).

Using the identity

$$(3.25)\qquad \frac{(\beta_l\gamma_l;q)_j}{(\beta_l;q)_j}(\beta_l;q)_{K_l}=(\beta_lq^j;q)_{K_l},$$

for $1\leqq l\leqq n$ and $j\geqq 0$, we rewrite (3.19) as

$$(3.26)\qquad \sum_{j=0}^{c}\frac{(\sigma^{-1};q)_j}{(q;q)_j}(\beta_1q^j;q)_{K_1}\cdots(\beta_nq^j;q)_{K_n}=(\sigma^{-1};q)_c.$$

Now if we expand $\prod_{l=1}^n(\beta_lq^j;q)_{K_l}$ in powers of q^j, we find

$$(3.27)\qquad \prod_{l=1}^{n}(\beta_lq^j;q)_{K_l}=1+\text{higher order terms in }q^j,$$

with highest order q^{jc}. An application of identity (3.20) completes the proof of Theorem 3.18.

Remark 3.28. There is a similar proof of Theorem 3.13 relying on a classical (ordinary) analogue of (3.20), which can be proved by induction from the ordinary binomial theorem. This classical analogue of (3.20) is given in an equivalent form by Project Ališaukas, Jucys and Jucys [1].

We also mention that Milne had discussed how a generalization of Dougall's theorem might lead to an identity similar to (3.14), just as in fact occurred. His conjectured identities are different than (3.14) [22].

4. Orthogonal polynomials in several variables. In this section we define a family of orthogonal polynomials in several variables which generalize the Racah polynomials of Wilson [26], [27]. These polynomials are orthogonal on a discrete set $\{(x_1,\cdots,x_n)\in\mathbb{Z}^n\,|\,x_i\geqq 0$ for $1\leqq i\leqq n$ and $\sum_{i=1}^n x_i=N\}$ for some nonnegative integer N and for $n\geqq 2$.

Recall the orthogonality relations (1.26a and b) satisfied by the multiplicity-free Racah coefficients above. Let notation be as in Proposition 1.25 and let $n\geqq 2$. We will denote $\lambda=[\lambda_1,\cdots,\lambda_n]$ for the highest weight λ of any irreducible representation space V_λ. Set $z_i=\mu_i'-m_i$ and $h_i=\nu_i-\mu_i$ for $1\leqq i\leqq n$ and denote $N=p''=\sum_{i=1}^n z_i=\sum_{i=1}^n h_i$. Finally we assume that $\nu_i>N+m_i$ and $m_i>N+\nu_{(i+1)}$ for $1\leqq i\leqq n$ and with $\nu_{(n+1)}=0$.

When we substitute the explicit form (2.4) and (2.14) of the Racah coefficients into identity (1.26a), we obtain

$$(4.1)\qquad \sum_{z_1+\cdots+z_n=N}\phi_\mu(\nu,m,m',m''|\mu')\phi_{\bar\mu}(\nu,m,m',m''|\mu')w(z)=\delta_{\bar\mu,\mu}M_\mu$$

where

$$\delta_{\bar\mu,\mu}=\begin{cases}1 & \text{if }\mu=\bar\mu,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$
\begin{aligned}
w(z)=&\prod_{1\leqq k<l\leqq n}(z_k+m_k-(z_l+m_l)+l-k)\\
&\cdot\prod_{1\leqq k<l\leqq n}\left[\frac{(m_k-\nu_l+l-k)_{z_k}}{(m_k-m_l+l-k+1)_{z_k}(m_l-m_k+k-l)_{(z_l+1)}}\right]\\
&\cdot\prod_{2\leqq k\leqq l\leqq n}(m_l-\nu_k+k-l)_{z_l}\cdot\prod_{l=1}^{n}\frac{1}{(m_l-\nu_1+1-l)_{z_l}(z_l)!},
\end{aligned}\tag{4.2}
$$

and

$$
\begin{aligned}
M_\mu=&\prod_{1\leqq k<l\leqq n}((\nu_k-h_k)-(\nu_l-h_l)+l-k)^{-1}\cdot\frac{1}{(h_1)!}\\
&\cdot\prod_{l=2}^{n}\left[\frac{(\nu_l-\nu_1+1-l)_{(h_1+1)}(h_l)!}{(\nu_1-\nu_l+l)_{h_l}}\right]\cdot\prod_{1\leqq k<l\leqq n}\frac{1}{(m_k-\nu_l+l-k)_{h_l}}\\
&\cdot\prod_{1\leqq k\leqq l\leqq n}\frac{1}{(m_l-\nu_k+k-l)_{h_l}}\\
&\cdot\prod_{2\leqq k<l\leqq n}(\nu_k-\nu_l+l-k+1)_{h_l}(\nu_l-\nu_k+k-l)_{(h_k+1)},
\end{aligned}\tag{4.3}
$$

and

$$
\begin{aligned}
\phi_\mu(\nu,m,m',m''|\mu')=&\sum_{\substack{y_l\geqq 0\\ 2\leqq l\leqq n}}\left\{\prod_{2\leqq \nu<s\leqq n}\left(\frac{\nu_s-\nu_r+r-s+y_r-y_s}{\nu_s-\nu_r+r-s}\right)\right.\\
&\left.\cdot\prod_{l=2}^{n}\prod_{k=1}^{n}\left[\frac{(\mu_k-\nu_l+l-k)_{y_l}(\mu'_k-\nu_l+l-k)_{y_l}}{(\nu_k-\nu_l+l-k+1)_{y_l}(m_k-\nu_l+l-k)_{y_l}}\right]\right\},
\end{aligned}\tag{4.4}
$$

with a similar formula for $\phi_{\bar{\mu}}$.

Similarly to the argument in the proof of Theorem 2.24 it follows that identity (4.1) is true for all ν, $m\in\mathbb{C}^n$ with the restriction that the denominators in (4.2)-(4.4) should not vanish.

We now rewrite (4.1) and set $\alpha_k=-m_{n-k+1}-k$, $\beta_k=-\nu_{n-k+1}-k$, $x_k=z_{n-k+1}$, $t_k=h_{n-k+1}$ for $1\leqq k\leqq n$.

PROPOSITION 4.5. *Let $n\geqq 2$ and N be a nonnegative integer. With notation as above and $\alpha,\beta\in\mathbb{C}^n$, then for $t=(t_1,\cdots,t_n)\in\mathbb{Z}^n$ such that $\sum_{i=1}^n t_i=N$ and $t_i\geqq 0$ for $1\leqq i\leqq n$, and similarly for $t'\in\mathbb{Z}^n$, we have*

$$
\sum_{x_1+\cdots+x_n=N}P_t(x;\alpha,\beta,N)P_{t'}(x;\alpha,\beta,N)w(x)=\delta_{t,t'}M_t\tag{4.6}
$$

where x_i is a nonnegative integer for $1\leqq i\leqq n$ and

$$
\delta_{t,t'}=\begin{cases}1 & \textit{if } t=t',\\ 0 & \textit{otherwise},\end{cases}
$$

and

$$
\begin{aligned}
w(x)=&\prod_{1\leqq k<l\leqq n}((\alpha_k-x_k)-(\alpha_l-x_l))\\
&\cdot\prod_{1\leqq k<l\leqq n}\left[\frac{(\beta_k-\alpha_l)_{x_l}}{(\alpha_k-\alpha_l+1)_{x_l}(\alpha_l-\alpha_k)_{(x_k+1)}}\right]\\
&\cdot\prod_{1\leqq k\leqq l\leqq n-1}(\beta_l-\alpha_k)_{x_k}\cdot\prod_{l=1}^{n}\frac{1}{(\beta_n-\alpha_l)_{x_l}(x_l)!}N!(\beta_n-\alpha_1)_N
\end{aligned}\tag{4.7}
$$

and

$$\begin{aligned} M_t = & \prod_{1\leqq k<l\leqq n} ((t_k+\beta_k)-(t_l+\beta_l))^{-1} \\ & \cdot \prod_{1\leqq k<l\leqq n} \left[\frac{(\beta_k-\beta_l+1)_{t_k}(\beta_l-\beta_k)_{(t_l+1)}}{(\beta_k-\alpha_l)_{t_k}}\right] \\ & \cdot \prod_{2\leqq k\leqq l\leqq n} \frac{1}{(\beta_l-\alpha_k)_{t_l}} \cdot \prod_{l=1}^{n} (\beta_l-\alpha_1)_{t_l}(t_l)! \frac{1}{N!(\beta_n-\alpha_1)_N} \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} P_t(x;\alpha,\beta,N) = & \frac{\prod_{l=1}^{n}(\beta_l-\beta_n+1)_{t_l}(\beta_l-\alpha_1)_{t_l}}{N!(\beta_n-\alpha_1)_N} \\ & \cdot F^{(n-1)}\left(\begin{array}{l} \begin{array}{cccc} \beta_1-\beta_2 & & & \\ \beta_1-\beta_3 & \beta_2-\beta_3 & & \\ \vdots & \vdots & & \\ \beta_1-\beta_{n-1} & \beta_2-\beta_{n-1} & \cdots & \beta_{n-2}-\beta_{n-1} \end{array} \\ \left|\begin{array}{cccccc} -t_1 & \cdots & -t_n-\beta_n+\beta_1 & x_1-\alpha_1+\beta_1 & \cdots & x_n-\alpha_n+\beta_1 \\ -t_1-\beta_1+\beta_2 & \cdots & -t_n-\beta_n+\beta_2 & x_1-\alpha_1+\beta_2 & \cdots & x_n-\alpha_n+\beta_2 \\ \vdots & & \vdots & \vdots & & \vdots \\ -t_1-\beta_1+\beta_{n-1} & \cdots & -t_n-\beta_n+\beta_{n-1} & x_1-\alpha_1+\beta_{n-1} & \cdots & x_n-\alpha_n+\beta_{n-1} \end{array}\right. \\ \left|\begin{array}{cccccc|c} 1 & \cdots & 1-\beta_n+\beta_1 & -\alpha_1+\beta_1 & \cdots & -\alpha_n+\beta_1 & 1 \\ 1-\beta_1+\beta_2 & \cdots & 1-\beta_n+\beta_2 & -\alpha_1+\beta_2 & \cdots & -\alpha_n+\beta_2 & 1 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 1-\beta_1+\beta_{n-1} & \cdots & 1-\beta_n+\beta_{n-1} & -\alpha_1+\beta_{n-1} & \cdots & -\alpha_n+\beta_{n-1} & 1 \end{array}\right. \end{array}\right) \end{aligned} \tag{4.9}$$

and similarly for $P_{t'}(x;\alpha,\beta,N)$.

Note that $P_t(x;\alpha,\beta,N)$ is a symmetric polynomial in the variables $x_1-\alpha_1$, $x_2-\alpha_2,\cdots,x_n-\alpha_n$. Hence $P_t(x;\alpha,\beta,N)$ may be expressed as a polynomial in the elementary symmetric functions $e_i(x_1-\alpha_1,\cdots,x_n-\alpha_n)$, $1\leqq i\leqq n$. Recall that for a set of variables $u_1,\cdots,u_n$ and $1\leqq i\leqq n$, one defines

$$e_i(u_1,\cdots,u_n)=\sum_{1\leqq j_1<j_2<\cdots<j_i\leqq n} u_{j_1}u_{j_2}\cdots u_{j_i}, \tag{4.10}$$

where the sum is over all subsets $\{j_1,\cdots,j_i\}\subseteq\{1,\cdots,n\}$ of cardinality i.

We will abbreviate $e_i(x_1-\alpha_1,\cdots,x_n-\alpha_n)$ by $e_i(x-\alpha)$ for $1\leqq i\leqq n$ and also abbreviate the set of nonconstant polynomials $\{e_2(x-\alpha),\cdots,e_n(x-\alpha)\}$ by $e(x-\alpha)$. Note that $e_1(x-\alpha)=\sum_{i=1}^n x_i-\alpha_i=N-\sum_{i=1}^n \alpha_i$.

DEFINITION 4.11. With α, β, x, t and t' as in Proposition 4.5 we define the Racah polynomial in the variables $e(x-\alpha)=\{e_2(x-\alpha),\cdots. e_n(x-\alpha)\}$ by

$$R_t(e(x-\alpha);\alpha,\beta,N)\equiv P_t(x;\alpha,\beta,N) \tag{4.12}$$

and the inner product

$$\begin{aligned} & \langle R_t(e(x-\alpha);\alpha,\beta,N), R_{t'}(e(x-\alpha);\alpha,\beta,N)\rangle \\ & \equiv \sum_{x_1+\cdots+x_n=N} R_t(e(x-\alpha);\alpha,\beta,N)R_{t'}(e(x-\alpha);\alpha,\beta,N)w(x). \end{aligned} \tag{4.13}$$

Remark 4.14. For $n=2$ we obtain up to a scalar factor the orthogonal polynomials defined in (4.1) and (4.2) of Wilson [27] by substituting $m_1=\gamma$, $m_2=-\delta-N$, $\mu_1=\beta+n-N-1$, $\mu_2=-n$, $\mu_1'=x+\gamma$, $\mu_2'=-x-\delta$, $\nu_1=\beta-1$, $\nu_2=0$ and $N=-(\alpha+1)$ in the notation of equations (4.1)-(4.3) here and with α, β, γ, δ defined as in [27]. (Use transformation (1.2) of [27].)

We also remark that for $n=2$, the R_t polynomials contain the Hahn polynomials as a limiting case (see [27]). However, for $n>2$, the Hahn polynomials in several variables defined by Karlin and McGregor [16] have a weight function (see [16, eq. (5.13)]) which in general does not appear to be the limit of the weight function $w(x)$ in (4.8) here. It is also not known what relation, if any, there is between the R_t polynomials in two variables and the orthogonal polynomials defined by Suslov [30].

Two important questions need to be answered about the R_t polynomials. First, are there polynomials in several variables orthogonal with respect to a continuous measure generalizing the Wilson polynomials [27], [32] just as the R_t polynomials generalize the discrete Racah polynomials in one variable? Second, are there q-analogues of the R_t polynomials (see [31], [32])?

5. Recurrence relation for the Racah polynomials. Let $t=(t_1,\cdots,t_n)\in\mathbb{Z}^n$ such that $t_i\geqq 0$ for all i, $1\leqq i\leqq n$ and $\sum_{i=1}^n t_i=N$ for some nonnegative integer N. We fix below α, $\beta\in\mathbb{C}^n$ and N, and abbreviate $R_t(e(x-\alpha);\alpha,\beta,N)$ by R_t. R_t is a polynomial of degree $N-t_n$ in the variables $e_2(x-\alpha),\cdots,e_n(x-\alpha)$ where $x=(x_1,\cdots,x_n)\in\mathbb{Z}^n$, $x_i\geqq 0$ and $\sum_{i=1}^n x_i=N$.

DEFINITION 5.1. Let ε_j be the n-tuple $(0,\cdots,0,1,0,\cdots,0)$ with 1 in the jth entry and 0 elsewhere. Then $t+\varepsilon_j-\varepsilon_n$ is the n-tuple obtained by adding 1 to the jth entry and subtracting 1 from the nth entry.

If $t_n>0$, then for all k, $1\leqq k\leqq n-1$, we will show that R_t satisfies the following recurrence relation

$$\begin{aligned}R_{t+\varepsilon_k-\varepsilon_n}&=\sum_{j=1}^{n-1}A_{kj}^{(t)}e_{j+1}(x-\alpha)R_t+A_{kn}^{(t)}R_t+\sum_{m=1}^{n-1}B_{km}^{(t)}R_{t-\varepsilon_m+\varepsilon_k}\\&\quad+\sum_{h=1}^{n-1}B_{kh}'^{(t)}R_{t-\varepsilon_k+\varepsilon_h}+C_k^{(t)}R_{t-\varepsilon_k+\varepsilon_n}\end{aligned}\tag{5.2}$$

where $A_{kj}^{(t)}$, $B_{km}^{(t)}$, $B_{kh}'^{(t)}$,$C_k^{(t)}$ are defined in (5.10), (5.12)-(5.14) and (5.19) below and are independent of x. We also set $R_{t-\varepsilon_j}\equiv 0$ if $t_j=0$, $1\leqq j\leqq n$.

We begin the proof of (5.2) with the following well-known lemma describing those polynomials which can appear nontrivially in a recurrence relation for R_t.

LEMMA 5.3. *If* $t_n>0$ *and* $2\leqq l\leqq n$, *then*

$$e_l(x-\alpha)R_t=\sum a_{t'}R_{t'}\tag{5.4}$$

as polynomials in the variables $e_2(x-\alpha),\cdots,e_n(x-\alpha)$, *where* $a_{t'}\in\mathbb{C}$ *and the sum is over all* $t'=(t_1',\cdots,t_n')\in\mathbb{Z}^n$, $t_i'\geqq 0$ *and* $\sum_{i=1}^n t_i=N$ *such that* $t_n+1\geqq t_n'\geqq t_n-1$.

Proof. We can write

$$e_l(x-\alpha)R_t=\sum a_{t'}R_{t'}$$

where $a_{t'}\in\mathbb{C}$ and the sum is over all $t'\in\mathbb{Z}^n$, $t_i'\geqq 0$ and $\sum_{i=1}^n t_i=N$. We need to show that if $a_{t'}\neq 0$, then $t_n+1\geqq t_n'\geqq t_n-1$. We have

$$a_{t'}=\langle e_l(x-\alpha)R_t,R_{t'}\rangle(M_{t'})^{-1}.\tag{5.5}$$

If $a_{t'}\neq 0$, then $\deg(e_l(x-\alpha)R_t)\geqq\deg R_{t'}$ in the variables $e_2(x-\alpha),\cdots,e_n(x-\alpha)$, equivalently $t'_n \geqq t_n - 1$. Similarly

$$a_{t'} = \langle R_t,\, e_l(x-\alpha)R_{t'}\rangle (M_{t'})^{-1}. \tag{5.6}$$

It follows that if $a_{t'} \neq 0$, then $\deg R_t \leqq \deg R_{t'} + 1$ or equivalently $t_n + 1 \geqq t'_n$. Q.E.D.

Proof of the recurrence relation (5.2). The leading term (L.T.) in the series for R_t is

$$\begin{aligned} &\frac{(t_n)!(\beta_n-\alpha_1)_{t_n}}{N!(\beta_n-\alpha_1)_N} \prod_{1\leqq i<j\leqq n-1} \frac{(\beta_i-\beta_j+t_i-t_j)}{(\beta_i-\beta_j)} \\ &\cdot \prod_{i=1}^{n-1} \left[\prod_{l=1}^{n} (\beta_i - t_l - \beta_l)_{t_i} (\beta_i + x_l - \alpha_l)_{t_i} \prod_{l=1}^{n-1} \frac{1}{(\beta_i-\beta_l+1)_{t_i}} \prod_{l=2}^{n} \frac{1}{(\beta_i-\alpha_l)_{t_i}} \right]. \end{aligned} \tag{5.7}$$

Assume $t_n \neq 0$ and consider $R_{t+\varepsilon_k-\varepsilon_n}$ for some integer k, $1 \leqq k \leqq n-1$. After comparing the leading terms of $R_{t+\varepsilon_k-\varepsilon_n}$ and R_t we obtain

$$\begin{aligned} R_{t+\varepsilon_k-\varepsilon_n} = &\prod_{l=1}^{n-1} \left[\frac{(\beta_n+t_n)-(\beta_l+t_l)}{(\beta_n+t_n-\beta_l)} \right] \prod_{l=2}^{n} \frac{1}{(\beta_k+t_k-\alpha_l)} \\ &\cdot \frac{(\beta_n+t_n-(\beta_k+t_k)-1)}{t_n(\beta_n+t_n-\alpha_1-1)} \left\{ \sum_{m=0}^{n} (\beta_k+t_k)^{n-m} e_m(x-\alpha) R_t \right\} \\ &+ \text{terms of degree less than } N - t_n + 1 \\ &\quad \text{in the variables } e_2(x-\alpha), \cdots, e_n(x-\alpha). \end{aligned} \tag{5.8}$$

To compute lower degree terms in the recurrence relation (5.2), we must consider the next to leading terms in R_t:

$$\begin{aligned} &\frac{(t_n)!(\beta_n-\alpha_1)_{t_n}}{N!(\beta_n-\alpha_1)_N} \sum_{m=1}^{n-1} \Bigg\{ \prod_{\substack{1\leqq i<j\leqq n-1 \\ i,j\neq m}} (\beta_i-\beta_j+t_i-t_j) \prod_{i=1}^{m-1} (\beta_i-\beta_m+t_i-t_m+1) \\ &\qquad \cdot \prod_{j=m+1}^{n-1} (\beta_m-\beta_j+t_m-1-t_j) \prod_{1\leqq i<j\leqq n-1} (\beta_i-\beta_j)^{-1} \\ &\qquad \cdot \prod_{\substack{i=1 \\ i\neq m}}^{n-1} \Bigg[\prod_{l=1}^{n} (\beta_i-t_l-\beta_l)_{t_i} (\beta_i+x_l-\alpha_l)_{t_i} \\ &\qquad\qquad \cdot \prod_{l=1}^{n-1} \frac{1}{(\beta_i-\beta_l+1)_{t_i}} \prod_{l=2}^{n} \frac{1}{(\beta_i-\alpha_l)_{t_i}} \Bigg] \\ &\qquad \cdot \prod_{l=1}^{n} (\beta_m-t_l-\beta_l)_{t_m-1} (\beta_m+x_l-\alpha_l)_{t_m-1} \\ &\qquad \cdot \prod_{l=1}^{n-1} \frac{1}{(\beta_m-\beta_l+1)_{t_m-1}} \prod_{l=2}^{n} \frac{1}{(\beta_m-\alpha_l)_{t_m-1}} \\ &\qquad\qquad \cdot (\beta_m-\beta_n+t_m)(\beta_m-\alpha_1+t_m-1) \Bigg\} \\ &= \sum_{m=1}^{n-1} \Bigg\{ \frac{t_m}{(t_n+1)(\beta_n+t_n-\alpha_1)} \prod_{\substack{i=1 \\ i\neq m}}^{n-1} \left[\frac{(\beta_i-(\beta_m+t_m))}{(\beta_i+t_i-(\beta_m+t_m))} \frac{(\beta_i+t_i-(\beta_n+t_n)-1)}{(\beta_i-(\beta_n+t_n)-1)} \right] \end{aligned} \tag{5.9}$$

$$\cdot \frac{(\beta_m + t_m - (\beta_n + t_n) - 2)}{(\beta_m - (\beta_n + t_n) - 1)} (\beta_m + t_m - \beta_n)$$

$$\left. \cdot (\beta_m + t_m - \alpha_1 - 1) \cdot (\text{L.T. of } R_{t-\varepsilon_m+\varepsilon_n}) \right\}.$$

Now let $(A_{ij}^{(t)})$ be the $n \times n$ matrix with

$$A_{nj}^{(t)} = \begin{cases} 1 & \text{if } j = n, \\ 0 & \text{otherwise} \end{cases} \tag{5.10a}$$

and

$$\begin{aligned} A_{ij}^{(t)} = \prod_{l=1}^{n-1} \left[\frac{(\beta_n + t_n) - (\beta_l + t_l)}{(\beta_n + t_n - \beta_l)} \prod_{l=2}^{n} \frac{1}{(\beta_i + t_i - \alpha_l)} \frac{(\beta + t_n - (\beta_i + t_i) - 1)}{t_n(\beta_n + t_n - \alpha_1 - 1)} \right. \\ \cdot \begin{cases} (\beta_i + t_i)^n + (\beta_i + t_i)^{n-1} e_1(x - \alpha) & \text{if } i < n \text{ and } j = n, \\ (\beta_i + t_i)^{n-j-1} & \text{if } i < n \text{ and } j < n. \end{cases} \end{aligned} \tag{5.10b}$$

It follows that if $t_n > 0$ and $1 \leqq k \leqq n - 1$, then

$$R_{t+\varepsilon_k-\varepsilon_n} - \sum_{j=1}^{n-1} A_{kj}^{(t)} e_{j+1}(x - \alpha) R_t - A_{kn}^{(t)} R_t$$

$$= \sum_{\substack{m=1 \\ m \neq k}}^{n-1} \left\{ \frac{t_m}{t_n(\beta_n + t_n - \alpha_1 - 1)} \prod_{\substack{i=1 \\ i \neq m,k}}^{n-1} \left[\frac{(\beta_i - (\beta_m + t_m))}{(\beta_i + t_i - (\beta_m + t_m))} \frac{(\beta_i + t_i - (\beta_n + t_n))}{(\beta_i - (\beta_n + t_n))} \right] \right.$$

$$\cdot \frac{(\beta_k - (\beta_m + t_m))(\beta_k + t_k - (\beta_n + t_n) + 1)}{(\beta_k + t_k - (\beta_m + t_m) + 1)(\beta_k - (\beta_n + t_n))}$$

$$\left. \cdot \frac{(\beta_m + t_m - (\beta_n + t_n) - 1)}{(\beta_m - (\beta_n + t_n))} (\beta_m + t_m - \beta_n)(\beta_m + t_m - \alpha_1 - 1) R_{t-\varepsilon_m+\varepsilon_k} \right\}$$

$$+ \frac{(t_k + 1)}{t_n(\beta_n + t_n - \alpha_1 - 1)} \prod_{\substack{i=1 \\ i \neq k}}^{n-1} \left[\frac{(\beta_i - (\beta_k + t_k) - 1)}{(\beta_i + t_i - (\beta_k + t_k) - 1)} \frac{(\beta_i + t_i - (\beta_n + t_n))}{(\beta_i - (\beta_n + t_n))} \right] \tag{5.11}$$

$$\cdot \frac{(\beta_k + t_k - (\beta_n + t_n))}{(\beta_k - (\beta_n + t_n))} (\beta_k + t_k - \beta_n + 1)(\beta_k + t_k - \alpha_1) R_t$$

$$+ \sum_{h,j,m=1}^{n-1} \left\{ \frac{-t_m}{(t_n + 1)(\beta_n + t_n - \alpha_1)} \prod_{\substack{i=1 \\ i \neq m}}^{n-1} \left[\frac{(\beta_i - (\beta_m + t_m))}{(\beta_i + t_i - (\beta_m + t_m))} \frac{(\beta_i + t_i - (\beta_n + t_n) - 1)}{(\beta_i - (\beta_n + t_n) - 1)} \right] \right.$$

$$\cdot \frac{(\beta_m + t_m - (\beta_n + t_n) - 2)}{(\beta_m - (\beta_n + t_n) - 1)} (\beta_m + t_m - \beta_n)(\beta_m + t_m - \alpha_1 - 1)$$

$$\left. \cdot A_{kj}^{(t)} (A^{(t-\varepsilon_m+\varepsilon_n)})_{jh}^{-1} R_{t-\varepsilon_m+\varepsilon_h} \right\}$$

$+$ terms of degree less than $N - t_n$

in the variables $e_2(x - \alpha), \cdots, e_n(x - \alpha)$,

where $(A^{(t-\varepsilon_m+\varepsilon_n)})_{jh}^{-1}$ is the (j, h) entry of the matrix inverse of $A^{(t-\varepsilon_m+\varepsilon_n)}$.

Now observe that if $t_m > 0$, then

(5.12)

$$\sum_{j=1}^{n-1} A_{kj}^{(t)}(A^{(t-\varepsilon_m+\varepsilon_n)})_{jh}^{-1}$$

$$= \prod_{l=1}^{n-1}\left[\frac{(\beta_n+t_n-(\beta_l+t_l))(\beta_n+t_n-\beta_l+1)}{(\beta_n+t_n-\beta_l)}\right]$$

$$\cdot \prod_{\substack{l=1\\ l\neq m}}^{n-1}\left[\frac{1}{((\beta_n+t_n)-(\beta_l+t_l)+1)}\right]\frac{1}{((\beta_n+t_n)-(\beta_m+t_m)+2)}\frac{(t_n+1)}{t_n}\frac{(\beta_n+t_n-\alpha_1)}{(\beta_n+t_n-\alpha_1-1)}$$

$$\cdot\begin{cases} 0 \quad \text{if } k\neq m \text{ and } h\neq k, \\[2mm] \dfrac{(\beta_n+t_n-(\beta_k+t_k)-1)}{(\beta_n+t_n-(\beta_k+t_k))} \quad \text{if } k\neq m \text{ and } h=k, \\[2mm] \dfrac{(\beta_n+t_n-(\beta_k+t_k)-1)}{(\beta_n+t_n-(\beta_k+t_k)+1)}\displaystyle\prod_{l=2}^{n}\frac{(\beta_k+t_k-\alpha_l+1)}{(\beta_k+t_k-\alpha_l)}\prod_{\substack{l=1\\ l\neq k}}^{n-1}\frac{(\beta_k+t_k-(\beta_l+t_l))}{(\beta_k+t_k-(\beta_l+t_l)-1)} \\ \hfill \text{if } k=m \text{ and } h=k, \\[2mm] \dfrac{(\beta_n+t_n-(\beta_k+t_k)-1)}{(\beta_n+t_n-(\beta_h+t_h))}\displaystyle\prod_{l=2}^{n}\frac{(\beta_h+t_h-\alpha_l)}{(\beta_k+t_k-\alpha_l)}\prod_{\substack{l=1\\ l\neq k,h}}^{n-1}\frac{(\beta_k+t_k)-(\beta_l+t_l)}{(\beta_h+t_h)-(\beta_l+t_l)} \\ \quad\cdot\dfrac{1}{(\beta_n+t_n-(\beta_k+t_k)+1)} \quad \text{if } k=m \text{ and } h\neq k, \end{cases}$$

where we have used the fact that, except for their nth rows and columns, both matrices $A^{(t)}$ and $A^{(t-\varepsilon_m+\varepsilon_n)}$ are the products of diagonal matrix and a Vandermonde matrix.

For $1\leqq k$, $m\leqq n-1$and $k\neq m$, we set

$$B_{km}^{(t)} = \frac{t_m}{t_n(\beta_n+t_n-\alpha_1-1)}\prod_{\substack{i=1\\ i\neq m,k}}^{n-1}\frac{(\beta_i-(\beta_m+t_m))(\beta_i+t_i-(\beta_n+t_n))}{(\beta_i+t_i-(\beta_m+t_m))(\beta_i-(\beta_n+t_n))}$$

$$\cdot\frac{(\beta_k-(\beta_m+t_m))(\beta_k+t_k-(\beta_n+t_n)+1)(\beta_m+t_m-(\beta_n+t_n)-1)}{(\beta_k+t_k-(\beta_m+t_m)+1)(\beta_k-(\beta_n+t_n))(\beta_m-(\beta_n+t_n))}$$

$$\text{(5.13a)} \qquad \cdot(\beta_m+t_m-\beta_n)(\beta_m+t_m-\alpha_1-1)-\frac{t_m}{(t_n+1)(\beta_n+t_n-\alpha_1)}$$

$$\cdot\prod_{\substack{i=1\\ i\neq m}}^{n-1}\left[\frac{(\beta_i-(\beta_m+t_m))}{(\beta_i+t_i-(\beta_m+t_m))}\frac{(\beta_i+t_i-(\beta_n+t_n)-1)}{(\beta_i-(\beta_n+t_n)-1)}\right]$$

$$\cdot\frac{(\beta_m+t_m-(\beta_n+t_n)-2)}{(\beta_m-(\beta_n+t_n)-1)}(\beta_m+t_m-\beta_n)$$

$$\cdot\frac{(\beta_m+t_m-\alpha_1-1)}{(\beta_n+t_n-(\beta_k+t_k))}(\beta_n+t_n-(\beta_k+t_k)-1).$$

If $1\leqq k\leqq n-1$, then set

$$B_{kk}^{(t)} = \frac{(t_k+1)}{t_n(\beta_n+t_n-\alpha_1-1)}\prod_{\substack{i=1\\ i\neq k}}^{n-1}\left[\frac{(\beta_i-(\beta_k+t_k)-1)}{(\beta_i+t_i-(\beta_k+t_k)-1)}\frac{(\beta_i+t_i-(\beta_n+t_n))}{(\beta_i-(\beta_n+t_n))}\right]$$

$$\text{(5.13b)} \qquad \cdot\frac{(\beta_k+t_k-(\beta_n+t_n))}{(\beta_k-(\beta_n+t_n))}(\beta_k+t_k-\beta_n+1)(\beta_k+t_k-\alpha_1).$$

Similarly for $1 \leqq k,\ h \leqq n-1$ and $k = m$, we set

$$B'^{(t)}_{kh} = \frac{-t_k}{(t_n+1)(\beta_n+t_n-\alpha_1)} \prod_{\substack{i=1 \\ i \neq k}}^{n-1} \left[\frac{(\beta_i-(\beta_k+t_k))}{(\beta_i+t_i-(\beta_k+t_k))} \frac{(\beta_i+t_i-(\beta_n+t_n)-1)}{(\beta_i-(\beta_n+t_n)-1)} \right]$$

$$\cdot \frac{(\beta_k+t_k-(\beta_n+t_n)-2)}{(\beta_k-(\beta_n+t_n)-1)} (\beta_k+t_k-\alpha_1-1) \left(\sum_{j=1}^{n-1} A^{(t)}_{kj} (A^{(t-\varepsilon_k+\varepsilon_n)})^{-1}_{jh} \right). \tag{5.14}$$

Now let $t' = (t'_1, \cdots, t'_n) \in \mathbb{Z}^n$ where $t'_i \geqq 0$ for all $1 \leqq i \leqq n$, $\sum_{i=1}^n t'_i = N$ and $t'_n = t_n + 1$. Then

$$\left\langle R_{t'}, R_{t+\varepsilon_k-\varepsilon_n} - \sum_{j=1}^{n-1} A^{(t)}_{kj} e_{j+1}(x-\alpha) R_t - A^{(t)}_{kn} R_t - \sum_{m=1}^{n-1} B^{(t)}_{km} R_{t-\varepsilon_m+\varepsilon_k} - \sum_{h=1}^{n-1} B'^{(t)}_{kh} R_{t-\varepsilon_k+\varepsilon_h} \right\rangle$$

$$= -\sum_{j=1}^{n-1} \langle R_{t'}, A^{(t)}_{kj} e_{j+1}(x-\alpha) R_t \rangle \tag{5.15}$$

$$= -\sum_{j=1}^{n-1} A^{(t)}_{kj} \langle e_{j+1}(x-\alpha) R_{t'}, R_t \rangle \tag{5.16}$$

$$= -\sum_{j,h=1}^{n-1} A^{(t)}_{kj} (A^{(t')})^{-1}_{jh} \langle R_{t'+\varepsilon_h-\varepsilon_n}, R_t \rangle. \tag{5.17}$$

From (5.12) it follows that $t' = t - \varepsilon_k + \varepsilon_n$ and expression (5.17)

$$= -\sum_{j=1}^{n-1} A^{(t)}_{kj} (A^{(t-\varepsilon_k+\varepsilon_n)})^{-1}_{jk} M_t, \tag{5.18}$$

where $\sum_{j=1}^{n-1} A^{(t)}_{kj} (A^{(t-\varepsilon_k+\varepsilon_n)})^{-1}_{jk}$ is computed in (5.12) and M_t in (4.8).

For $1 \leqq k \leqq n-1$ and $t_k \neq 0$ we set

$$C^{(t)}_k = -M_t (M_{t-\varepsilon_k+\varepsilon_n})^{-1} \sum_{j=1}^{n-1} A^{(t)}_{kj} (A^{(t-\varepsilon_k+\varepsilon_n)})^{-1}_{jk}. \tag{5.19}$$

If $t_k = 0$, set $C^{(t)}_k \equiv 0$.

Applying Lemma 5.3, this completes the proof of the recurrence relation (5.2).

6. Duality for Racah polynomials. The main result in this section is

PROPOSITION 6.1. *With notation as in Proposition* 4.5 *and where* $\bar{\beta} = (-\beta_n, -\beta_{n-1}, \cdots, -\beta_1)$ *and similarly for* $\bar{\alpha}$, $\bar{x}$ *and* $\bar{t}$, *then we have*

$$P_t(x; \alpha, \beta, N) = P_{-\bar{x}}(-\bar{t}; \bar{\beta}, \bar{\alpha}, N). \tag{6.2}$$

The proof of Proposition 6.1 is a consequence of the following

LEMMA 6.3. *With assumptions as in Proposition* 1.25 *we have*

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} = \begin{bmatrix} m'' & \bar{\nu} & \bar{\mu} \\ m' & \bar{m} & \bar{\mu}' \end{bmatrix}. \tag{6.4}$$

Proof. Applying identities (1.22) and (1.31), we obtain

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} = \sum_{(m'')\otimes(m)\otimes(m')} \langle (\nu), T_\nu(1 \otimes T_\mu)((m'')\otimes(m)\otimes(m')) \rangle$$

$$\cdot \langle (\nu), T_\nu(T_{\mu'} \otimes 1)((m'')\otimes(m)\otimes(m')) \rangle \tag{6.5}$$

with the sum over all Gelfand states $(m'')\otimes(m)\otimes(m')$ of $V_{m''} \otimes V_m \otimes V_{m'}$ and (ν) is any state in V_ν.

$$(6.6)\qquad \begin{aligned} &= \sum_{\substack{(\mu),(\mu'),(m''),\\ (m),(m')}} \langle(\nu),\, T_\nu((m'')\otimes(\mu))\rangle\langle(\mu),\, T_\mu((m)\otimes(m'))\rangle \\ &\qquad \cdot \langle(\nu),\, T_\nu((\mu')\otimes(m'))\rangle\langle(\mu'),\, T_{\mu'}((m'')\otimes(m))\rangle \end{aligned}$$

where the sum is over all states (μ') of $V_{\mu'}$, etc.

Let $m''=[p'',0,\cdots,0]$ and

$$(m'')=\begin{pmatrix}[p'',0,\cdots,0]\\ (m'')_{n-1}\end{pmatrix};$$

then in the notation of Proposition A.6 we have

$$(6.7)\qquad \langle(\nu),\, T_\nu((m'')\otimes(\mu))\rangle = \left\langle (\nu) \middle| \left\langle \begin{matrix}(\Gamma)_{n-1}\\ [p'',0,\cdots,0]\\ (m'')_{n-1}\end{matrix}\right\rangle \middle| (\mu)\right\rangle$$

where (Γ) is the unique operator pattern (see [17]) such that $[\nu]_n=\Delta(\Gamma)+[\mu]_n$. From Proposition A.6 it follows that

$$(6.8)\quad \langle(\nu),\, T_\nu((m'')\otimes(\mu))\rangle=(-1)^{p''+\varphi(\mu)+\varphi(\nu)}\left[\frac{d(\nu)}{d(\mu)}\right]^{1/2}\langle(\bar\mu),\, T_{\bar\mu}((m'')\otimes(\bar\nu))\rangle,$$

where $\varphi(\mu)$ and $\varphi(\nu)$ are defined in Definition A.3 and $d(\nu)$, $d(\mu)$ in (2.18).

From (6.6) and (6.8) we find

$$(6.9)\qquad \begin{aligned} d(\nu)\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} &= \sum_{\substack{(\nu),(\mu),(\mu')\\ (m''),(m),(m')}} \langle(\nu),\, T_\nu((m'')\otimes(\mu))\rangle \\ &\qquad \cdot\langle(\mu),\, T_\mu((m)\otimes(m'))\rangle\langle(\nu),\, T_\nu((\mu')\otimes(m'))\rangle \\ &\qquad \cdot\langle(\mu'),\, T_{\mu'}((m'')\otimes(m))\rangle \\ &= \frac{d(\nu)}{d(m)}\sum_{\substack{(\bar\nu),(\bar\mu),(\bar\mu')\\ (m''),(\bar m),(m')}} \langle(\bar\mu),\, T_{\bar\mu}((m'')\otimes(\bar\nu))\rangle \\ &\qquad \cdot\langle(\bar m),\, T_{\bar m}((m')\otimes(\bar\mu))\rangle\langle(\bar\mu'),\, T_{\bar\mu'}((m')\otimes(\bar\nu))\rangle \\ &\qquad \cdot\langle(m),\, T_{\bar m}((m'')\otimes(\bar\mu'))\rangle = d(\nu)\begin{bmatrix} m'' & \bar\nu & \bar\mu \\ m' & \bar m & \bar\mu' \end{bmatrix} \end{aligned}$$

where the sum is over all states $(\nu)\in V_\nu$, etc. This completes the proof of Lemma 6.3.

In order to apply identity (2.4) to (6.4) above, we shall need the following definition and lemma.

Definition 6.10. If $\lambda=(\lambda_1,\cdots,\lambda_n)\in\mathbb{Z}^n$ then $\lambda+k^n=(\lambda_1+k,\lambda_2+k,\cdots,\lambda_n+k)$ for any integer k. If (λ) is a Gelfand pattern, i.e. an array λ_{ij} for $1\leqq i\leqq j\leqq n$, then $(\lambda+k^n)$ is the Gelfand pattern whose (i,j) entry, $1\leqq i\leqq j\leqq n$, is $\lambda_{ij}+k$.

Lemma 6.11. *With notation as in Lemma* 6.3 *we have*

$$(6.12)\qquad \begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix}=\begin{bmatrix} m'' & m+k^n & \mu'+k^n \\ m' & \nu+k^n & \mu+k^n \end{bmatrix}$$

for any integer k.

Proof. In the notation of expression (6.7) and Definition 6.10 it follows from [6, Lemma 2.29] that

$$(6.13)\quad \left\langle (\nu) \middle| \left\langle \begin{matrix}(\Gamma)_{n-1}\\ [p'',0,\cdots,0]\\ (m'')_{n-1}\end{matrix}\right\rangle \middle| (\mu)\right\rangle = \pm\left\langle (\nu+k^n) \middle| \left\langle \begin{matrix}(\Gamma)_{n-1}\\ [p'',0,\cdots,0]\\ (m'')_{n-1}\end{matrix}\right\rangle \middle| (\mu+k^n)\right\rangle,$$

where the sign in (6.13) is always positive or always negative for all choices of Gelfand states (ν), (m'') and (μ) in V_ν, $V_{m''}$ and V_μ (assuming both sides of (6.13) do not vanish).

The phase convention (1.8) guarantees that for at least one choice of such (ν), (m'') and (μ) the sign in (6.13) is positive. Hence the sign is positive for all choices of (ν), (m'') and (μ). Now identity (6.12) follows from (6.6), (6.7) and (6.13).

Proof of Proposition 6.1. With assumptions on m'', m', m, ν, μ, μ' as in (4.1), we have by Lemmas 6.3 and 6.11 the equality

$$\begin{bmatrix} m'' & m & \mu' \\ m' & \nu & \mu \end{bmatrix} = \begin{bmatrix} m'' & \bar{\nu}+k^n & \bar{\mu}'+k^n \\ m' & \bar{m}+k^n & \bar{\mu}+k^n \end{bmatrix} \tag{6.14}$$

where $k=\nu_1$. After substituting the explicit form (2.4) and (2.14) of the Racah coefficients in (6.14) above, then by an elementary computation one obtains the identity

$$P_t(x;\alpha,\beta,N) = P_{-\bar{x}}(-\bar{t};\bar{\beta}-k^n,\bar{\alpha}-k^n,N) \tag{6.15}$$

where $\alpha_l=-m_{n-l+1}-l$, $\beta_l=-\nu_{n-l+1}-l$, $x_l=\mu'_{n-l+1}-m_{n-l+1}$ and $t_l=\nu_{n+l+1}-\mu_{n+l+1}$ for $1\leqq l\leqq n$. Observing that

$$P_{-\bar{x}}(-\bar{t};\bar{\beta}-k^n,\bar{\alpha}-k^n,N) = P_{-\bar{x}}(-\bar{t};\bar{\beta};\bar{\alpha},N), \tag{6.16}$$

then identity (6.2) follows for all $\alpha,\beta\in\mathbb{C}^n$ (such that both sides of (6.2) are defined) by an argument similar to that in the proof of Theorem 2.24.

Remark 6.17. Lemma 6.3 is a well-known symmetry relation for multiplicity free Racah coefficients (see [12] and [29]).

7. An identity for Racah polynomials. The generalized Biedenharn–Elliott identity, Proposition 1.50, implies an identity for Racah polynomials which is similar to an addition theorem. In the notation of formula (1.51) we set $\delta=(n,n-1,\cdots,1)$, $\bar{m}_1-\delta=\alpha$, $\bar{m}-\delta=\beta$, $\bar{m}_1-\bar{m}_{123}=x+y$, $\bar{m}_{12}-\bar{m}_{123}=x$, $\bar{m}_1-\bar{m}_{12}=y$, $\bar{m}_{14}-\bar{m}=t$, $\bar{m}_{14}-\bar{m}_{124}=t-u$, $\bar{m}_{124}-\bar{m}=u$, $\sum_{i=1}^n(m-m_{14})_i=N$, $\sum_{i=1}^n(m_{124}-m_{14})_i=N-N'$, $\sum_{i=1}^n(m-m_{124})_i=N'$, where for λ, $\lambda'\in\mathbb{C}^n$, then $(\lambda+\lambda')_i=\lambda_i+\lambda'_i$, $1\leqq i\leqq n$.

Applying identities (2.4), (2.13), (2.14), (2.17a) and (2.17e) to (1.51) and continuing for $\alpha,\beta\in\mathbb{C}^n$, one obtains the following:

PROPOSITION 7.1. *With notation as in Proposition* 4.5, *let* $\alpha,\beta\in\mathbb{C}^n$, $N\geqq N'\geqq 0$ *be integers and* $x,y,t\in\mathbb{Z}^n$ *such that* $\sum_{i=1}^n t_i=N$, $\sum_{i=1}^n x_i=N'$, $\sum_{i=1}^n y_i=N-N'$, *where* x_i, y_i, $t_i\geqq 0$ *for* $1\leqq i\leqq n$. *We have* (*when both sides are defined*)

$$P_t(x+y;\alpha,\beta,N) = \sum_{t_i\geqq u_i} f(y,\alpha,\beta,t,u)P_u(x;\alpha-y,\beta,N')P_{t-u}(y;\alpha,\beta+u,N-N') \tag{7.2}$$

where the sum is over all $u\in\mathbb{Z}^n$ *such that* $t_i\geqq u_i\geqq 0$ *for* $1\leqq i\leqq n$ *and* $\sum_{i=1}^n u_i=N'$ *and where*

$$\begin{aligned} f(y,\alpha,\beta,t,u) &= \binom{N}{N'}^{-1}\frac{((\beta+u)_n-\alpha_1)_{N-N'}}{((\beta+u)_n-\alpha_1)_{N-u_n}} \\ &\quad\cdot((\beta+u)_n-(\alpha-y)_1)_{N'-u_n}\prod_{l=1}^n\binom{t_l}{u_l}\prod_{2\leqq k\leqq l\leqq n-1}\frac{(\beta_l-(\alpha-y)_k)_{u_l}}{(\beta_l-\alpha_k)_{u_l}} \\ &\quad\cdot\prod_{1\leqq k<l\leqq n}\left[\frac{(\beta_k-(\alpha-y)_l)_{u_k}(\beta_k-(\beta+t)_l)_{u_k}(\beta_l-(\beta+t)_k)_{(t-u)_k}}{(\beta_k-\alpha_l)_{u_k}(\beta_k-(\beta+u)_l)_{u_k}((\beta+u)_l-(\beta+t)_k)_{(t-u)_k}}\right]. \end{aligned} \tag{7.3}$$

Appendix A. Let $m=[m_1,\cdots,m_n]$ be a $U(n)$ highest weight and define

$$\bar{m}=m^*=[-m_n,\cdots,-m_1]. \tag{A.1}$$

If (m) is a Gelfand state of V_m define $(\bar{m})$ to be the Gelfand state of $V_{\bar{m}}$ with Gelfand pattern

$$\text{(A.2)} \qquad \bar{m}_{ij} = -m_{j-i+1,j}, \qquad 1 \leqq i \leqq j \leqq n.$$

It is important to note that $(\bar{m})$ is in general *not* the dual basis vector to (m) with respect to the Gelfand-Zetlin basis of V_m.

Now let

$$(m) = \begin{pmatrix} [m]_n \\ (m)_{n-1} \end{pmatrix}, \qquad (m') = \begin{pmatrix} [m]_n \\ (m')_{n-1} \end{pmatrix}$$

be Gelfand states of V_m and assume that the highest weight m is a partition. In [17] and in [6, eq. (1.10)] the "boson polynomial"

$$B\begin{pmatrix} (m')_{n-1} \\ [m]_n \\ [m]_{n-1} \end{pmatrix}$$

is defined. Also in [6, eq. (1.14)] a "dual boson polynomial"

$$B\begin{pmatrix} (\bar{m}')_{n-1} \\ [\bar{m}]_n \\ (\bar{m})_{n-1} \end{pmatrix} = \overline{B\begin{pmatrix} (m') \\ [m]_n \\ (m)_{n-1} \end{pmatrix}}$$

is defined. This notation for the dual boson polynomial is not consistent with Louck [17] or with definition (A.2) above, because in [6] we set $(\bar{m}) \in V_{\bar{m}}$ to be the dual basis vector to the state (m) with respect to the Gelfand-Zetlin basis of V_m. However the dual basis for $V_{\bar{m}}$ is not actually the Gelfand-Zetlin basis for $V_{\bar{m}}$ as given by applying the lowering operators of Nagel and Moshinsky [23] to the highest weight state of $V_{\bar{m}}$. For a similar reason

$$\overline{B\begin{pmatrix} (m')_{n-1} \\ [m]_n \\ (m)_{n-1} \end{pmatrix}} = \mathcal{M}(m)^{1/2} \overline{\Phi\begin{pmatrix} [m']_{n-1} \\ [m]_n \\ (m)_{n-1} \end{pmatrix}} \in \mathcal{A}$$

is not a Gelfand-Zetlin state vector for the $U(n) \times U(n)$ representation acting on $V_{\bar{m}} \otimes V_{\bar{m}} \subseteq \mathcal{A}$, where

$$\overline{\Phi\begin{pmatrix} (m')_{n-1} \\ [m]_n \\ (m)_{n-1} \end{pmatrix}} \in \mathcal{A}$$

is the dual basis vector to

$$\Phi\begin{pmatrix} (m')_{n-1} \\ [m]_n \\ (m)_{n-1} \end{pmatrix} \in \mathcal{C}$$

in the pairing of the algebras $\mathcal{C}$ and $\mathcal{A}$ of creation and annihilation operators as described in [6, § 1]. We desire that

$$\Phi\begin{pmatrix} (\bar{m}')_{n-1} \\ [\bar{m}]_n \\ (\bar{m})_{n-1} \end{pmatrix}$$

should be a $U(n) \times U(n)$ Gelfand-Zetlin state vector in $V_{\bar{m}} \otimes V_{\bar{m}}$.

DEFINITION A.3. Let $n \geqq 2$. With notation as above we define

$$B\begin{pmatrix}(\bar{m}')_{n-1}\\ [\bar{m}]_n\\ (\bar{m})_{n-1}\end{pmatrix} \equiv (-1)^{\varphi(m')-\varphi(m)} B\overline{\begin{pmatrix}(m')_{n-1}\\ [m]_n\\ (m)_{n-1}\end{pmatrix}}, \tag{A.4}$$

where if (m) is a $U(n)$ Gelfand pattern then

$$\varphi(m) = \sum_{j=1}^{n} \sum_{i=1}^{j} m_{ij}.$$

LEMMA A.5. *With $n \geqq 2$ and notation as above*

$$\Phi\begin{pmatrix}(\bar{m}')_{n-1}\\ [\bar{m}]\\ (\bar{m})_{n-1}\end{pmatrix} = \mathcal{M}(m)^{-1/2} B\begin{pmatrix}(\bar{m}')_{n-1}\\ [\bar{m}]_n\\ (\bar{m})_{n-1}\end{pmatrix} \in \mathcal{A}$$

is a Gelfand-Zetlin state vector in $V_{\bar{m}} \otimes V_{\bar{m}}$.

Proof. After applying (1.10), (1.14), (2.17) and (1.17a and b) of [6] (recalling the different notation in [6]), this is essentially identity (2.112) of Louck [17]. The proof of identity (2.112) is as follows. Identity (2.112) is certainly true when $(m') = (m)$ are highest weight states of V_m. Then applying lowering operations as in (2.98) and (2.99) of [17] and using (2.114) of [17], one proves (2.112) by induction for arbitrary Gelfand states (m') and (m) of V_m.

We have the following proposition.

PROPOSITION A.6. *Let p be a nonnegative integer and*

$$(p) = \begin{pmatrix}[p, 0, \cdots, 0]_n\\ (p)_{n-1}\end{pmatrix}$$

be a Gelfand state of $V_{[p,0,\cdots,0]_n}$. Let $[m]_n$ and $[M]_n$ be $U(n)$ highest weights and assume that $[m]_n$ and $[M]_n$ are partitions. Let (m) and (M) be states in $V_{[m]_n}$ and $V_{[M]_n}$ respectively. Let $V_{[M]_n}$ occur in $V_{[p,0,\cdots,0]_n} \otimes V_{[m]_n}$ and let

$$(\Gamma) = \begin{pmatrix}(\Gamma)_{n-1}\\ [p, 0, \cdots, 0]_n\end{pmatrix}$$

be the unique operator pattern (see [17] or [6]) such that $[M]_n = \Delta(\Gamma) + [M]_n$. For $n \geqq 2$ and with notation as above, we have

$$\begin{aligned}&\left\langle (M) \middle| \left\langle \begin{matrix}(\Gamma)_{n-1}\\ [p, 0, \cdots, 0]_n\\ (p)_{n-1}\end{matrix} \right\rangle \middle| (m) \right\rangle\\ &= (-1)^{p+\varphi(M)-\varphi(m)} \left[\frac{d([M]_n)}{d([m]_n)}\right]^{1/2} \left\langle (\bar{m}) \middle| \left\langle \begin{matrix}(\Gamma_{n-1}\\ [p, 0, \cdots, 0]_n\\ (p)_{n-1}\end{matrix} \right\rangle \middle| (\bar{M}) \right\rangle\end{aligned} \tag{A.7}$$

where, for any $U(n)$ highest weight $\lambda = [\lambda_1, \cdots, \lambda_n]$,

$$d(\lambda) = \frac{\prod_{1 \leqq i < j \leqq n} (p_{in} - p_{jn})}{\prod_{i=1}^{n-1} i!} \tag{A.8}$$

and $p_{in} = \lambda_i + n - i$ for $1 \leqq i \leqq n$.

Proof. Let

$$(m') = \begin{pmatrix}[m]_n\\ (m')_{n-1}\end{pmatrix}$$

be the highest weight state in $V_{[m]_n}$ and

$$\begin{pmatrix}[p,0,\cdots,0]_n\\ (0)_{n-1}\end{pmatrix}$$

be the lowest weight state in $V_{[p,0,\cdots,0]_n}$. With notation as in [6] we have

$$(\mathcal{M}([M]_n)\mathcal{M}([m]_n))^{1/2}\left\langle\begin{pmatrix}(m')_{n-1}\\ [M]_n\\ (M)_{n-1}\end{pmatrix}\middle| B\begin{pmatrix}(0)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{pmatrix}(A)\middle|\begin{pmatrix}(m')_{n-1}\\ [m]_n\\ (m)_{n-1}\end{pmatrix}\right\rangle$$

$$=\left\langle B\begin{pmatrix}(m')_{n-1}\\ [M]_n\\ (M)_{n-1}\end{pmatrix}(A), B\begin{pmatrix}(0)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{pmatrix}(A)B\begin{pmatrix}(m')_{n-1}\\ [m]_n\\ (m)_{n-1}\end{pmatrix}(A)\right\rangle \tag{A.9}$$

$$=\left\langle \overline{B\begin{pmatrix}(0)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{pmatrix}(A)}\,\overline{B\begin{pmatrix}(m')_{n-1}\\ [m]_n\\ (m)_{n-1}\end{pmatrix}(A)}, \overline{B\begin{pmatrix}(m')_{n-1}\\ [M]_n\\ (M)_{n-1}\end{pmatrix}(A)}\right\rangle \tag{A.10}$$

$$=(-1)^{p+\varphi(M)-\varphi(m)}\left\langle \overline{B\begin{pmatrix}(0)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{pmatrix}(A)} \cdot B\begin{pmatrix}(\bar m')_{n-1}\\ [\bar m]_n\\ (\bar m)_{n-1}\end{pmatrix}(\bar A), B\begin{pmatrix}(\bar m')_{n-1}\\ [\bar M]_n\\ (\bar M)_{n-1}\end{pmatrix}(\bar A)\right\rangle \tag{A.11}$$

by Definition A.3 and Lemma A.5, and

$$=(-1)^{p+\varphi(M)-\varphi(m)}\frac{d([M]_n)\mathcal{M}([M]_n)}{d([m]_n)} \cdot\left\langle\begin{pmatrix}[\bar m]_n\\ (\bar m)_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[\bar M]_n\\ (\bar M)_{n-1}\end{pmatrix}\right\rangle \cdot\left\langle\begin{pmatrix}[\bar m]_n\\ (\bar m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\ (0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[\bar M]_n\\ (\bar m')_{n-1}\end{pmatrix}\right\rangle \tag{A.12}$$

by [6, eq. (2.24)].

From [6, eq. (2.12)] we also have

$$(\mathcal{M}([M]_n)\mathcal{M}([m]_n))^{1/2}\left\langle\begin{pmatrix}(m')_{n-1}\\ [M]_n\\ (M)_{n-1}\end{pmatrix}\middle| B\begin{pmatrix}(0)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{pmatrix}(A)\middle|\begin{pmatrix}(m')_{n-1}\\ [m]_n\\ (m)_{n-1}\end{pmatrix}\right\rangle$$

$$=\mathcal{M}([M]_n)\left\langle\begin{pmatrix}[M]_n\\ (M)_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\ (p)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[m]_n\\ (m)_{n-1}\end{pmatrix}\right\rangle \cdot\left\langle\begin{pmatrix}[M]_n\\ (m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\ (0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[m]_n\\ (m)_{n-1}\end{pmatrix}\right\rangle. \tag{A.13}$$

If we set $(p)_{n-1}=(0)_{n-1}$ and $(M)_{n-1}=(m)_{n-1}=(m')_{n-1}$ and equate (A.12) and (A.13), we obtain

$$\begin{aligned}&\frac{d([M]_n)}{d([m]_n)}\left\langle\begin{pmatrix}[\bar m]_n\\(\bar m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\(0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[\bar M]_n\\(\bar m')_{n-1}\end{pmatrix}\right\rangle^2\\ &=\left\langle\begin{pmatrix}[M]_n\\(m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\(0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[m]_n\\(m)_{n-1}\end{pmatrix}\right\rangle^2.\end{aligned}\tag{A.14}$$

From formula (2.13) above for reduced Wigner coefficients and formula (3.8) of [6] which expresses a $U(n)$ Wigner coefficient in terms of reduced Wigner coefficients, one checks that both Wigner coefficients appearing in (A.14) are postive real numbers. It follows that

$$\begin{aligned}&\left[\frac{d([M]_n)}{d([m]_n)}\right]^{1/2}\left\langle\begin{pmatrix}[\bar m]_n\\(\bar m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\(0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[\bar M]_n\\(\bar m')_{n-1}\end{pmatrix}\right\rangle\\ &=\left\langle\begin{pmatrix}[M]_n\\(m')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_{n-1}\\ [p,0,\cdots,0]_n\\(0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}[m]_n\\(m')_{n-1}\end{pmatrix}\right\rangle>0.\end{aligned}\tag{A.15}$$

After equating (A.12) and (A.13) and applying (A.15), we obtain identity (A.7). Q.E.D.

PROPOSITION A.16. *Let $n\geqq 2$. With notation as in equations* (2.10)-(2.14) *above, where h, h' are $U(n+1)$ highest weights and q, q' are $U(n)$ highest weights, we have*

$$\begin{aligned}\left\langle\begin{pmatrix}h\\q\end{pmatrix}\middle|\begin{bmatrix}[p,0,\cdots,0]_{n+1}\\ [p',0,\cdots,0]_n\end{bmatrix}\middle|\begin{pmatrix}h'\\q'\end{pmatrix}\right\rangle&=(-1)^{p'}\left[\frac{d(h)\,d(q')}{d(h')\,d(q)}\right]^{1/2}\\ &\cdot\left\langle\begin{pmatrix}\bar h'\\ \bar q'\end{pmatrix}\middle|\begin{bmatrix}[p,0,\cdots,0]_{n+1}\\ [p',0,\cdots,0]_n\end{bmatrix}\middle|\begin{pmatrix}\bar h\\ \bar q\end{pmatrix}\right\rangle\end{aligned}\tag{A.17}$$

where $d(h)$, $d(h')$ are defined by formula (A.8) (*as dimensions of $U(n+1)$ representations*) *and similarly for $d(q)$, $d(q')$* (*as dimensions of $U(n)$ representations*).

Proof. Consider the Wigner coefficient:

$$\left\langle\begin{pmatrix}h\\q\\(q'')_{n-1}\end{pmatrix}\middle|\left\langle\begin{matrix}(\Gamma)_n\\ [p,0,\cdots,0]_{n+1}\\ [p',0,\cdots,0]_n\\(0)_{n-1}\end{matrix}\right\rangle\middle|\begin{pmatrix}h'\\q'\\(q'')_{n-1}\end{pmatrix}\right\rangle\tag{A.18}$$

where

$$(\Gamma)_{n+1}=\begin{pmatrix}[p,0,\cdots,0]_{n+1}\\(\Gamma)_n\end{pmatrix}$$

is the unique operator pattern such that

$$\Delta(\Gamma)_{n+1}+h'=h\quad\text{and}\quad\begin{pmatrix}q'\\(q'')_{n-1}\end{pmatrix}$$

is a $U(n)$ highest weight state in $V_{q'}$.

Applying formula (3.8) of [6], we find

$$\left\langle\!\!\left\langle \begin{pmatrix} h \\ q \\ (q'')_{n-1} \end{pmatrix} \right| \left\langle \begin{matrix} (\Gamma)_n \\ [p,0,\cdots,0]_{n+1} \\ [p',0,\cdots,0]_n \\ (0)_{n-1} \end{matrix} \right\rangle \left| \begin{pmatrix} h' \\ q' \\ (q'')_{n-1} \end{pmatrix} \right\rangle\!\!\right\rangle$$

$$\text{(A.19)} \qquad = \left\langle \begin{pmatrix} h \\ q \end{pmatrix} \right\| \begin{bmatrix} [p,0,\cdots,0]_{n+1} \\ [p',0,\cdots,0]_n \end{bmatrix} \left\| \begin{pmatrix} h' \\ q' \end{pmatrix} \right\rangle$$

$$\cdot \left\langle \begin{pmatrix} q \\ (q'')_{n-1} \end{pmatrix} \right| \left\langle \begin{matrix} (\gamma)_{n-1} \\ [p',0,\cdots,0]_n \\ (0)_{n-1} \end{matrix} \right\rangle \left| \begin{pmatrix} q' \\ (q'')_{n-1} \end{pmatrix} \right\rangle$$

where

$$(\gamma)_n = \begin{pmatrix} [p',0,\cdots,0]_n \\ (\gamma)_{n-1} \end{pmatrix}$$

is the unique operator pattern such that $q' + \Delta(\gamma)_n = q$. Similarly we have

$$\left\langle\!\!\left\langle \begin{pmatrix} \bar{h}' \\ \bar{q}' \\ (\bar{q}'')_{n-1} \end{pmatrix} \right| \left\langle \begin{matrix} (\Gamma)_n \\ [p,0,\cdots,0]_{n+1} \\ [p',0,\cdots,0]_n \\ (0)_{n-1} \end{matrix} \right\rangle \left| \begin{pmatrix} \bar{h} \\ \bar{q} \\ (\bar{q}'')_{n-1} \end{pmatrix} \right\rangle\!\!\right\rangle$$

$$\text{(A.20)} \qquad = \left\langle \begin{pmatrix} \bar{h}' \\ \bar{q}' \end{pmatrix} \right\| \begin{bmatrix} [p,0,\cdots,0]_{n+1} \\ [p',0,\cdots,0]_n \end{bmatrix} \left\| \begin{pmatrix} \bar{h} \\ \bar{q} \end{pmatrix} \right\rangle$$

$$\cdot \left\langle \begin{pmatrix} \bar{q}' \\ (\bar{q}'')_{n-1} \end{pmatrix} \right| \left\langle \begin{matrix} (\gamma)_{n-1} \\ [p',0,\cdots,0]_n \\ (0)_{n-1} \end{matrix} \right\rangle \left| \begin{pmatrix} \bar{q} \\ (\bar{q}'')_{n-1} \end{pmatrix} \right\rangle .$$

Now applying Proposition A.6 to relate the $U(n+1)$ and $U(n)$ Wigner coefficients appearing in (A.19) and (A.20) and using that

$$\text{(A.21)} \qquad \left\langle \begin{pmatrix} q \\ (q'')_{n-1} \end{pmatrix} \right| \left\langle \begin{matrix} (\gamma)_{n-1} \\ [p',0,\cdots,0]_n \\ (0)_{n-1} \end{matrix} \right\rangle \left| \begin{pmatrix} q' \\ (q'')_{n-1} \end{pmatrix} \right\rangle \neq 0$$

as in (A.15), we obtain identity (A.17).

REFERENCES

[1] S. J. Ališaukas, A.-A. A. Jucys and A. P. Jucys, *On the symmetric tensor operators of the unitary groups*, J. Math. Phys., 13 (1972), pp. 1329–1333.

[2] S. J. Ališaukas, V. V. Vanagas and A. P. Jucys, *Relation between isoscalar factors and recoupling matrices of unitary group representations*, Dokl. Akad. Nauk SSSR, 197 (1971), pp. 804–805. (In Russian.)

[3] G. E. Andrews, *Problems and prospects for basic hypergeometric functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191–224.

[4] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[5] L. C. Biedenharn, *An identity satisfied by Racah coefficients*, J. Math. Phys., 31 (1953), pp. 287–293.

[6] L. C. Biedenharn, R. A. Gustafson and S. C. Milne, *U(n) Wigner coefficients, the path sum formula and invariant G-functions*, Advances in Appl. Math., 6 (1985), pp. 291–349.

[7] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics: Theory and Applications*, Vol. 8 in Encylopedia of Mathematics and its Applications, G.-C. Rota, ed., Addison-Wesley, Reading, MA, 1981.

[8] ———, *The Racah-Wigner Algebra in Quantum Theory*, Vol. 9, in Encylopedia of Mathematics and its Applications, G.-C. Rota, ed., Addison-Wesley, Reading, MA, 1981.

[9] E. Chacón, M. Ciftan and L. C. Biedenharn, *On the evaluation of the multiplicity-free Wigner coefficients of $U(n)$*, J. Math. Phys., 13 (1972), pp. 577-590.

[10] J. R. Derome and W. T. Sharp, *Racah algebra for an arbitrary group*, J. Math. Phys., 6 (1965), pp. 1584-1590.

[11] J. Dougall, *On Vandermonde's theorem and some more general expansions*, Proc. Edinburgh Math. Soc., 25 (1907), pp. 114-132.

[12] A. R. Edmonds, *Angular Momentum in Quantum Mechanics*, second edition, Princeton Univ. Press, Princeton, NJ, 1960.

[13] J. P. Elliott, *Theoretical studies in nuclear structure V. The matrix elements of non-central forces with an application to the 2p-shell*, Proc. Roy. Soc. A, 218 (1953), pp. 345-370.

[14] W. J. Holman iii, *Summation theorems for hypergeometric series in $U(n)$*, this Journal, 11 (1980), pp. 523-532.

[15] W. J. Holman iii, L. C. Biedenharn and J. D. Louck, *On hypergeometric series well-poised in $SU(n)$*, this Journal, 7 (1976), pp. 529-541.

[16] S. Karlin and J. McGregor, *Linear growth models with many types and multidimensional Hahn polynomials*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 261-288.

[17] J. D. Louck, *Recent progress toward a theory of tensor operators in the unitary groups*, Amer. J. Phys., 38 (1970), pp. 3-42.

[18] J. D. Louck and L. C. Biedenharn, *On the structure of the canonical tensor operators in the unitary groups* III. *Further developments of the boson polynomials and their implications*, J. Math. Phys., 14 (1973), pp. 1336-1357.

[19] ———, *Identity satisfied by the Racah coefficients of $U(n)$*, J. Math. Phys., 12 (1971), pp. 173-177.

[20] S. C. Milne, *Hypergeometric series well-poised in $SU(n)$ and a generalization of Biedenharn's G-functions*, Advances in Math., 36 (1980), pp. 161-211.

[21] ———, *An elementary proof of the Macdonald identities for $A_l^{(1)}$*, Advances in Math., 57 (1985), pp. 34-70.

[22] ———, *Private communication.*

[23] J. G. Nagel and M. Moshinsky, *Operators that lower or raise irreducible vector spaces of U_{n-1} contained in an irreducible vector space of U_n*, J. Math. Phys., 6 (1965), pp. 682-694.

[24] G. N. Watson, *A new proof of the Rogers-Ramanujan identities*, J. London Math. Soc. 4 (1929), pp. 4-9.

[25] F. J. W. Whipple, *On well-poised series, generalized hypergeometric series having parameters in pairs, each pair with the same sum*, Proc. London Math. Soc. (2), 24 (1926), pp. 247-263.

[26] J. A. Wilson, *Hypergeometric series recurrence relations and some new orthogonal functions*, Ph.D. thesis, Univ. Wisconsin, Madison, 1978.

[27] ———, *Some hypergeometric orthogonal polynomials*, this Journal, 11 (1980), pp. 690-701.

[28] M. K. F. Wong, *On the structure of the multiplicity-free Wigner coefficients of $U(n)$*, J. Math. Phys., 17 (1976), pp. 1558-1569.

[29] ———, *Multiplicity-free 6-j symbols and Weyl coefficients of $U(n)$: Explicit evaluation*, J. Math. Phys., 19 (1978), pp. 1635-1643.

[30] S. K. Suslov, *The 9j-symbols as orthogonal polynomials in two discrete variables*, Sov. J. Nucl. Phys., 38(4), October 1983, pp. 662-663 (English translation); pp. 1102-1104 (Russian).

[31] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, this Journal, 10 (1979), pp. 1008-1016.

[32] ———, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem Amer. Math. Soc., 54, Number 319, 1985.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

020

QUADRATURE SUMS INVOLVING *p*TH POWERS OF POLYNOMIALS*

D. S. LUBINSKY†, A. MATɇ AND P. NEVAI§

Abstract. R. Askey's problem of estimating quadrature sums involving pth powers of polynomials $(0<p<\infty)$ in terms of integrals, is solved. In its simplest form, the method can extend the large sieve of number theory to sums involving pth powers, rather than just squares, of trigonometric polynomials. Further, the method yields estimates whenever the abscissas in the quadrature have a suitable spacing and the weights have suitable bounds. In particular, it may be applied to generalized Jacobi weights and to Freud weights.

Key words. quadrature sums, pth powers, orthogonal polynomials, Freud weights, the large sieve, generalized Jacobi weights

AMS(MOS) subject classifications. Primary 41A55; secondary 42C05

1. Introduction. The large sieve of number theory is an inequality for trigonometric polynomials of degree at most n,

$$S_n(\tau)=\sum_{k=-n}^{n} a_k e^{ik\tau}, \qquad \tau\in[0,2\pi],$$

which states that

$$\sum_{j=1}^{m}|S_n(\tau_j)|^2\leqq\{2n+\delta^{-1}\}(2\pi)^{-1}\int_0^{2\pi}|S_n(\tau)|^2\,d\tau, \tag{1.1}$$

whenever $0\leqq\tau_1<\tau_2<\cdots<\tau_M\leqq 2\pi$ and

$$\delta=\min\{\tau_2-\tau_1,\tau_3-\tau_2,\cdots,\tau_m-\tau_{m-1},2\pi-(\tau_m-\tau_1)\}>0. \tag{1.2}$$

See Montgomery [16, p. 548 and p. 559, Thm. 3], but note the different notation.

It is the purpose of this paper to extend (1.1) to sums of the form $\sum_{j=1}^{n}|S_n(\tau_j)|^p$, $p>0$, and to show how such sums may be estimated using L_2 techniques. While such sums have been considered by Davenport and Halberstam [4] and Forti and Viola [5], their estimates did not include the case $0<p<1$, and the upper bounds involved terms such as $\sum_{k=0}^{n-1}|a_k|^p(k+2)^{p-2}$, rather than $\int_0^{2\pi}|S_n(\tau)|^p\,d\tau$. In this paper, we shall also solve a problem of Askey [2] on quadrature sums involving pth powers of polynomials. Before discussing the latter, we first illustrate the method in the case of trigonometric polynomials. Let

$$D_n(\tau)=\sum_{k=-n}^{n} e^{ik\tau}, \qquad \tau\in[0,2\pi). \tag{1.3}$$

* Received by the editors March 4, 1985, and in revised form October 16, 1985.

† National Research Institute for Mathematical Sciences, CSIR, PO Box 395, Pretoria 0001, Republic of South Africa. This author began working on this material during a visit to Ohio State University, Columbus, Ohio 43210.

‡ Department of Mathematics, Brooklyn College of the City University of New York, Brooklyn, New York 11210. The work of this author was supported in part by National Science Foundation grant MCS-81-00673 and by the PSC-CUNY Research Award Program of New York under grant 662043.

§ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The work of this author was supported in part by National Science Foundation grant MCS-83-00882.

Applying (1.1) to $D_n(x-\tau)$, and noting that

$$\int_0^{2\pi} |D_n(\tau)|^2\,d\tau = 2\pi(2n+1), \tag{1.4}$$

we obtain for any $x \in [0, 2\pi)$,

$$\sum_{j=1}^{m} |D_n(x-\tau_j)|^2 \leqq \{2n+\delta^{-1}\}(2n+1). \tag{1.5}$$

We remark that it is unnecessary to use the large sieve to derive an inequality of the form (1.5), and we pause in our exposition to outline the proof of such an inequality. First, note the straightforward bound

$$|D_n(\tau)| \leqq \min\{2n+1, \pi/|\tau|\}, \qquad |\tau| \leqq \pi.$$

Let $\varepsilon = (2n+1)^{-1}+\delta$ and assume (as we may) that all $x-\tau_j$ $(j=1,2,\cdots,m)$ lie in $[-\pi,\pi]$. Then

$$\begin{aligned}\sum_{j=1}^{m} |D_n(x-\tau_j)|^2 &\leqq \sum_{|x-\tau_j|\geqq\varepsilon} (\pi/|x-\tau_j|)^2 + \sum_{|x-\tau_j|<\varepsilon} (2n+1)^2\\ &\leqq 2(2n+1)^2 + 2\sum_{j\delta\geqq\varepsilon} (\pi/(j\delta))^2 + 2\sum_{j\delta<\varepsilon} (2n+1)^2.\end{aligned}$$

Here we have used the fact that in each interval of length δ, there is at most one point of the form $x-\tau_j$. Then, using the definition of ε,

$$\begin{aligned}\sum_{j=1}^{m} |D_n(x-\tau_j)|^2 &\leqq 2(2n+1)^2 + 2(\pi/\delta)^2 \sum_{j\geqq((2n+1)\delta)^{-1}+1} j^{-2} + 2(2n+1)^2\{\varepsilon/\delta+1\}\\ &\leqq 2(\pi/\delta)^2((2n+1)\delta) + 2(2n+1)^2\{2+((2n+1)\delta)^{-1}\}\\ &\leqq 22(2n+1)\{n+\delta^{-1}\}.\end{aligned}$$

This establishes (1.5) except for a factor of 22.

We now return to our main theme. Let $p>0$. Theorem 6 in Maté and Nevai [14] shows that for any trigonometric polynomial $S_n(\tau)$ as above,

$$|S_n(\tau)|^p \leqq (2+2np)e(8\pi)^{-1} \int_0^{2\pi} |S_n(u)|^p\,du, \qquad \tau \in [0, 2\pi]. \tag{1.6}$$

Let k denote the smallest positive integer such that $kp \geqq 2$. Applying (1.6) to $S_n(\tau)D_n^k(x-\tau)$ with x fixed, we obtain

$$|S_n(\tau)D_n^k(x-\tau)|^p \leqq (2+2n(k+1)p)e(8\pi)^{-1} \int_0^{2\pi} |S_n(u)D_n^k(x-u)|^p\,du.$$

Now setting $\tau = x$, and noting that $kp = \{kp-2\}+2$, and

$$|D_n(x-u)| \leqq 2n+1 = D_n(0),$$

we obtain

$$|S_n(x)|^p \leqq (2+2n(k+1)p)e(8\pi)^{-1}(2n+1)^{-2} \int_0^{2\pi} |S_n(u)|^p D_n^2(x-u)\,du. \tag{1.7}$$

Next, $(k-1)p<2$, so that

$$2+2n(k+1)p < 2+2n(2p+2) < (2n+1)2(p+1)$$

and hence

$$|S_n(x)|^p \leqq (p+1)e(4\pi)^{-1}(2n+1)^{-1} \int_0^{2\pi} |S_n(u)|^p D_n^2(x-u)\,du. \tag{1.8}$$

Applying Jensen's inequality (Zygmund [24, p. 24]) to (1.8), and noting (1.4), we obtain the following proposition.

THEOREM 1. *Let* $0<p<\infty$. *Let* $\psi(t)$ *be convex, nonnegative and nondecreasing in* $[0,\infty)$. *For any trigonometric polynomial* $S_n(t)$ *of degree at most n,*

$$\psi(|S_n(x)|^p)\leqq(2\pi)^{-1}(2n+1)^{-1}\int_0^{2\pi}\psi(|S_n(u)|^p(p+1)e/2)D_n^2(x-u)\,du, \tag{1.9}$$
$$x\in[0,2\pi].$$

If $0<p<1$, t^p is concave, rather than convex, and so it is meaningful to include the pth power in (1.9). Applying (1.5) to (1.9), we obtain

THEOREM 2. *Assuming the notation of* (1.2) *and Theorem* 1,

$$\sum_{j=1}^{m}\psi(|S_n(\tau_j)|^p)\leqq(2n+\delta^{-1})(2\pi)^{-1}\int_0^{2\pi}\psi(|S_n(u)|^p(p+1)e/2)\,du. \tag{1.10}$$

Theorems 1 and 2 are useful in trigonometric interpolation and approximation.

In 1969, Askey [2, p. 553] posed the following problem: Let $P_{n-1}(x)$ be any polynomial of degree at most $n-1$. Let $\{x_{nk}\}_{k=1}^{n}$ be the zeros of the polynomials orthogonal with respect to $d\alpha(x)$, a positive measure on $[-1,1]$, and let $\{\lambda_{nk}\}_{k=1}^{n}$ be the corresponding Christoffel numbers. When is

$$\sum_{k=1}^{n}\lambda_{nk}|P_{n-1}(x_{nk})|^p\leqq C\int_{-1}^{1}|P_{n-1}(x)|^p\,d\alpha(x), \tag{1.11}$$

where C is independent of P_{n-1} and n? Such inequalities are essential in various problems in approximation theory, and in particular, in investigating mean convergence of Lagrange interpolation (Askey [1], Bonan [3], Knopfmacher and Lubinsky [9], Nevai [18], [19], [20], [22], [23]).

For $p=2$, the Gauss quadrature formula establishes (1.11) with $C=1$ and with equality of both sides. Askey proved (1.11) for certain Jacobi weights for $p\geqq1$. Subsequently, Nevai [20, pp. 167-168] proved inequalities generalizing (1.11) for $p\geqq1$ and for generalized Jacobi weights on $[-1,1]$. For the Hermite weight, estimates such as (1.11) appear in Nevai [20] with the range of summation suitably restricted. Similar estimates for generalized Hermite weights appear in Bonan [3] and for Freud weights in Knopfmacher and Lubinsky [9].

To date, no one has succeeded in dealing with the case $0<p<1$. We shall fill this gap, using much the same method as that used to prove Theorems 1 and 2. Furthermore, we can treat general quadrature sums with suitably spaced abscissas, rather than just Gauss quadrature sums. As before, estimation of sums involving pth powers is reduced to estimation of a sum involving an even integer power of a specific polynomial. The specific polynomial (which is the analogue of $D_n(x-t)$ above) is

$$K_n(v,x,t)=\sum_{j=0}^{n-1}p_j(v,x)p_j(v,t), \tag{1.12}$$

the kernel function associated with the Chebyshev weight

$$v(t)=\begin{cases}(1-t^2)^{-1/2}, & t\in(-1,1),\\ 0, & t\notin(-1,1).\end{cases} \tag{1.13}$$

Here $p_j(v,x)$, $j=0,1,2,\cdots$, are the orthonormal Chebyshev polynomials associated with v. Finally, the sum involving $K_n(v,x,t)$ may be estimated using the large sieve or using methods of Nevai [20, p. 167].

Before stating the results for weights on $[-1,1]$, we need the definition of the generalized Jacobi weights (GJ), studied in [20], [21], [23]. Throughout, $C, C_1, C_2, \cdots$, denote positive constants independent of n and x, and of all polynomials P of degree at most n, or at most a positive constant times n. Further, we use, o, O as in [20]. Thus, for example, $f(x) \sim g(x)$, if there exist C_1 and C_2 such that $C_1 \leqq f(x)/g(x) \leqq C_2$ for all relevant x.

DEFINITION 3. Let $-1 = t_1 > t_2 > \cdots > t_N = 1$ and $\Gamma_k > -1$, $k = 1, 2, \cdots, N$. Let $w(t) = 0$, $t \notin [-1, 1]$ and

$$w(t) \sim \prod_{k=1}^{N} |t - t_k|^{\Gamma_k}, \qquad t \notin [-1, 1]. \tag{1.14}$$

Then we say that w is a generalized Jacobi weight and write $w \in \mathrm{GJ}$. The following quantities are associated with w:

$$\bar{w}_n(t) = (\sqrt{1-t} + 1/n)^{2\Gamma_1 + 1} \left\{ \prod_{k=2}^{N-1} (|t - t_k| + 1/n)^{\Gamma_k} \right\} (\sqrt{1+t} + 1/n),^{2\Gamma_n + 1}, \tag{1.15}$$
$$t \in [-1, 1], \qquad n = 1, 2, \cdots,$$

and $\bar{w}_n(t) = 0$, $t \notin [-1, 1]$. Further for $0 < p < \infty$ and $n = 1, 2, 3, \cdots$,

$$\lambda_n(w, p, x) = \inf \int_{-1}^{1} \frac{|P(u)|^p w(u)\, du}{|P(x)|^p}, \qquad x \in [-1, 1], \tag{1.16}$$

where the inf is over all polynomials P of degree at most $n-1$. In particular,

$$\lambda_n(w, x) = \lambda_n(w, 2, x), \qquad x \in [-1, 1].$$

THEOREM 4. *Let $w \in GJ$, $0 < p$, $L < \infty$ and l be a positive integer. Further, let $\psi(t)$ be convex, nonnegative and nondecreasing in $[0, \infty)$. Then for all polynomials $P(x)$ of degree at most ln,*

$$\psi(|P(x)|^p) \bar{w}_n(x) \leqq C_1 n^{-L+1} \int_{-1}^{1} \psi(C_2 |P(u)|^p) |K_n(v, x, u)|^L w(u)\, du, \tag{1.17}$$
$$x \in [-1, 1].$$

THEOREM 5. *Let $w \in GJ$, $0 < p < \infty$ and let l be a positive integer. Further, let $\psi(t)$ be convex, nonnegative and nondecreasing in $[0, \infty)$. Given*

$$-1 \leqq y_m < y_{m-1} < \cdots < y_1 \leqq 1,$$

write

$$\theta_j = \arccos(y_j) \in [0, \pi], \qquad j = 1, 2, \cdots, m$$

and let

$$\delta = \min\{\theta_2 - \theta_1, \theta_3 - \theta_2, \cdots, \theta_m - \theta_{m-1}\} > 0. \tag{1.18}$$

Then for all polynomials $P(x)$ of degree at most ln,

$$\sum_{j=1}^{m} \bar{w}_n(y_j) \psi(|P(y_j)|^p) \leqq C_1 \{n + \delta^{-1}\} \int_{-1}^{1} \psi(C_2 |P(u)|^p) w(u)\, du, \tag{1.19}$$

and

$$\sum_{j=1}^{m} \lambda_n(w, y_j) \psi(|P(y_j)|^p) \leqq C_3 \{1 + (n\delta)^{-1}\} \int_{-1}^{1} \psi(C_2 |P(u)|^p) w(u)\, du. \tag{1.20}$$

The constants C_1, C_2 and C_3 are independent of m, $y_1, y_2, \cdots, y_m$, δ, n and P.

COROLLARY 6. *Let $\{x_{nj}\}$ and $\{\lambda_{nj}\}$ denote the abscissas and Christoffel numbers in the Gauss quadrature for the weight $w \in GJ$. With the notation of the previous theorem,*

$$\sum_{j=1}^{n} \lambda_{nj}\psi(|P(x_{nj})|^p) \leqq C_1 \int_{-1}^{1} \psi(C_2|P(u)|^p)w(u)\,du, \tag{1.21}$$

for all polynomials P of degree at most ln.

In the case $\psi(t)=t$ and $p \geqq 1$, Corollary 6 is implied by Nevai [20, Thm. 9.25, p. 168]. Further for $\psi(t)=t$, Corollary 6 ensures that Nevai [23, (24), p. 675] is correct as stated, even for $0<p<1$.

Finally, we state the analogues of the above results for weights with the whole real line as support. First, however, we need a definition of $\mathscr{F}$, a suitable class of Freud weights:

DEFINITION 7. Let $W(x)=\exp(-Q(x))$, $x \in \mathbb{R}$, where Q is even and continuous in $\mathbb{R}$, and $Q''(x)$ is continuous and nonnegative for large positive x. Then we say W is a Freud weight and write $W \in \mathscr{F}$ if either

(1.22) I. $Q''(x)$ is positive and nondecreasing for large positive x, or

II. There exists $\alpha \in (1,2)$ and C and C_1 such that

$$xQ''(x)/Q'(x) \leqq C, \qquad x \in (C_1, \infty), \tag{1.23}$$

and

$$\lim_{x\to\infty} xQ'(x)/Q(x) = \alpha. \tag{1.24}$$

Associated with W are the following quantities: For large enough n, q_n denotes the positive root of the equation

$$q_nQ'(q_n) = n. \tag{1.25}$$

Further, given a nonnegative integer j, $0<p\leqq\infty$ and $n=j+1, j+2, \cdots$, we define the generalized Christoffel functions

$$\lambda_{n,p}(W,j,x) = \inf \|PW\|_{L_p(\mathbb{R})}/|P^{(j)}(x)|, \qquad x \in \mathbb{R}, \tag{1.26}$$

where the inf is over all polynomials P of degree at most n. We also define the classical Christoffel functions

$$\lambda_n(W^2,x) = \lambda_{n,2}^2(W,0,x) = \inf \int_{-\infty}^{\infty} \frac{(PW(u))^2\,du}{P^2(x)}, \qquad n=1,2,\cdots, \quad x \in \mathbb{R}. \tag{1.27}$$

The essential feature of $W \in \mathscr{F}$ is the following: Lower bounds are available for the generalized Christoffel functions (Freud [6], [8], Lubinsky [13], Levin and Lubinsky [10], [11] and inequalities are available which connect the L_p norms of weighted polynomials over finite and infinite ranges (Nevai [17], Mhaskar and Saff [15], Lubinsky [12], [13]). As concrete examples of weights $W \in \mathscr{F}$ satisfying I, we mention $\exp(-|x|^\alpha(\log|x|)^\beta)$, $\alpha \geqq 2$, $\beta \geqq 0$, and $\exp(-\exp(|x|^\alpha))$, $\alpha>0$, suitably modified for small $|x|$. As examples of weights $W \in \mathscr{F}$ satisfying II, we mention $\exp(-|x|^\alpha(\log|x|)^\beta)$, $1<\alpha<2$, $\beta \in \mathbb{R}$, suitably modified for small $|x|$.

THEOREM 8. *Let $W \in \mathscr{F}$, $0<p<\infty$, and let l be a positive integer. Let ψ be convex, nonnegative and nondecreasing in $[0,\infty)$. There exists a positive constant C^* with the following property: Let*

$$-C^*q_n \leqq y_m < y_{m-1} < \cdots < y_1 \leqq C^*q_n,$$

and

$$\delta = \min\{y_j - y_{j-1}: j=2,3,\cdots,m\} > 0. \tag{1.28}$$

Then for all polynomials P of degree at most ln,

$$(1.29)\qquad \sum_{j=1}^{m} \psi(|PW|^p(y_j)) \leqq C_1\{n/q_n+\delta^{-1}\} \int_{-\infty}^{\infty} \psi(C_2|PW|^p(u))\,du.$$

The constants C_1 and C_2 are independent of m, δ, $\{y_j\}$, n and P. If in addition (1.23) *holds, then we may take C^* arbitrarily large, but C_1 and C_2 will depend on C^*.*

We shall prove Theorem 8 using the kernel function for the Chebyshev weight. By using the kernel function for the weight W^2, we can also estimate sums of the form

$$\sum_{j=1}^{m} \psi(|P(y_j)|^p)\,W^2(y_j),$$

but we omit the more complicated proof.

For the Hermite weight, the following corollary is related to results in Nevai [22] and for generalized Hermite weights to results in Bonan [3]. For Freud weights, a similar result appears in Knopfmacher and Lubinsky [9], but for all cases, the following corollary is new when $0<p<1$.

COROLLARY 9. *Let $\{x_{nj}\}$ and $\{\lambda_{nj}\}$ denote the abscissas and Christoffel numbers in the Gauss quadrature for the weight W^2, where $W\in\mathcal{F}$. Let $-\infty<r<2$. With the notation of the previous theorem, there exists C^* such that*

$$(1.30)\qquad \sum_{|x_{nj}|\leqq C^*q_n} \lambda_{nj}|P(x_{nj})|^p W^{-r}(x_{nj}) \leqq C_1 \int_{-\infty}^{\infty} |P(u)|^p W^{2-r}(u)\,du,$$

for all polynomials P of degree at most ln.

Note that even when (1.23) holds, we could not prove (1.30) for the full sum extended over $j=1,2,\cdots,n$. The reason for this is that upper bounds for the Christoffel functions are not available near x_{n1} and x_{nn}.

The paper is organized as follows: In § 2, we prove the results for the finite interval, namely Theorems 4 and 5 and Corollary 6, and in § 3, we prove Theorem 8 and Corollary 9 for weights on the infinite interval.

2. Weights on [−1, 1]. The proof of Theorems 4 and 5 and Corollary 6 will use estimates on the L_p Christoffel functions:

LEMMA 2.1. *Let $w\in GJ$ and $0<p<\infty$. Then for $n=1,2,3,\cdots$,*

$$(2.1)\qquad \lambda_n(w,p,x)\sim\lambda_n(w,x)\sim n^{-1}\bar{w}_n(x),\qquad x\in[-1,1].$$

Proof. See Nevai [20, Thm. 6.3.28, p. 120], and note that if $w\in GJ$ and $w_1\in GJ$ and $w\sim w_1$, in $[-1,1]$, then $\lambda_n(w,p,x)\sim\lambda_n(w,x)$ in $[-1,1]$. □

We can now prove Theorem 4 in a special case:

Proof of Theorem 4 *when $\psi(t)=t$ and L is a positive even integer.* Let L be a fixed positive even integer. By definition of the Christoffel functions, we have for all polynomials P of degree at most ln, and for $|t|\leqq 1$,

$$(2.2)\qquad \begin{aligned} |P(t)|^p &\leqq \lambda_{ln}^{-1}(w,p,t)\int_{-1}^{1}|P(u)|^p w(u)\,du\\ &\leqq Cn\bar{w}_n^{-1}(t)\int_{-1}^{1}|P(u)|^p w(u)\,du, \end{aligned}$$

by Lemma 2.1, and as $\bar{w}_{ln}(x)\sim\bar{w}_n(x)$, $|x|\leqq 1$. Let k be the smallest integer such that

$kp \geqq L$. Applying (2.2) to the polynomial $P(t)K_n^k(v, x, t)$ with x fixed, we obtain

$$(2.3) \qquad |P(t)|^p |K_n(v, x, t)|^{kp} \leqq Cn\bar{w}_n^{-1}(t) \int_{-1}^{1} |P(u)|^p |K_n(v, x, u)|^{kp} w(u)\, du, \qquad x, t \in [-1, 1].$$

Next, we note that

$$(2.4) \qquad |K_n(v, x, t)| \leqq C_1 n, \qquad |x|, |t| \leqq 1,$$

and

$$(2.5) \qquad K_n(v, x, x) \sim n, \qquad |x| \leqq 1$$

(Nevai [20, p. 79, p. 108]). Setting $t = x$ in (2.3), and using (2.4) and (2.5) and $kp = L + \{kp - L\}$ with $kp - L \geqq 0$, we obtain for $|x| \leqq 1$,

$$(2.6) \qquad |P(x)|^p \leqq Cn^{-L+1}\bar{w}_n^{-1}(x) \int_{-1}^{1} |P(u)|^p K_n^L(v, x, u) w(u)\, du,$$

which is Theorem 4 in the special case mentioned. □

In order to complete the proof of Theorem 4, we must apply Jensen's inequality to (2.6). For this to be possible, we must obtain upper bounds for $\int_{-1}^{1} K_n^L(v, x, u) w(u)\, du$:

THEOREM 2.2. *Let $w \in GJ$. Let L be a positive even integer such that*

$$(2.7) \qquad L > 1 + \max\{2\Gamma_1 + 1, \Gamma_2, \Gamma_3, \cdots, \Gamma_{N-1}, 2\Gamma_N + 1\}.$$

Then with the notation of (1.12) *to* (1.16),

$$(2.8) \qquad \int_{-1}^{1} \{K_n(v, x, u)/K_n(v, x, x)\}^L w(u)\, du \sim n^{-1}\bar{w}_n(x), \qquad x \in [-1, 1].$$

Proof. In the special case $N = 2$, so that w is a Jacobi weight, this is Theorem 6.3.10 in Nevai [20, p. 109]. Let

$$I_n(x) = \int_{-1}^{1} \{K_n(v, x, u)/K_n(v, x, x)\}^L w(u)\, du, \qquad |x| \leqq 1, \quad n = 1, 2, \cdots.$$

By Lemma 2.1,

$$I_n(x) \geqq Cn^{-1}\bar{w}_n(x), \qquad |x| \leqq 1.$$

Thus we must prove

$$(2.9) \qquad I_n(x) \leqq C_1 n^{-1}\bar{w}_n(x), \qquad |x| \leqq 1.$$

First, note from (1.15) and (2.7) that

$$(2.10) \qquad \bar{w}_n(t) \geqq C_2(1/n)^{L-1}, \qquad |t| \leqq 1.$$

Let $\varepsilon = (1/4)\min_{i \neq j} |t_i - t_j|$, so that for $|x| \leqq 1$, the interval $(x - \varepsilon, x + \varepsilon)$ contains at most one of $\{t_1, t_2, \cdots, t_N\}$. Given $|x| \leqq 1$, we let t_c denote the element of $\{t_1, t_2, \cdots, t_N\}$ that is closest to x. If this does not uniquely define t_c, we take t_c to be the closest element from the left of x. We have

$$(2.11) \qquad w(u) \sim |u - t_c|^{\Gamma_c}, \qquad u \in (x - \varepsilon, x + \varepsilon) \cap [-1, 1],$$

with the constants in $\sim$ independent of $x\in[-1,1]$. Next, it follows from Lemma 6.3.8 in Nevai [20, p. 108] that

$$|K_n(v,x,u)|\leqq C/\{|x-u|+1/n\},\qquad |x|,|u|\leqq 1. \tag{2.12}$$

By (2.5), (2.11) and (2.12),

$$I_n(x)\leqq n^{-L}\int_{|x-u|\geqq\varepsilon}(C/\varepsilon)^L w(u)\,du \tag{2.13}$$

$$+C_3 n^{-L}\int_{|x-u|<\varepsilon}|u-t_c|^{\Gamma_c}(|x-u|+1/n)^{-L}\,du$$

$$\leqq C_4 n^{-1}\bar{w}_n(x)+C_3 n^{-L}J_n(x), \tag{2.14}$$

by (2.10) and where $L_n(x)$ denotes the second integral in (2.13), so that

$$J_n(x)=\int_{-\varepsilon}^{\varepsilon}|(x-t_c)-t|^{\Gamma_c}/(|t|+1/n)^L\,dt. \tag{2.15}$$

Suppose first that $2\leqq c\leqq N-1$, so that $t_c\in(-1,1)$. Let $z=x-t_c$. We consider two cases:

Case I. $|z|\geqq 1/n$. Let $\sigma=\operatorname{sign}(z)$. We see from (2.15) that

$$J_n(x)\leqq\left\{\int_{|t|\geqq|z|/2}+\int_{|t|\leqq|z|/2}\right\}|z-t|^{\Gamma_c}/(|t|+1/n)^L\,dt$$

$$\leqq|z|^{\Gamma_c+1}\int_{|u|\geqq 1/2}|\sigma-u|^{\Gamma_c}/(|uz|+1/n)^L\,du+C|z|^{\Gamma_c}\int_{|t|\leqq|z|/2}(|t|+1/n)^{-L}\,dt,$$

by the substitution $t=|z|u$ in the first integral, and as $|z-t|\sim|z|$ for t in the second integral. Then we see that

$$J_n(x)\leqq|z|^{\Gamma_c+1-L}\int_{|u|\geqq 1/2}|\sigma-u|^{\Gamma_c}|u|^{-L}\,du+|z|^{\Gamma_c}2C\int_{1/n}^{2}u^{-L}\,du$$

$$\leqq C|z|^{\Gamma_c}n^{L-1}, \tag{2.16}$$

as $L>1+\Gamma_c$ and $|z|\geqq 1/n$. Finally, we note that as $|z|=|x-t_c|\geqq 1/n$,

$$\bar{w}_n(x)\sim(|x-t_c|+1/n)^{\Gamma_c}\sim|z|^{\Gamma_c}$$

and hence (2.14) and (2.16) yield (2.9).

Case II. $|z|<1/n$. Making the substitution $t=u/n$ in (2.15), we see that

$$J_n(x)=n^{L-1-\Gamma_c}\int_{-n\varepsilon}^{n\varepsilon}|nz-u|^{\Gamma_c}/(|u|+1)^L\,du$$

$$\leqq n^{L-1-\Gamma_c}\max\left\{\int_{-\infty}^{\infty}|s-u|^{\Gamma_c}/(|u|+1)^L\,du\colon |s|\leqq 1\right\}$$

$$\leqq Cn^{L-1}\bar{w}_n(x),$$

as $|z|=|x-t_c|<1/n$ implies that

$$\bar{w}_n(x)\sim(|x-t_c|+1/n)^{\Gamma_c}\sim n^{-\Gamma_c}.$$

Then (2.14) again yields (2.9).

Finally, we must show that (2.9) holds when $t_c\in\{t_1,t_N\}$. Proceeding as at (2.13) and (2.14), we see

$$I_n(x)\leqq Cn^{-1}\bar{w}_n(x)+C\int_{|x-u|<\varepsilon}\{K_n(v,x,u)/K_n(v,x,x)\}^L|u-t_c|^{\Gamma_c}\,du.$$

$$\leqq Cn^{-1}\bar{w}_n(x)+Cn^{-1}(|x-t_c|^{1/2}+1/n)^{2\Gamma_c+1}$$

by Theorem 6.3.10 in Nevai [20, p. 109] as $t_c = \pm 1$, and (2.9) again follows. □

We can now complete the proof of Theorem 4:

Proof of Theorem 4 *in the general case.* Let L be a positive even integer satisfying (2.7). By (2.5) and (2.8), we may rewrite (2.6) as

$$|P(x)|^p \leqq C \int_{-1}^{1} |P(u)|^p K_n^L(v, x, u) w(u)\, du \Big/ \int_{-1}^{1} K_n^L(v, x, u) w(u)\, du,$$

for all polynomials P of degree at most ln. Applying Jensen's inequality (Zygmund [24, p. 24]), and then using (2.8), we obtain (1.17) for all sufficiently large positive even integers L. It then follows that (1.17) holds for all $L > 0$, since if $0 < L' < L$,

$$|K_n(v, x, u)|^L \leqq C n^{L-L'} |K_n(v, x, u)|^{L'}, \qquad |x|, |u| \leqq 1. \qquad \square$$

We shall need the full generality of the following lemma in the next section.

LEMMA 2.3. *Let* $-\infty < y_1 < y_2 < \cdots < y_m < \infty$ *and* $Y \geqq \max_j |y_j|$. *Let*

$$\theta_j = \arccos(y_j/Y) \in [0, \pi], \qquad j = 1, 2, \cdots, m, \tag{2.17}$$

and let

$$\delta = \min\{\theta_{j+1} - \theta_j : 1 \leqq j \leqq m-1\}. \tag{2.18}$$

Then for $|u| \leqq Y$,

$$\sum_{j=1}^{m} K_n^2(v, y_j/Y, u/Y) \leqq (8/\pi^2) n\{n + \delta^{-1}\}, \tag{2.19}$$

where $K_n(v, x, t)$ *and* v *are as in* (1.12) *and* (1.13).

Proof. Note that

$$p_0(v, x) = \pi^{-1/2} \quad \text{and} \quad p_j(v, x) = (2/\pi)^{1/2} \cos(j \arccos x),$$

$j = 1, 2, 3, \cdots, x \in [-1, 1]$. Let $\eta, \theta \in [0, \pi]$. If $'$ denotes that the first term in the sum is multiplied by $\frac{1}{2}$,

$$\begin{aligned} K_n(v, \cos\theta, \cos\eta) &= \frac{2}{\pi} \sum_{j=0}^{n-1}{}' \cos(j\theta) \cos(j\eta) \\ &= \frac{1}{\pi} \sum_{j=0}^{n-1}{}' \{\cos j(\eta + \theta) + (-1)^j \cos j(\eta - \theta + \pi)\} \\ &= \operatorname{Re}\{S_n(\eta + \theta) + S_n^*(\eta - \theta + \pi)\}, \end{aligned} \tag{2.20}$$

where

$$S_n(t) = \frac{1}{\pi} \sum_{j=0}^{n-1}{}' e^{ijt} \quad \text{and} \quad S_n^*(t) = \frac{1}{\pi} \sum_{j=0}^{n-1}{}' (-1)^j e^{ijt}.$$

Now let $|u| \leqq Y$ and write $\eta = \arccos(u/Y) \in [0, \pi]$. We see from (2.17) and (2.20) that

$$\sum_{j=1}^{m} K_n^2(v, y_j/Y, u/Y) \leqq 2 \sum_{j=1}^{m} \{|S_n(\eta + \theta_j)|^2 + |S_n^*(\eta - \theta_j + \pi)|^2\}.$$

Note that $\theta_j + \eta \in [0, 2\pi]$ and $\eta - \theta_j + \pi \in [0, 2\pi]$, $j = 1, 2, \cdots, m$, and further that the numbers $\theta_j + \eta$, $j = 1, 2, \cdots, m$ lie in an interval of length at most π, as do the numbers $\eta - \theta_j + \pi$, $j = 1, 2, \cdots, m$. We may thus apply the large sieve, namely (1.1), to S_n and S_n^* to deduce

$$\begin{aligned} \sum_{j=1}^{m} K_n^2(v, y_j/Y, u/Y) &\leqq 2\{2n + \delta^{-1}\} \frac{1}{2\pi} \int_0^{2\pi} \{|S_n(t)|^2 + |S_n^*(t)|^2\}\, dt \\ &\leqq 4n\pi^{-2}(2n + \delta^{-1}). \end{aligned} \qquad \square$$

An alternative proof of Lemma 2.3 may be based on the method in [20, pp. 167-168].

Proof of Theorem 5. First, (1.19) follows easily from Theorem 4 and Lemma 2.3 with $Y=1$, while (1.20) follows from (1.19) and Lemma 2.1.

Proof of Corollary 6. Let $\theta_{nj}=\arccos(x_{nj})\in(0,\pi), j=1,2,\cdots,n$, and assume the zeros x_{nj} are ordered so that $\theta_{nj}>\theta_{n,j-1}, j=2,3,\cdots,n$. By Theorem 9.22 in Nevai [20, p. 167],

$$\theta_{nj}-\theta_{n,j-1}\sim 1/n, \qquad j=1,2,3,\cdots,n.$$

Hence (1.21) follows directly from (1.20). □

3. Weights on ℝ. First, we need some properties of q_n, defined by (1.25).

LEMMA 3.1. *Let $W\in\mathcal{F}$.*

(i) *For large enough n,*

$$1\leqq q_{2n}/q_n\leqq 2.$$

(ii) *If* (1.23) *holds, then there exists $C>1$ such that*

$$q_{2n}/q_n\geqq C>1, \quad \textit{for } n \textit{ large enough.}$$

(iii) *Suppose Q'' satisfies* (1.22). *Let ξ_n denote the positive root of the equation*

$$\xi_n^2 Q''(\xi_n)=n, \tag{3.1}$$

for n large enough. Then

$$q_n\sim\xi_n, \quad \textit{for } n \textit{ large enough.} \tag{3.2}$$

Proof. (i), (ii). These follow exactly as in Freud [8, p. 22].

(iii). By monotonicity of Q'', for n large enough,

$$\begin{aligned} q_nQ'(q_n)=n=\xi_n^2Q''(\xi_n)&\leqq\xi_n\int_{\xi_n}^{2\xi_n}Q''(u)\,du\\ &=\xi_n(Q'(2\xi_n)-Q'(\xi_n))\leqq\xi_nQ'(2\xi_n),\end{aligned}$$

for n large, by (1.22). As $uQ'(u)$ is nondecreasing for large u, we deduce that $q_n\leqq 2\xi_n$. Further for some large enough A,

$$\begin{aligned} q_nQ'(q_n)=n=\xi_n^2Q''(\xi_n)&\geqq\xi_n\int_A^{\xi_n}Q''(u)\,du\\ &=\xi_n(Q'(\xi_n)-Q'(A^+))\\ &\geqq\xi_nQ'(\xi_n)/2, \quad \text{for } n \text{ large enough.}\end{aligned}$$

We deduce that $q_n\geqq\xi_n/2$, for n large enough, as $Q'(u)$ and $uQ'(u)$ are nondecreasing for large enough u. □

LEMMA 3.2 (infinite-finite range inequality). *Let $W\in\mathcal{F}$. Let $0<p\leqq\infty$. There exists a constant C and a positive integer n_0 such that for $n\geqq n_0$ and all polynomials P of degree at most n,*

$$\|PW\|_{L_p(\mathbb{R})}\leqq C\|PW\|_{L_p(-22q_n,22q_n)}. \tag{3.3}$$

Proof. Take $g=1$ in [12, Thm. A, p. 264] and note that $11q_{2n}\leqq 22qn$ by Lemma 3.1(i). □

Next, we need estimates of the Christoffel functions:

LEMMA 3.3. *Let $W\in\mathcal{F}$. Let $0<p\leqq\infty$, and let j be a nonnegative integer. There exists a constant C^* such that*

$$\lambda_{n,p}(W,j,x)\sim(q_n/n)^{j+1/p}W(x), \qquad |x|\leqq C^*q_n. \tag{3.4}$$

If (1.23) *also holds, then*

$$(3.5) \qquad \lambda_{n,p}(W, j, x) \geqq C(q_n/n)^{j+1/p} W(x), \qquad x \in \mathbb{R}.$$

Proof of Theorem 5. First, (1.19) follows easily from Theorem 4 and Lemma 2.3 this follows from Levin and Lubinsky [11, Thm. 3.5]. Suppose now that W satisfies I in Definition 7, so that $Q''(x)$ is positive and nondecreasing for large enough x. Then Lemma 3.1(iii) and Theorem 3.1 in Lubinsky [13] show that there exists C^* such that for $|x| \leqq C^* q_n$,

$$(3.6) \qquad \lambda_{n,p}(W, j, x) \geqq C_1(q_n/n)^{j+1/p} W(x), \qquad |x| \leqq C^* q_n.$$

To prove the upper bounds to match the lower bounds in (3.6), we apply Theorem 3.4 in Lubinsky [13]. Noting that $Q'(x)$ is positive and increasing for large x, we must show the following:

Given $\eta > 0$, there exists $\varepsilon > 0$ and $C > 0$ such that

$$(3.7) \qquad Q'(\varepsilon\xi)/Q'(\xi) < \eta, \qquad \xi > C.$$

There exist $C_1 > 0$ and $C_2 > 0$ such that for $\xi \geqq C_1$,

$$(3.8) \qquad 3\xi Q'(\xi)(\log(|x|/\xi))/Q(x) < 1, \qquad |x| \geqq C_2\xi.$$

The condition [13, (3.10)] holds trivially as $Q'' \geqq 0$. To prove (3.7), we note that if $0 < \varepsilon < 1$, and A is large enough,

$$\begin{aligned} Q'(\varepsilon\xi)/Q'(\xi) &= \left(\int_A^{\varepsilon\xi} Q''(u)\,du + Q'(A+)\right)\Big/\left(\int_A^{\xi} Q''(u)\,du + Q'(A+)\right) \\ &\leqq (\varepsilon\xi Q''(\varepsilon\xi) + Q'(A+))/((\xi - \varepsilon\xi)Q''(\varepsilon\xi) + Q'(A+)) \\ &\to \varepsilon/(1-\varepsilon), \qquad \xi \to \infty. \end{aligned}$$

Next, if $x \geqq 30\xi > C_3$,

$$\begin{aligned} Q(x) &= \int_\xi^x Q'(u)\,du + Q(\xi) > Q'(\xi)(x - \xi) \\ &= \xi Q'(\xi)(29/30)(x/\xi) \geqq 3\xi Q'(\xi)\log(x/\xi), \end{aligned}$$

as $\log u \leqq u/6$ if $u \geqq 30$. Thus (3.8) holds. Then Theorem 3.4 in [13] shows that (3.6) holds with $\geqq$ replaced by $\leqq$, and then (3.4) follows.

Suppose now that (1.23) holds. By Lemma 3.1(ii), we can choose a fixed positive integer k such that $C^* q_{nk} \geqq 22 q_n$, for n large enough. Then (3.6) shows that for $|x| \leqq 22 q_n$,

$$\lambda_{n,p}(W, j, x) \geqq \lambda_{kn,p}(W, j, x) \geqq C_1(q_{kn}/(kn))^{j+1/p} W(x) \geqq C_2(q_n/n)^{j+1/p} W(x).$$

Using the definition (1.26) of $\lambda_{n,p}(W, j, x)$, we see that this last inequality may be rewritten in the following equivalent form: For all polynomials P of degree at most n,

$$\|P^{(j)} W\|_{L_\infty(-22q_n, 22q_n)} \leqq C_1(n/q_n)^{j+1/p} \|PW\|_{L_p(\mathbb{R})}.$$

Then Lemma 3.2 yields

$$\|P^{(j)} W\|_{L_\infty(\mathbb{R})} \leqq C_2(n/q_n)^{j+1/p} \|PW\|_{L_p(\mathbb{R})},$$

and (3.5) follows. □

The following lemma is an analogue of Theorem 4:

LEMMA 3.4. *Let $W \in \mathscr{F}$, $0<p<\infty$ and let l be a positive integer. Further, let $\psi(t)$ be convex, nonnegative and nondecreasing in $[0,\infty)$. There exist C_1, C_2, B and C^* such that*

$$\psi(|PW|^p(x)) \leqq C_1(q_n n)^{-1} \int_{-Bq_n}^{Bq_n} \psi(C_2|PW|^p(u))K_n^2(v, x/(Bq_n), u/(Bq_n))\,du, \tag{3.9}$$

for all $|x| \leqq C^* q_n$ and all polynomials P of degree at most ln. If (1.23) *holds, we may take C^* arbitrarily large, but C, C_1 and C_2 will depend on C^*.*

Proof. Let k be the smallest positive integer such that $kp \geqq 4$. Let

$$B = 44(l+k).$$

It follows easily from $q_{2n} \leqq 2q_n$ (Lemma 3.1(i)) that for any positive integer j, we have

$$q_{jn} \leqq 2jq_n, \quad n \text{ large enough.}$$

Hence

$$22q_{(l+k)n} \leqq Bq_n, \quad n \text{ large enough.} \tag{3.10}$$

Let P have degree at most ln. For each fixed x, $P(t)K_n^k(v, x/(Bq_n), t/(Aq_n))$ has degree less than $(l+k)n$ in t, and Lemma 3.3 shows that there exists C^* such that for $|t| \leqq C^* q_n$,

$$\begin{aligned}
&|PW|^p(t)|K_n^k(v, x/(Bq_n), t/(Bq_n))|^p \\
&\leqq C_2(n/q_n) \int_{-\infty}^{\infty} |PW|^p(u)|K_n(v, x/(Bq_n), u/(Bq_n))|^{kp}\,du \\
&\leqq C_3(n/q_n) \int_{-Bq_n}^{Bq_n} |PW|^p(u)|K_n(v, x/(Bq_n), u/(Bq_n))|^{kp}\,du,
\end{aligned}$$

by Lemma 3.2 and (3.10). As the constants are independent of x, we can set $t = x$, and use (2.4) and (2.5) (as in the proof of Theorem 4—see (2.3) to (2.6)) to deduce that

$$|PW|^p(x) \leqq C_4(n^3 q_n)^{-1} \int_{-Bq_n}^{Bq_n} |PW|^p(u)K_n^4(v, x/(Bq_n), u/(Bq_n))\,du, \tag{3.11}$$

for $|x| \leqq C^* q_n$. To obtain (3.9) from (3.11), we must apply Jensen's inequality. First note that

$$\int_{-Bq_n}^{Bq_n} K_n^4(v, x/(Bq_n), u/(Bq_n))\,du = Bq_n \int_{-1}^{1} K_n^4(v, x/(Bq_n), t)\,dt \sim n^3 q_n, \tag{3.12}$$

for $|x| \leqq Bq_n/2$, by Theorem 2.2 with $w(u) \equiv 1$. We may then use Jensen's inequality and (3.11) and (3.12), in the same way as in the proof of Theorem 4, to deduce (3.9). If (1.23) holds, Lemma 3.3 ensures that lower bounds for $\lambda_{n,p}(W, j, x)$ hold for all $x \in \mathbb{R}$, as obvious modifications of the above argument yield the stronger stated result. □

Proof of Theorem 8. In view of (3.9), it suffices to show that

$$(q_n n)^{-1} \sum_{j=1}^{m} K_n^2(v, y_j/(Bq_n), u/(Bq_n)) \leqq C_3\{n/q_n + \delta^{-1}\}, \tag{3.13}$$

for all $|u| \leqq Bq_n$, with C_3 independent of m, $\{y_j\}$, n, δ and u, provided

$$-C^* q_n \leqq y_m < y_{m-1} < \cdots < y_1 \leqq C^* q_n$$

with C^* small enough and with δ as in (1.28). We can assume that $C^* \leqq B/2$, so that $|y_j/(Bq_n)| \leqq 1/2$, $j = 1, 2, \cdots, m$. Then as d/dx (arc cos $(x)) \sim -1$ for $|x| \leqq 1/2$, we see that

$$\arccos(y_j/(Bq_n)) - \arccos(y_{j-1}/(Bq_n)) \geqq C_4(y_{j-1} - y_j)/q_n \geqq C_4\delta/q_n,$$

by (1.21). Then Lemma 2.3 with $Y = Bq_n$ shows that (3.13) is true. □

Before proving Corollary 9, we need a result on the spacing of the zeros or orthogonal polynomials. Let $p_n(W^2; x)$, $n = 0, 1, 2, \cdots$, denote the orthonormal polynomials for the weight W^2.

LEMMA 3.5. *Let $W \in \mathcal{F}$. There exists C^* such that*

$$x_{n,j-1} - x_{n,j} \sim q_n/n, \qquad |x_{nj}| \leqq C^* q_n.$$

Proof. The proof of the upper bounds on $x_{n,j-1} - x_{n,j}$ in Freud [7, pp. 293-294] requires only suitable upper bounds for $\lambda_n(W^2, x)$ and convexity of Q. The proof of the lower bounds for $x_{n,j-1} - x_{n,j}$ in Freud [8, p 36] uses only suitable upper and lower bounds for $K_n(W^2, x, x)$ and suitable upper bounds for

$$\sum_{j=0}^{n-1} (p_j'(W^2, x))^2 = 1/\{\lambda_{n,2}(W, 1, x)\}^2,$$

as well as the monotonicity of Q'. As Lemma 3.3 yields the desired upper and lower bounds for $\lambda_{n,p}(W, j, x)$ and $\lambda_n(W^2, x)$, the result follows. □

Proof of Corollary 9. If C^* is small enough, Lemma 3.3 shows that

$$\begin{aligned}\sum_{|x_{nj}| \leqq C^* q_n} \lambda_{nj} |P(x_{nj})|^p W^{-r}(x_{nj}) &\leqq C_1(q_n/n) \sum_{|x_{nj}| \leqq C^* q_n} |P(x_{nj})|^p W^{2-r}(x_{nj}) \\ &= C_1(q_n/n) \sum_{|x_{nj}| \leqq C^* q_n} |PW^*(x_{nj})|^p\end{aligned}$$

where $W^*(x) = \exp(-Q^*(x))$ and $Q^*(x) = ((2-r)/p)Q(x)$. If q_n^* denotes the root of the equation $q_n^* Q^{*\prime}(q_n) = n$ for large enough, Lemma 3.1(i) shows $q_n \sim q_n^*$. If $\delta_n = \max\{x_{n,j-1} - x_{n,j}: |x_{nj}| \leqq C^* q_n\}$, Theorem 8 shows that

$$\sum_{|x_{nj}| \leqq C^* q_n} \lambda_{nj} |P(x_{nj})|^p W^{-r}(x_{nj}) \leqq C_2\{1 + (\delta_n n/q_n)^{-1}\} \int_{-\infty}^{\infty} |PW^*(u)|^p \, du$$

and then Lemma 3.5 yields the result. □

REFERENCES

[1] R. ASKEY, *Mean convergence of orthogonal series and lagrange interpolation*, Acta Math. Acad. Sci. Hung., 23 (1972), pp. 71-85.

[2] ———, *Proposed problems*, Proceedings of the Conference on Constructive Theory of Functions, G. Alexits and S. B. Stechkin, eds., Akadémiai Kiadó, Budapest, 1972, p. 533.

[3] S. BONAN, *Weighted mean convergence of Lagrange interpolation*, Ph.D. thesis, Columbus, Ohio, 1982.

[4] H. DAVENPORT AND H. HALBERSTAM, *The values of a trigonometric polynomial at well spaced points*, Mathematika, 13 (1966), pp. 91-96.

[5] M. FORTI AND C. VIOLA, *On the large sieve type estimates for the Dirichlet series operator*, Proc. Symposium in Pure Mathematics, 24, American Mathematical Society, Providence, RI, 1973, pp. 31-49.

[6] G. FREUD, *On polynomial approximation with respect to general weights*, in Lecture Notes in Mathematics 399, H. G. Garnir, K. R. Unni and J. H. Williamson, eds., Springer, Berlin, 1974, pp. 149-179.

[7] ———, *On the theory of one-sided weighted polynomial approximation*, in Approximation Theory and Functional Analysis, P. L. Butzer, J. P. Kahane and B. Sz-Nagy, eds., Birkhaüser, Basel, 1974, pp. 285-303.

[8] G. FREUD, *On Markov-Bernstein type inequalities and their applications*, J. Approx. Theory, 19 (1977), pp. 22-37.

[9] A. KNOPFMACHER AND D. S. LUBINSKY, *Mean convergence of Lagrange interpolation for Freud's weights with application to product integration rules*, J. Comp. Appl. Math., to appear.

[10] A. L. LEVIN AND D. S. LUBINSKY, *Canonical products and the weights* $\exp(-|x|^{\alpha})$, $\alpha > 1$, *with applications*, J. Approx. Theory, to appear.

[11] ———, *Weights on the real line that admit good relative polynomial approximation, with applications*, J. Approx. Theory, to appear.

[12] D. S. LUBINSKY, *A weighted polynomial inequality*, Proc. AMS., 92 (1984), pp. 263-267.

[13] ———, *Estimates of Freud-Christoffel functions for some weights with the whole real line as support*, J. Approx. Theory, 44 (1985), pp. 86-91.

[14] A. MATÉ AND P. NEVAI, *Bernstein's inequality in* L_p *for* $0<p<1$ *and* $(C, 1)$ *bounds for orthogonal polynomials*, Ann. Math., 111 (1980), pp. 145-154.

[15] H. N. MHASKAR AND E. B. SAFF, *Extremal problems for polynomials with exponential weights*, Trans. AMS, 285 (1984), pp. 203-234.

[16] H. L. MONTGOMERY, *The analytic principle of the large sieve*, Bull. AMS, 84 (1978), pp. 547-567.

[17] P. NEVAI, *Polynomials orthogonal on the real axis to the weight* $|x|^{\alpha} \exp(-|x|^{\beta})$. I, Acta Math. Acad. Sci. Hung, 24 (1973), pp. 335-342. (In Russian.)

[18] ———, *Mean convergence of Lagrange interpolation*, J. Approx. Theory, 18 (1976), pp. 363-377.

[19] ———, *Lagrange interpolation at zeros or orthogonal polynomials*, in Approximation Theory II, G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 163-201.

[20] ———, *Orthogonal Polynomials*, Memoirs of the AMS, 18, 1979, 213.

[21] ———, *Bernstein's Inequality in* L_p *for* $0<p<1$, J. Approx. Theory, 27 (1979), pp. 239-243.

[22] ———, *Mean convergence of Lagrange interpolation* II, J. Approx. Theory, 30 (1980), pp. 263-276.

[23] ———, *Mean convergence of Lagrange interpolation* III, Trans. AMS, 282 (1984), pp. 669-698.

[24] A. ZYGMUND, *Trigonometric Series*, Vol. 1, Cambridge Univ. Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

021

JACOBI POLYNOMIALS ASSOCIATED WITH SELBERG INTEGRALS*

K. AOMOTO†

Abstract. An extension of Selberg's beta integral is evaluated and used to obtain a new integral representation for Jacobi polynomials.

Key words. Jacobi polynomials, Selberg integrals

In his work with Y. Kanie on the Fock space representation of the Virasoro algebra, Prof. A. Tsuchiya mentioned the importance of computing the following integral $J_\chi = J_\chi^{(N)}(\lambda, \lambda', \lambda'')$:

$$J_\chi = \int_G \chi(x_3, \cdots, x_N) \prod_{j=3}^{N} x_j^{\lambda'}(1-x_j)^{\lambda''} \prod_{3 \leqq i<j \leqq N} |x_i - x_j|^\lambda \, dx_3 \wedge \cdots \wedge dx_N. \tag{1}$$

The integral is taken over the $(N-2)$-dimensional domain G: $0 \leqq x_j \leqq 1$, $3 \leqq j \leqq N$, where we put $x_1 = 0$ and $x_2 = 1$, and $\chi(x_3, \cdots, x_N)$ is an arbitrary symmetric polynomial. For convergence take $\lambda' > -1$, $\lambda'' > -1$ and for simplicity $\lambda > 0$ (see (Ts)). The one dimensionality of the twisted de Rham cohomology associated with J_χ implies that J_χ / J_1 is equal to a product of certain rational functions of λ, λ', and λ'' (see [Al, p. 177]). However, it seems to be rather difficult to get explicit formulae. J_1 is known, since it is A. Selberg's celebrated formula. See [Se], [Ma], [Mo]. Conjectured q-extensions of J_1 are stated in [As], [Ma], [Mo], [Ra].

In this note we want to show that if $\chi = \prod_{j=3}^N (x_j - t)$, then J_χ / J_1 is equal to

$$P_{N-2}^{(\alpha,\beta)}(1-2t) \frac{(N-2)!}{\prod_{j=3}^N (\alpha + \beta + N + j - 4)} \tag{2}$$

for $\alpha = -1 + 2(\lambda' + 1)/\lambda$, $\beta = -1 + 2(\lambda'' + 1)/\lambda$, where $P_n^{(\alpha,\beta)}(x)$ denotes the Jacobi polynomial of degree n. In the case $\lambda = 2$, this formula reduces to a standard one for orthogonal polynomials; see [Sz, (2.2.10) and § 4.2].

We denote by Φ the function $\prod_{j=3}^N x_j^{\lambda'}(1-x_j)^{\lambda''} \prod_{3 \leqq i<j \leqq N} |x_i - x_j|^\lambda$, by $d\tau$ the $(N-2)$-form $dx_3 \wedge \cdots \wedge dx_N$ and abbreviate by (i, j) the difference $x_i - x_j$. Recall that $x_1 = 0$ and $x_2 = 1$. We start by proving the following.

LEMMA 1.

$$\int_G \Phi \frac{(3,1)(4,1) \cdots (j,1)}{(k,3)} d\tau \tag{3}$$

$$= 0 \quad \textit{for } 4 \leqq k \leqq j \tag{3a}$$

$$= -\frac{1}{2} \int_G \Phi(4,1) \cdots (j,1) \, d\tau \quad \textit{for } j+1 \leqq k \leqq N. \tag{3b}$$

Proof. The domain of integration, is invariant under an arbitrary permutation of the arguments $(3, 4, \cdots, N)$. When $4 \leqq k \leqq j$, the integrand changes sign when 3 and k are transposed. This shows that (3a) holds.

* Received by the editors August 14, 1985; accepted for publication January 7, 1986.

† Department of Mathematics, Nagoya University, Chikusa-ku, Nagoya 464, Japan.

On the other hand, when $j+1\leqq k\leqq N$ the same transposition transforms as follows

$$\frac{(3,1)}{(k,3)}\to\frac{(k,1)}{(3,k)}=-1-\frac{(3,1)}{(k,3)},\tag{4}$$

so the left-hand side of (3) is equal to

$$-\int_G \Phi(4,1)\cdots(j,1)\,d\tau-\int_G \Phi\frac{(3,1)(4,1)\cdots(j,1)}{(k,3)}\,d\tau,\tag{5}$$

and this completes the proof of (3b).

LEMMA 2.

$$\int_G \Phi\frac{(3,1)^2(4,1)\cdots(j,1)}{(k,3)}\,d\tau\tag{6}$$

$$=-\frac{1}{2}\int_G \Phi(3,1)\cdots(j,1)\,d\tau\quad \textit{for } 4\leqq k\leqq j\tag{6a}$$

$$=-\int_G \Phi(3,1)\cdots(j,1)\,d\tau\quad \textit{for } j+1\leqq k\leqq N.\tag{6b}$$

Proof. Transposition of 3 and k transforms

$$\frac{(3,1)^2(k,1)}{(k,3)}\to\frac{(k,1)^2(3,1)}{(3,k)}=-(3,1)(k,1)-\frac{(3,1)^2(k,1)}{(k,3)}\quad\text{for } 4\leqq k\leqq j,\tag{7}$$

while

$$\frac{(3,1)^2}{(k,3)}=-(3,1)+\frac{(3,1)(k,1)}{(k,3)}\quad\text{for } j+1\leqq k\leqq N.\tag{8}$$

The last term changes sign under the transposition of 3 and k, so this integral vanishes. These equalities imply Lemma 2.

LEMMA 3. *For* $3\leqq j\leqq N$

$$-(2,1)\lambda''\int_G \Phi\frac{(4,1)\cdots(j,1)}{(3,2)}\,d\tau=\left(\lambda'+\lambda''+1+\frac{1}{2}(N-j)\lambda\right)\times\int_G \Phi(4,1)\cdots(j,1)\,d\tau.\tag{9}$$

Proof. By Stokes' formula

$$\begin{aligned}O&=\int_G d\{\Phi(3,1)\cdots(j,1)\,dx_4\wedge\cdots\wedge dx_N\}\\&=(\lambda'+1)\int_G \Phi(4,1)\cdots(j,1)\,d\tau\\&\quad+\lambda''\int_G \Phi\frac{(3,1)\cdots(j,1)}{(3,2)}\,d\tau\\&\quad-\lambda\sum_{k=4}^{N}\int_G \Phi\frac{(3,1)\cdots(j,1)}{(k,3)}\,d\tau.\end{aligned}\tag{10}$$

In view of Lemma 1, this is equal to

$$\begin{aligned}\lambda''\int_G \Phi\frac{(2,1)(4,1)\cdots(j,1)}{(3,2)}\,d\tau\\ +\left(\lambda'+\lambda''+1+\frac{1}{2}(N-j)\lambda\right)\int_G \Phi(4,1)\cdots(j,1)\,d\tau\end{aligned}\tag{11}$$

which implies Lemma 3.

LEMMA 4 (Recurrence formula). *For* $3\leqq j\leqq N$,

$$\begin{aligned}\left(\lambda'+\lambda''+2+\frac{1}{2}(j-3)\lambda+(N-j)\lambda\right)\int_G \Phi(3,1)\cdots(j,1)\,d\tau\\ =(2,1)\left(\lambda'+1+\frac{1}{2}(N-j)\lambda\right)\int_G \Phi(4,1)\cdots(j,1)\,d\tau.\end{aligned}\tag{12}$$

Proof. By Stokes' formula again,

$$\begin{aligned}O&=\int_G d\{\Phi(3,1)^2(4,1)\cdots(j,1)\,dx_4\wedge\cdots\wedge dx_N\}\\ &=(\lambda'+2)\int_G \Phi(3,1)\cdots(j,1)\,d\tau\\ &\quad+\lambda''\int_G \Phi\frac{(3,1)^2(4,1)\cdots(j,1)}{(3,2)}\,d\tau\\ &\quad-\lambda\sum_{k=4}^{N}\int_G \Phi\frac{(3,1)^2(4,1)\cdots(j,1)}{(k,3)}\,d\tau.\end{aligned}\tag{13}$$

In view of Lemmas 2 and 3, and seeing that

$$\frac{(3,1)^2}{(3,2)}=(3,1)+(2,1)+\frac{(2,1)^2}{(3,2)}\tag{14}$$

we get

$$\begin{aligned}O&=\left[\lambda'+\lambda''+2+\frac{1}{2}(j-3)\lambda+(N-j)\lambda\right]\int_G \Phi(3,1)\cdots(j,1)\,d\tau\\ &\quad+\lambda''(2,1)\int_G \Phi(4,1)\cdots(j,1)\,d\tau\\ &\quad+\lambda''(2,1)^2\int_G \Phi\frac{(4,1)\cdots(j,1)}{(3,2)}\,d\tau\\ &=\left[\lambda'+\lambda''+2+\frac{1}{2}(j-3)\lambda+(N-j)\lambda\right]\int \Phi(3,1)\cdots(j,1)\,d\tau\\ &\quad-(2,1)\left(\lambda'+1+\frac{1}{2}(N-j)\lambda\right)\int_G \Phi(4,1)\cdots(j,1)\,d\tau\end{aligned}\tag{15}$$

which implies Lemma 4.

Successive applications of Lemma 4 lead to the following theorem.

THEOREM 1.

$$\text{(16)} \qquad \int_G \Phi(3,1)\cdots(j,1)\,d\tau = \int_G \Phi\,d\tau \prod_{i=3}^{j} \frac{\left(\lambda'+1+\frac{1}{2}(N-i)\lambda\right)}{\left(\lambda'+\lambda''+2+\lambda\left(N-\frac{i}{2}-\frac{3}{2}\right)\right)}.$$

This identity and the symmetry of Φ give the following.

THEOREM 2.

$$\text{(17)} \qquad \begin{aligned} \int_G \Phi \prod_{j=3}^{N} (x_j-t)\,d\tau &= \sum_{r=2}^{N} (-t)^{N-r}\binom{N-2}{r-2}\prod_{j=3}^{r} \frac{\left[\lambda'+1+\frac{1}{2}(N-j)\lambda\right]}{\left[\lambda'+\lambda''+2+\lambda\left(N-\frac{j}{2}-\frac{3}{2}\right)\right]} \\ &= \frac{(N-2)!}{\prod_{j=3}^{N}(\alpha+\beta+N-4+j)} P_{N-2}^{(\alpha,\beta)}(1-2t) \end{aligned}$$

for $\alpha=(2(\lambda'+1)/\lambda)-1$ *and* $\beta=(2(\lambda''+1)/\lambda)-1$, *where* $P_n^{(\alpha,\beta)}(x)$ *denotes the Jacobi polynomial of degree n defined by*

$$\text{(18)} \qquad \begin{aligned} P_n^{(\alpha,\beta)}(x) &= \frac{1}{n!}\sum_{\nu=0}^{n}\binom{n}{\nu}(n+\alpha+\beta+1)\cdots(n+\alpha+\beta+\nu), \\ &(\alpha+\nu+1)\cdots(\alpha+n)\left(\frac{x-1}{2}\right)^{\nu}. \end{aligned}$$

It has been remarked by Prof. R. Askey that if we take $\chi=\prod_{i=3}^{r}(x_i-t)$, then the quotient J_χ/J_1 is also equal to a constant multiple of $P_{r-2}^{(N-r+\alpha,N-r+\beta)}(1-2t)$. This easily follows from (16).

He also pointed out the following: If λ, λ', λ'' are all nonnegative integers, then (16) implies the following recurrence relation:

$$J_1^{(N)}(\lambda,\lambda',\lambda'') = \frac{\prod_{k=1}^{\lambda'}\prod_{j=3}^{N}\left\{k+\frac{1}{2}(N-j)\lambda\right\}\prod_{k=1}^{\lambda''}\prod_{j=3}^{N}\left\{k+\frac{1}{2}(N-j)\lambda\right\}}{\prod_{k=1}^{\lambda'+\lambda''}\prod_{j=3}^{N}\left\{k+1+\lambda\left(N-\frac{j}{2}-\frac{3}{2}\right)\right\}} J_1^{(N)}(\lambda,0,0).$$

This equality is nothing else than the key formula (5) in [Se]. By the change of variables $x_j=x_3t_j$ for $j\geqq 4$, we have

$$J_1^{(N)}(\lambda,0,0) = \frac{1}{(N-2)\left(1+\frac{N-3}{2}\lambda\right)} J_1^{(N-1)}(\lambda,0,\lambda).$$

The proof of Selberg's formula can then be completed along the same line as in [Se] using Carlson's theorem.

Note added in proof. K. W. J. Kadell has recently proved a conjecture of Askey for a q-analogue of Selberg's integral.

REFERENCES

[Al] K. Aomoto, *Configurations and invariant theory of Gauss-Manin systems*, in Adv. Study in Pure Math., 4, Group Representations and Systems of Differential Equations, K. Okamoto, ed., Kinokuniya, Tokyo and North-Holland, Amsterdam, 1984, pp. 165-179.

[A2] ———, *Gauss-Manin system of integrals of difference products*, J. Math. Soc. Jap., submitted.

[As] R. Askey, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, this Journal, 11 (1980), pp. 938-951.

[Ma] I. G. Macdonald, *Some conjectures for root systems*, this Journal, 13 (1982), pp. 988-1007.

[Mo] W. G. Morris II, *Constant term identities for finite and affine root systems, conjectures and theorems*, Ph.D. dissertation, Univ. Wisconsin, Madison, 1982.

[Ra] M. Rahman, *Another conjectured q-Selberg integral*, this Journal, 17 (1986), pp. 1267-1279.

[Se] A. Selberg, *Bemerkninger om et multipelt integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71-78.

[Sz] G. Szegö, *Orthogonal polynomials*, Amer. Math. Soc. Colloq. Publ., 4th edition, 1975.

[Ts] A. Tsuchiya and T. Kanie, *Fock space representation of the Virasoro algebra—Intertwining operators*, Pub. RIMS, Kyoto Univ., 22 (1986), pp. 259-327.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

022

DIRICHLET AVERAGES OF $x^t \log x$*

B. C. CARLSON†

Abstract. A neglected class of special functions may be described as Dirichlet averages of $x^t \log x$ or equivalently as derivatives of hypergeometric R-functions with respect to the degree of homogeneity. Special cases include the derivative of a Legendre function with respect to the degree and the derivative of Gauss's hypergeometric function with respect to a numerator parameter. There are connections with the logarithmic derivative of the gamma function, with Euler's dilogarithm, and with ${}_3F_2$; some special cases of ${}_3F_2$ are thereby evaluated. Applications include a two-point boundary-value problem, mean values, series expansions of elliptic integrals, integral tables, and several physics problems. The discussion of various properties emphasizes series expansions, quadratic transformations, inequalities, and evaluation of special cases, including certain cases of the derivative of a Legendre function with respect to the degree.

Key words. Dirichlet averages, logarithms, hypergeometric functions, R-functions, elliptic integrals, Legendre functions

AMS (MOS) subject classifications. Primary 33A30; secondary 33A25, 33A45

1. Introduction. A class of special functions that seems not to have been discussed systematically arises in the asymptotic expansion of elliptic integrals and in various other problems. The prototype of this class is the derivative of a Legendre function P_ν with respect to the degree ν. The general case is the derivative of the multivariate hypergeometric function R_t with respect to the degree t of homogeneity. It can equivalently be thought of as a Dirichlet average of $\partial x^t/\partial t = x^t \log x$. After defining this average, denoted by L_t, we shall list some places where it occurs.

We recall first the definition of the R-function [3, (5.9-1)]. Let b and z be k-tuples with components in the open right-half complex plane, $\mathbb{C}_>$. The Dirichlet average of x^t, $t \in \mathbb{C}$, is defined for $k \geqq 2$ by

$$(1.1) \qquad R_t(b, z) = \int (u \cdot z)^t \, d\mu_b(u),$$

where $u \cdot z = \sum_{i=1}^k u_i z_i$ is a convex combination of the variables $z_1, \cdots, z_k$ and μ_b is a Dirichlet measure (i.e. a multivariate beta distribution with possibly complex parameters $b_1, \cdots, b_k$). The integration extends over all nonnegative weights $u_1, \cdots, u_k$ whose sum is unity. If $k = 1$ we define $R_t(b, z) = z^t$. The analytic continuation [3, Sec. 6.8] of $R_t(b, z)/\Gamma(c)$, $c = \sum_{i=1}^k b_i$, is entire in t and b and holomorphic in z on $\mathbb{C}_0^k$, where $\mathbb{C}_0$ is the complex plane slit along the nonpositive real axis.

The subject of this paper is the function L_t defined by

$$(1.2) \qquad L_t(b, z) = \frac{\partial}{\partial t} R_t(b, z).$$

If b, $z \in \mathbb{C}_>^k$, $k \geqq 2$, we see from (1.1) that

$$(1.3) \qquad L_t(b, z) = \int (u \cdot z)^t \log (u \cdot z) \, d\mu_b(u).$$

* Received by the editors October 10, 1985; accepted for publication (in revised form) January 23, 1986. This work was supported by the Director of Energy Research, Office of Basic Energy Sciences. The Ames Laboratory is operated for the U.S. Department of Energy by Iowa State University under contract W-7405-ENG-82.

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

That is, L_t is a Dirichlet average of $x^t \log x$. The case $t=0$ is discussed briefly in [3, Ex. 5.9-12 and p. 305]. If $k=1$ we define $L_t(b, z)=z^t \log z$. We shall often display the components of b and z by writing $L_t(b_1, \cdots, b_k; z_1, \cdots, z_k)$. If $k=2$, $b=(\beta, \beta')$ and $z=(x, y)$, (1.3) becomes

$$L_t(\beta, \beta'; x, y)=\frac{\Gamma(\beta+\beta')}{\Gamma(\beta)\Gamma(\beta')}\int_0^1 [ux+(1-u)y]^t \log[ux+(1-u)y] \cdot u^{\beta-1}(1-u)^{\beta'-1}\,du, \tag{1.4}$$

where β, β', x, y have positive real parts and t is unrestricted. The case in which x or y is 0 will be discussed in § 4.

We list some examples to show that L_t is worth studying.

Example 1. The difference between two values of the logarithmic derivative of the gamma function is

$$\psi(\beta)-\psi(\gamma)=L_0(\beta, \gamma-\beta; 1, 0), \quad \operatorname{Re}\beta>0, \quad \gamma\neq 0, -1, -2, \cdots. \tag{1.5}$$

Example 2. The derivative of Gauss' hypergeometric function with respect to a numerator parameter is

$$\frac{\partial}{\partial\alpha}\,{}_2F_1(\alpha, \beta; \gamma; z)=-L_{-\alpha}(\beta, \gamma-\beta; 1-z, 1). \tag{1.6}$$

If α is a negative integer, the ${}_2F_1$-function is a polynomial; if α is close to a negative integer, the function can be approximated to first order with the help of the derivative. Equation (1.6) can easily be generalized from ${}_2F_1$ to Appell's F_1 or Lauricella's F_D by expressing these functions in terms of R_t (e.g. see [3, Ex. 6.3-5]).

Example 3. The derivative of an associated Legendre function with respect to its degree is

$$\frac{\partial}{\partial\nu}P_\nu^{-\mu}(\cos\theta)=\frac{\sin^\mu\theta}{2^\mu\Gamma(1+\mu)}L_{\nu-\mu}(\mu+1/2, \mu+1/2; e^{i\theta}, e^{-i\theta}). \tag{1.7}$$

This derivative is useful in antenna theory [13, pp. 114–115]. Special cases are evaluated in closed form in (6.14) and (6.19).

Example 4. For any real t and positive x, a, b, the solution of the two-point boundary-value problem $y''=x^t \log x$, $y(a)=y(b)=0$, is

$$y(x)=\tfrac{1}{2}(x-a)(x-b)L_t(1, 1, 1; x, a, b). \tag{1.8}$$

Closed formulas are given in (8.10), (8.12), and (8.14).

Example 5. Let $x_1, \cdots, x_k$ be positive numbers and $w_1, \cdots, w_k$ positive weights with $\sum w_i=1$. Then $\exp L_0(cw_1, \cdots, cw_k; x_1, \cdots, x_k)$, $c>0$, is a homogeneous mean value of the x_i that increases strictly with c unless the x_i are all equal. It tends to the weighted geometric mean $\prod x_i^{w_i}$ as $c\to 0+$ and the weighted arithmetic mean $\sum w_i x_i$ as $c\to+\infty$.

Example 6. Series expansions of elliptic integrals and some other R-functions [6], [5] near their singularities involve L_t. For example [6, (1.20)],

$$\begin{aligned}(zw)^{1/2}\int_0^\infty [(t+x)(t+y)(t+z)(t+w)]^{-1/2}\,dt &= \sum_{n=0}^\infty [-L_n(1/2, 1/2; x, y)R_n(1/2, 1/2; z^{-1}, w^{-1}) \\ &\qquad - R_n(1/2, 1/2; x, y)L_n(1/2, 1/2; z^{-1}, w^{-1})]\end{aligned} \tag{1.9}$$

where x and y are nonnegative, z and w are positive, and $0<\max\{x,y\}/\min\{z,w\}<1$. Convergence is fast near the logarithmic singularity at $x=y=0$. A closed formula for $L_n(1/2, 1/2; x, y)$, which was not known when [6] was written, is given in (6.13).

Example 7. Integrals with a logarithm in the integrand are met in various applied problems and can sometimes be put in the form (1.4). Two such integrals occurring in diffusion problems associated with crystal growth [14, (17), (36)] [11, (A4)] are constant multiples of $L_0(1/2, 1/2+1/n; z^2-1, z^2)$ and $L_0(1+a, 1-a; z-1, z+1)$, where n is a positive integer and a is an angle in units of π. In connection with the first example, Seeger [14, p. 6] remarks that he "did not succeed in expressing [the concentration] in terms of well-investigated functions." A more complicated integral (not found in tables) containing a logarithm and a Bessel function is

$$\begin{aligned}\int_0^\infty e^{-pt}\log(t)t^\nu J_\nu(xt)\,dt &= \pi^{-1/2}\Gamma(\nu+1/2)(2x)^\nu(p^2+x^2)^{-\nu-1/2}\\ &\quad\cdot[\psi(2\nu+1)-\log(p^2+x^2)+\tfrac{1}{2}L_0(\nu+1/2, 1/2; p^2, p^2+x^2)],\end{aligned}\tag{1.10}$$

where p, x, and $\nu+1/2$ are positive. An example with an integrand not containing a logarithm is

$$\int_0^1 u^{-1}(1-u)^{a-1}[{}_2F_1(a, b; c; ux)-1]\,du = -L_0(b, c-b; 1-x, 1),\tag{1.11}$$

where Re $a>0$, $|\arg(1-x)|<\pi$, and $c\neq 0, -1, -2, \cdots$. The case $a=b=\frac{1}{2}$, $c=1$ occurs in the theory of neutron stars [8, Appendix B].

Equation (1.5) is a special case of (4.4). Equations (1.6) and (1.7) follow, respectively, from differentiating [3, (5.9-12) and (6.8-19)]. It can be verified that (1.8) satisfies the differential equation by using [3, (5.6-5), (5.6-10), (6.3-3)] to prove a more general result for $y''=f(x)$. Example 5 comes from Theorem 7.2 or from [1, (2.6) and Thm. 1] and [7, Thm. 4]. Equation (1.9) is taken from [6, (1.18), (3.17)]. Equation (1.10) is proved by differentiating [3, Ex. 5.10-3] with respect to λ and using [3, (5.9-23)] and Equations (2.11), (6.4), and (2.3) of this paper. Equation (1.11) is derived by integrating the ${}_2F_1$-series term by term and comparing with (5.13).

This paper does not discuss higher derivatives of R_t with respect to t nor double Dirichlet averages of $x^t\log x$, although the latter occur in some problems such as the two-dimensional Ising model.

2. General properties.

Several important properties of L_t follow immediately from the theory of Dirichlet averages or from properties of R_t:

(2.1) $L_t(b, z)/\Gamma(c)$, $c=\sum_{i=1}^k b_i$, is entire in t and b and holomorphic in z on $\mathbb{C}_0^k$. See (1.2).

(2.2) $L_t(b_1, \cdots, b_k; z_1, \cdots, z_k)$ is symmetric in the indices $1, \cdots, k$ [3, Thm. 5.2-3].

(2.3) A vanishing parameter b_i can be omitted along with the corresponding variable z_i [3, (6.3-3)].

(2.4) Equal variables can be replaced by a single variable if the corresponding parameters are replaced by their sum [3, (5.2.-3)].

In particular, if all variables are equal, then

$$L_t(b_1, \cdots, b_k; x, \cdots, x) = L_t(c, x) = x^t\log x.$$

From (1.3) we obtain

$$
\begin{aligned}
L_t(b, \lambda z) &= \lambda^t L_t(b, z) + \lambda^t R_t(b, z) \log \lambda \\
&= \lambda^t L_t(b, z) + R_t(b, \lambda z) \log \lambda, \\
L_0(b, \lambda z) &= L_0(b, z) + \log \lambda.
\end{aligned} \tag{2.5}
$$

Differentiation of [3, (6.8-15)] gives

$$L_t(b, z) = -\left(\prod_{i=1}^{k} z_i^{-b_i}\right) L_{-c-t}(b, z^{-1}), \tag{2.6}$$

where $c = \sum_{i=1}^{k} b_i \neq 0, -1, -2, \cdots$; $z \in \mathbb{C}_0^k$; and $z^{-1} = (z_1^{-1}, \cdots, z_k^{-1})$.

Since L_t is a Dirichlet average, it satisfies a system of Euler-Poisson equations [3, (5.4-2)],

$$[(z_i - z_j) D_i D_j + b_i D_j - b_j D_i] L_t(b, z) = 0, \qquad i, j = 1, 2, \cdots, k, \tag{2.7}$$

where $D_i = \partial/\partial z_i$. Differentiating [3, (5.9-2)] with respect to t, we find also the inhomogeneous differential equation

$$\sum_{i=1}^{k} z_i D_i L_t = t L_t + R_t \tag{2.8}$$

and the differential-difference equation

$$\sum_{i=1}^{k} D_i L_t = t L_{t-1} + R_{t-1}. \tag{2.9}$$

If we define $x + z = (x + z_1, \cdots, x + z_k)$, where z is independent of x, then (2.9) implies

$$\frac{d}{dx} L_t(b, x + z) = t L_{t-1}(b, x + z) + R_{t-1}(b, x + z). \tag{2.10}$$

Some useful relations are peculiar to the case of two variables:

$$L_t(\beta, \beta'; x, y) = \log(xy) R_t(\beta, \beta'; x, y) - x^{t+\beta'} y^{t+\beta} L_{-\beta-\beta'-t}(\beta', \beta; x, y), \tag{2.11}$$

where $x, y \in \mathbb{C}_0$ and $\beta + \beta' \neq 0, -1, -2, \cdots$. This follows from (2.6) and (2.5) or alternatively from [3, (5.9-21)].

$$
\begin{aligned}
\frac{\partial}{\partial t} R_{-a}(u + t, v - t; x, y) &= x^{-a} L_{t-v}(u + v - a, a; 1, y/x) \\
&= -y^{-a} L_{-u-t}(a, u + v - a; x/y, 1),
\end{aligned} \tag{2.12}
$$

where $x, y \in \mathbb{C}_0$ and $u + v \neq 0, -1, -2, \cdots$. On the left side use homogeneity to arrive at arguments 1 and y/x and then apply [3, (5.9-20)]. The second member equals the third by (2.6) and (2.2). We shall use (2.12) to derive (6.4) and (6.5).

In addition to the representation of L_t by the multiple integral (1.3), a representation by a single integral is obtained by differentiating the corresponding representation [3, (6.8-6), Ex. 6.8-8] of R_t:

$$
\begin{aligned}
B(a, a')\{L_{-a}(b, z) &+ [\psi(a') - \psi(a)] R_{-a}(b, z)\} \\
&= \int_0^{\infty} t^{a'-1} \log(t) \prod_{i=1}^{k} (t + z_i)^{-b_i}\, dt \\
&= -\int_0^{\infty} t^{a-1} \log(t) \prod_{i=1}^{k} (1 + t z_i)^{-b_i}\, dt,
\end{aligned} \tag{2.13}
$$

where B is the beta function, $a+a'=\sum_{i=1}^k b_i$, a and a' have positive real parts, and $z\in\mathbb{C}_0^k$. For integrals with finite limits of integration, we find by differentiating [3, (8.1-2)] with respect to a that

$$\begin{aligned}\int_x^y (t-x)^{a-1}(y-t)^{a'-1}\log\left(\frac{y-t}{t-x}\right)\prod_{i=1}^k (z_i+w_it)^{-b_i}\,dt \\ =B(a,a')(y-x)^{a+a'-1}\prod_{i=1}^k (z_i+w_ix)^{-b_i} \\ \cdot\left\{L_{-a}\left(b,\frac{z+wy}{z+wx}\right)+[\psi(a')-\psi(a)]R_{-a}\left(b,\frac{z+wy}{z+wx}\right)\right\} \\ =B(a,a')(y-x)^{a+a'-1}\prod_{i=1}^k (z_i+w_iy)^{-b_i} \\ \cdot\left\{-L_{-a'}\left(b,\frac{z+wx}{z+wy}\right)+[\psi(a')-\psi(a)]R_{-a'}\left(b,\frac{z+wx}{z+wy}\right)\right\}\end{aligned} \tag{2.14}$$

where the conditions on the parameters are the same as for (2.13), $y>x$, $(z+wy)/(z+wx)$ denotes the k-tuple with ith component $(z_i+w_iy)/(z_i+w_ix)$, and $z_i+w_it\in\mathbb{C}_0$ for every $t\in[x,y]$ and every $i=1,\cdots,k$. The equality $a+a'=\sum_{i=1}^k b_i$ can always be satisfied by choice of b_k if we put $z_k=1$ and $w_k=0$.

An integral containing the confluent hypergeometric function S (the Dirichlet average of e^x) can be evaluated by differentiating [3, (5.10-11)] with respect to a:

$$\int_0^\infty t^{a-1}\log(t)S(b,-tz)\,dt=\Gamma'(a)R_{-a}(b,z)-\Gamma(a)L_{-a}(b,z), \tag{2.15}$$

where a and all components of z have positive real parts and $\sum_{i=1}^k b_i\neq 0,-1,-2,\cdots$. The functions of Kummer, Bessel and Whittaker are expressed in terms of S by [3, (5.8-7), (5.8-23), (5.12-20 to 27)].

3. Relations between associated functions. Two or more L-functions of k variables are said to be associated if the parameters $t, b_1,\cdots,b_k$ of each function differ by integers from the corresponding parameters of the other functions.

THEOREM 3.1. *Between any $k+1$ associated L-functions of k variables there exists a linear (possibly inhomogeneous) relation with coefficients that are polynomials in the variables and parameters. The inhomogeneous term, if any, is a linear combination of R-functions with polynomial coefficients.*

Proof. There is a linear homogeneous relation between any $k+1$ associated R-functions [3, Thm. 8.4-3] in which the coefficients are polynomials in the parameters as well as the variables. By (1.2), differentiation with respect to t produces a linear relation of the kind described between the corresponding L-functions.

Let $w_i=b_i/c$, $c=\sum_{i=1}^k b_i$, and let e_i be a k-tuple with unity in the ith place and zeros elsewhere. We list first some homogeneous relations:

$$L_t(b,z)=\sum_{i=1}^k w_iL_t(b+e_i,z), \tag{3.1}$$

$$L_{t+1}(b,z)=\sum_{i=1}^k w_iz_iL_t(b+e_i,z), \tag{3.2}$$

$$(z_i-z_j)L_t(b-e_h,z)+(z_j-z_h)L_t(b-e_i,z)+(z_h-z_i)L_t(b-e_j,z)=0. \tag{3.3}$$

The first and third are special cases of [3, (5.6-4), (5.6-11)], while the second comes from differentiating [3, (5.9-6)].

Differentiation of [3, (5.9-8), (5.9-9), (5.9-10)] yields, for $i=1,2,\cdots,k$,

$$(c-1)L_t(b-e_i,z)=(c+t-1)L_t(b,z)-tz_iL_{t-1}(b,z)+R_t(b,z)-z_iR_{t-1}(b,z), \tag{3.4}$$

$$D_iL_t(b,z)=w_itL_{t-1}(b+e_i,z)+w_iR_{t-1}(b+e_i,z), \tag{3.5}$$

$$(z_iD_i+b_i)L_t(b,z)=w_i(c+t)L_t(b+e_i,z)+w_iR_t(b+e_i,z). \tag{3.6}$$

From [3, Ex. 5.9-6] we see that

$$t(z_i-z_j)L_{t-1}(b,z)+(c-1)[L_t(b-e_i,z)-L_t(b-e_j,z)]+(z_i-z_j)R_{t-1}(b,z)=0, \tag{3.7}$$

$$\begin{aligned}(c+t-1)(z_i-z_j)L_t(b,z)+(c-1)[z_jL_t(b-e_i,z)-z_iL_t(b-e_j,z)]\\ +(z_i-z_j)R_t(b,z)=0.\end{aligned} \tag{3.8}$$

In the case of two variables, differentiation of [3, (5.9-24), (5.9-25)] gives

$$w_i(z_j-z_i)L_t(b+e_i,z)=z_jL_t(b,z)-L_{t+1}(b,z), \qquad j=3-i, \quad i=1,2, \tag{3.9}$$

$$\begin{aligned}(c+t)L_{t+1}(b,z)-\sum_{i=1}^{2}(b_i+t)z_iL_t(b,z)+tz_1z_2L_{t-1}(b,z)\\ =-R_{t+1}(b,z)+(z_1+z_2)R_t(b,z)-z_1z_2R_{t-1}(b,z)\\ =\frac{b_1b_2}{c(c+1)}(z_1-z_2)^2R_{t-1}(b+e_1+e_2,z)\\ =\frac{(z_1-z_2)^2}{t(t+1)}D_1D_2R_{t+1}(b,z).\end{aligned} \tag{3.10}$$

In (3.10), where $c=b_1+b_2$, the second and third equalities come from two applications of [3, (5.9-24)] and [3, (5.9-9)], respectively. The case with t a nonnegative integer and $b=(1/2,1/2)$ was quoted in [6, (2.24)], where it was used to show that the quantity $\lambda_n(1/2,1/2;x,y)$ defined by [6, (1.13)] is a homogeneous polynomial of degree $2n$ in $x^{1/2}$ and $y^{1/2}$ containing $(x^{1/2}-y^{1/2})^2$ as a factor. That conclusion is confirmed by (6.13) of the present paper, which yields the formula

$$\begin{aligned}\lambda_n(1/2,1/2;x,y)=2\sum_{s=1}^{n}\binom{n}{s}\binom{n+s}{s}[\psi(n+1+s)-\psi(n+1)]\\ \cdot\left(\frac{x^{1/2}-y^{1/2}}{2}\right)^{2s}(xy)^{(n-s)/2}.\end{aligned} \tag{3.11}$$

4. The case of a zero argument. If all components of z except one, say z_k, are fixed, both R_t and L_t have a branch point at $z_k=0$. (In the exceptional case where b_k is a nonpositive integer, L_t is a polynomial in z_k, for b_k can be raised to 0 by successive applications of (3.4) and then omitted by (2.3). Equation (5.17) is an example. Similar remarks apply to R_t.) If $\operatorname{Re}(b_1+\cdots+b_{k-1}+t)>0$, R_t and L_t have finite limits as $z_k\to 0$ in $\mathbb{C}_0$ with $|\arg z_k|$ bounded away from π, and we define the value of the function at $z_k=0$ to be this limit. By an extension [3, (8.3-4)] of Gauss' theorem for a hypergeometric function with unit argument,

$$\begin{aligned}R_t(b_1,\cdots,b_k;z_1,\cdots,z_{k-1},0)\\ =\frac{\Gamma(c)\Gamma(c-b_k+t)}{\Gamma(c+t)\Gamma(c-b_k)}R_t(b_1,\cdots,b_{k-1};z_1,\cdots,z_{k-1}),\end{aligned} \tag{4.1}$$

where $c=b_1+\cdots+b_k \neq 0, -1, -2, \cdots$; $\operatorname{Re}(c-b_k+t)>0$; and $z_1, \cdots, z_{k-1} \in \mathbb{C}_0$. Logarithmic differentiation with respect to t shows that

$$\begin{aligned}&\frac{L_t(b_1, \cdots, b_k; z_1, \cdots, z_{k-1}, 0)}{R_t(b_1, \cdots, b_k; z_1, \cdots, z_{k-1}, 0)}\\ &\quad=\psi(c-b_k+t)-\psi(c+t)+\frac{L_t(b_1, \cdots, b_{k-1}; z_1, \cdots, z_{k-1})}{R_t(b_1, \cdots, b_{k-1}; z_1, \cdots, z_{k-1})},\end{aligned} \tag{4.2}$$

with the further conditions that $c+t$, $c-b_k \neq 0, -1, -2, \cdots$. Although (4.2) will be used in the remark following Theorem 7.1, a better form for other purposes is obtained by ordinary differentiation of (4.1):

(4.3)

$$\begin{aligned}&L_t(b_1, \cdots, b_k; z_1, \cdots, z_{k-1}, 0)\\ &\quad=\frac{\Gamma(c)\Gamma(c-b_k+t)}{\Gamma(c+t)\Gamma(c-b_k)}\{L_t(b_1, \cdots, b_{k-1}; z_1, \cdots, z_{k-1})\\ &\qquad\qquad+[\psi(c-b_k+t)-\psi(c+t)]R_t(b_1, \cdots, b_{k-1}; z_1, \cdots, z_{k-1})\},\end{aligned}$$

with the same conditions of validity as for (4.1). If $k=2$ this reduces to

$$L_t(\beta, \gamma-\beta; x, 0)=\frac{\Gamma(\gamma)\Gamma(\beta+t)}{\Gamma(\gamma+t)\Gamma(\beta)}x^t[\psi(\beta+t)-\psi(\gamma+t)+\log x], \tag{4.4}$$

where β, γ, $\gamma+t\neq 0$, -1, $-2, \cdots$ and $\operatorname{Re}(\beta+t)>0$. The elliptic integral (1.9) is equivalent to Legendre's first integral if $w=\infty$, and its series expansion then contains the special case of (4.4) with $\beta=\frac{1}{2}$, $\gamma=1$, and t a nonnegative integer. This case was quoted in [6, (1.10)], with the difference of ψ-functions given by [6, (1.11)] or [3, (8.3-15)]. Other special cases of (4.4) are (1.5) and

$$L_0(1, m; x, 0)=\log x-(1+1/2+1/3+\cdots+1/m), \tag{4.5}$$

where m is a positive integer. If $x=m$, this tends to the negative of the Euler-Mascheroni constant as $m\to\infty$.

5. Series expansions. Let all parameters and variables be complex, define $1-z=(1-z_1, \cdots, 1-z_k)$ and $|1-z|=\max_i |1-z_i|$, and assume $|1-z|<1$ and $b_1+\cdots+b_k\neq 0$, -1, $-2, \cdots$. The series expansion [3, (5.9-4)]

$$R_t(b, z)=\sum_{s=0}^{\infty}\frac{(-t)_s}{s!}R_s(b, 1-z) \tag{5.1}$$

is the Dirichlet average of the binomial series

$$x^t=\sum_{s=0}^{\infty}\frac{(-t)_s}{s!}(1-x)^s, \qquad |1-x|<1, \tag{5.2}$$

where $(-t)_0=1$ and $(-t)_s=(-t)(-t+1)(-t+2)\cdots(-t+s-1)$, $s=1, 2, 3, \cdots$. Differentiation with respect to t is essentially the same for both series. By [3, proof of Cor. 6.3-4] both converge uniformly for $|t|<T$, where T may be arbitrarily large, and so the series may be differentiated term by term. Since $(-t)_0=1$ we assume s to be a positive integer and find

$$\begin{aligned}(d/dt)(-t)_s&=(-1)(-t+1)(-t+2)\cdots(-t+s-1)\\ &\quad+(-t)(-1)(-t+2)\cdots(-t+s-1)+\cdots\\ &\quad+(-t)(-t+1)\cdots(-t+s-2)(-1).\end{aligned} \tag{5.3}$$

If $(-t)_s \neq 0$ we can rewrite this as

$$(5.4)\qquad \begin{aligned}(d/dt)(-t)_s &= (-t)_s\left(\frac{1}{t}+\frac{1}{t-1}+\cdots+\frac{1}{t+1-s}\right)\\ &= (-t)_s[\psi(t+1)-\psi(t+1-s)].\end{aligned}$$

If $t=n$ where $n=0,1,\cdots,s-1$, then (5.4) is indeterminate, but all terms on the right side of (5.3) vanish except one, leaving

$$(5.5)\qquad [(d/dt)(-t)_s]_{t=n} = (-1)^{n+1}n!(s-n-1)!, \qquad n=0,1,\cdots,s-1.$$

Therefore, if $t\neq 0,1,2,\cdots$, differentiation of (5.1) yields

$$(5.6)\qquad L_t(b,z)=\sum_{s=1}^{\infty}\frac{(-t)_s}{s!}[\psi(t+1)-\psi(t+1-s)]R_s(b,1-z), \qquad |1-z|<1.$$

On the other hand, if $n=0,1,2,\cdots$, we find

$$(5.7)\qquad \begin{aligned}L_n(b,z) &= \sum_{s=1}^{n}\frac{(-n)_s}{s!}[\psi(n+1)-\psi(n+1-s)]R_s(b,1-z)\\ &\quad +(-1)^{n+1}n!\sum_{s=n+1}^{\infty}\frac{(s-n-1)!}{s!}R_s(b,1-z), \qquad |1-z|<1.\end{aligned}$$

The first sum is empty if $n=0$, whence

$$(5.8)\qquad L_0(b,z)=-\sum_{s=1}^{\infty}\frac{1}{s}R_s(b,1-z), \qquad |1-z|<1,$$

while

$$(5.9)\qquad \begin{aligned}L_1(b,z) &= -R_1(b,1-z)+\sum_{s=2}^{\infty}\frac{1}{s(s-1)}R_s(b,1-z)\\ &= L_0(b,z)+\sum_{s=2}^{\infty}\frac{1}{s-1}R_s(b,1-z), \qquad |1-z|<1.\end{aligned}$$

The second equality in (5.9) is generalized by

$$(5.10)\qquad \sum_{s=1}^{\infty}\frac{1}{s}R_{N+s}(b,z)=\sum_{n=0}^{N}(-1)^{n+1}\binom{N}{n}L_n(b,1-z),$$

where $|z|<1$ and $N=0,1,2,\cdots$. This equation is proved by taking the Dirichlet average of the case in which all components of z are equal.

If all components of z are equal, (5.7) reduces to

$$(5.11)\qquad \begin{aligned}&x^n\log x-\sum_{s=1}^{n}\frac{(-n)_s}{s!}[\psi(n+1)-\psi(n+1-s)](1-x)^s\\ &\quad =(-1)^{n+1}n!\sum_{s=n+1}^{\infty}\frac{(s-n-1)!}{s!}(1-x)^s, \qquad |1-x|<1,\\ &\quad =\frac{(x-1)^{n+1}}{n+1}\,{}_2F_1(1,1;n+2;1-x)\\ &\quad =\frac{(x-1)^{n+1}}{n+1}R_{-1}(1,n+1;x,1), \qquad |\arg x|<\pi.\end{aligned}$$

We shall use (5.11) to prove (5.15).

An explicit formula [3, (6.2-1)] and a recurrence relation [4, (A.6)] are available for computing the polynomials R_s in (5.6) and (5.7). Moreover we may use (2.5) to make one component of z equal to unity; then R_s has one vanishing component that can be eliminated by [3, (6.2-5), (6.2-6)]. If $k=2$ this leaves a power series in one variable:

$$L_t(\beta, \gamma-\beta; 1-x, 1)=\sum_{s=1}^{\infty}\frac{(-t)_s(\beta)_s}{(\gamma)_s s!}[\psi(t+1)-\psi(t+1-s)]x^s, \tag{5.12}$$

where $|x|<1$, $\gamma\neq 0, -1, -2, \cdots$, and $t\neq 0, 1, 2, \cdots$. From (5.7) we obtain similarly, for $n=0, 1, 2, \cdots$ and $\gamma\neq 0, -1, -2, \cdots$,

$$\begin{aligned} &L_n(\beta, \gamma-\beta; 1-x, 1)-\sum_{s=1}^{n}\frac{(-n)_s(\beta)_s}{(\gamma)_s s!}[\psi(n+1)-\psi(n+1-s)]x^s \\ &=(-1)^{n+1}n!\sum_{s=n+1}^{\infty}\frac{(s-n-1)!(\beta)_s}{(\gamma)_s s!}x^s, \qquad |x|<1, \\ &=\frac{(\beta)_{n+1}(-x)^{n+1}}{(\gamma)_{n+1}(n+1)}\,{}_3F_2(1, 1, \beta+n+1; n+2, \gamma+n+1; x), \qquad |\arg(1-x)|<\pi, \end{aligned} \tag{5.13}$$

which reduces to (5.11) if $\beta=\gamma$. An important case of (5.13) is

$$\begin{aligned} &L_n(1, \gamma-1; x, 1)-\sum_{s=1}^{n}\frac{(-n)_s}{(\gamma)_s}[\psi(n+1)-\psi(n+1-s)](1-x)^s \\ &=\frac{n!}{(\gamma)_{n+1}}(x-1)^{n+1}\,{}_2F_1(1, 1; \gamma+n+1; 1-x) \\ &=\frac{n!}{(\gamma)_{n+1}}(x-1)^{n+1}R_{-1}(1, \gamma+n; x, 1), \end{aligned} \tag{5.14}$$

where n is a nonnegative integer and $|\arg x|<\pi$. If γ is a positive integer, say $\gamma=1+m$, we can sum the ${}_2F_1$-series by (5.11) to get

$$\begin{aligned} L_n(1, m; x, 1)&=\sum_{s=1}^{n}\frac{(-n)_s}{(m+1)_s}[\psi(n+1)-\psi(n+1-s)](1-x)^s+\frac{n!m!}{(n+m)!}(x-1)^{-m} \\ &\cdot\left\{x^{n+m}\log x-\sum_{s=1}^{n+m}\frac{(-n-m)_s}{s!}[\psi(n+m+1)-\psi(n+m+1-s)](1-x)^s\right\}, \end{aligned} \tag{5.15}$$

where n and m are nonnegative integers, $|\arg x|<\pi$, and $x\neq 1$. If $n=0$ the first sum is empty and a change of index in the second sum gives, with the help of (2.5),

$$\begin{aligned} L_0(1, m; x, y)&=\log y+\left(\frac{x}{x-y}\right)^m\log\frac{x}{y} \\ &+\sum_{s=0}^{m-1}\binom{m}{s}[\psi(1+s)-\psi(1+m)]\left(\frac{y}{x-y}\right)^s, \end{aligned} \tag{5.16}$$

where m is a nonnegative integer, $x, y\in\mathbb{C}_0$, and $x\neq y$. This result generalizes (4.5) and will be used to prove (6.18).

The series in (5.12) terminates if β is a nonpositive integer, and the conditions of validity may then be relaxed:

$$L_t(-n, \gamma+n; x, 1)=\sum_{s=1}^{n}\frac{(-n)_s(-t)_s}{(\gamma)_s s!}[\psi(t+1)-\psi(t+1-s)](1-x)^s, \tag{5.17}$$

where x is unrestricted, n is a nonnegative integer, $(\gamma)_n \neq 0$, and $(-t)_n \neq 0$. The last condition is required because (5.4) was used to prove (5.12). This result will be used to prove (6.12).

6. Quadratic transformations. The R-function of two variables has exactly two independent quadratic transformations [3, (6.9-7), (6.10-1)],

$$R_{2t}(\beta, \beta; x, y) = R_t(\beta + t, 1/2 - t; A, G), \tag{6.1}$$

$$R_t(\beta, \beta; x^2, y^2) = R_t(2\beta + t, 1/2 - \beta - t; A, G), \tag{6.2}$$

where A and G denote the squared arithmetic and geometric means of x and y,

$$A = \left(\frac{x+y}{2}\right)^2, \qquad G = xy, \tag{6.3}$$

and where x and y have positive real parts and $\beta + 1/2 \neq 0, -1, -2, \cdots$. When the right-hand sides of (6.1) and (6.2) are differentiated with respect to t, one term arises from the subscript and a second term from the t-dependence of the b-parameters. Using (1.2) and (2.12) we find, with the same conditions of validity,

$$2L_{2t}(\beta, \beta; x, y) = L_t(\beta + t, 1/2 - t; A, G) - G^t L_{-t-\beta}(-t, \beta + 1/2 + t; A/G, 1), \tag{6.4}$$

$$L_t(\beta, \beta; x^2, y^2) = C + D, \tag{6.5}$$

where, for convenience in applications, we give several forms for C and D that are connected by (2.5), (2.11), (6.2), and [3, (5.9-19)]:

$$\begin{aligned} C &= L_t(2\beta + t, 1/2 - \beta - t; A, G) \\ &= \log(G) R_t(\beta, \beta; x^2, y^2) + G^t L_t(2\beta + t, 1/2 - \beta - t; A/G, 1) \\ &= \log(A) R_t(\beta, \beta; x^2, y^2) \\ &\quad - A^{1/2-\beta} G^{t+\beta-1/2} L_{-t-\beta-1/2}(1/2 - \beta - t, 2\beta + t; A/G, 1) \\ &= \log(AG) R_t(\beta, \beta; x^2, y^2) \\ &\quad - A^{1/2-\beta} G^{2t+2\beta} L_{-t-\beta-1/2}(1/2 - \beta - t, 2\beta + t; A, G), \end{aligned} \tag{6.6}$$

$$\begin{aligned} D &= -G^t L_{-t-2\beta}(-t, \beta + 1/2 + t; A/G, 1) \\ &= \log(G) R_t(\beta, \beta; x^2, y^2) - G^{2t+2\beta} L_{-t-2\beta}(-t, \beta + 1/2 + t; A, G) \\ &= -\log(A) R_t(\beta, \beta; x^2, y^2) + A^{1/2-\beta} L_{t+\beta-1/2}(\beta + 1/2 + t, -t; A, G) \\ &= -\log(A/G) R_t(\beta, \beta; x^2, y^2) \\ &\quad + A^{1/2-\beta} G^{t+\beta-1/2} L_{t+\beta-1/2}(\beta + 1/2 + t, -t; A/G, 1). \end{aligned} \tag{6.7}$$

If $t = 0$ the second term in both (6.4) and (6.5) vanishes by (2.3):

$$\begin{aligned} &2L_0(\beta, \beta; x, y) = L_0(\beta, 1/2; A, G), \\ &L_0(\beta, \beta; x^2, y^2) = L_0(2\beta, 1/2 - \beta; A, G), \\ &L_0(1/2, 1/2; x^2, y^2) = \log A = 2 \log \frac{x+y}{2}. \end{aligned} \tag{6.8}$$

Putting $t = -\beta$ and using the first form of C and the second form of D, we find

$$L_{-\beta}(\beta, \beta; x^2, y^2) = \log(xy) R_{-\beta}(\beta, \beta; x^2, y^2). \tag{6.9}$$

If $t=1/2-\beta$ the second term in the third form of C vanishes by (2.3), and the third form of D then gives

$$L_{1/2-\beta}(\beta,\beta;x^2,y^2)=\left(\frac{x+y}{2}\right)^{1-2\beta}L_0\left(1,\beta-1/2;\left(\frac{x+y}{2}\right)^2,xy\right). \tag{6.10}$$

From $L_{1/2-\beta}$ we can obtain $L_{-1/2-\beta}$ in terms of L_0 by (2.11) and [3, Ex. 6.10-12]. We shall use (6.10) to prove (6.18).

From the third form of C and the first form of D, we find

$$\begin{aligned}L_n(1/2+m,1/2+m;x^2,y^2)&=\log(A)R_n(1/2+m,1/2+m;x^2,y^2)\\&\quad-A^{-m}G^{n+m}L_{-n-m-1}(-n-m,n+2m+1;A/G,1)\\&\quad-G^nL_{-n-2m-1}(-n,n+m+1;A/G,1).\end{aligned} \tag{6.11}$$

If m and n are nonnegative integers, the last two terms (which are equal if $m=0$) have terminating series expansions (5.17). The result is

$$\begin{aligned}&L_n(1/2+m,1/2+m;x^2,y^2)\\&\quad=2\log\left(\frac{x+y}{2}\right)R_n(1/2+m,1/2+m;x^2,y^2)\\&\quad+\left(\frac{x+y}{2}\right)^{-2m}\sum_{s=1}^{n+m}\binom{n+m}{s}\frac{(n+m+1)_s}{(m+1)_s}\\&\quad\cdot[\psi(n+m+1+s)-\psi(n+m+1)]\left(\frac{x-y}{2}\right)^{2s}(xy)^{n+m-s}\\&\quad+\sum_{s=1}^{n}\binom{n}{s}\frac{(n+2m+1)_s}{(m+1)_s}[\psi(n+2m+1+s)-\psi(n+2m+1)]\left(\frac{x-y}{2}\right)^{2s}(xy)^{n-s}.\end{aligned} \tag{6.12}$$

By (6.2) and [3, (5.9-11)] the term in R_n can be omitted if, in the last sum, $2\log[(x+y)/2]$ is added to the terms in square brackets and the lower limit of summation is changed from 1 to 0. For example, the case $m=0$ is

$$\begin{aligned}&L_n(1/2,1/2;x^2,y^2)\\&\quad=2\sum_{s=0}^{n}\binom{n}{s}\binom{n+s}{s}\left[\psi(n+1+s)-\psi(n+1)+\log\frac{x+y}{2}\right]\left(\frac{x-y}{2}\right)^{2s}(xy)^{n-s},\end{aligned} \tag{6.13}$$

which generalizes the third equation in (6.8). This formula is useful for calculating the terms of the series (1.9). The polynomial $R_n(1/2,1/2;x^2,y^2)$ occurring in that series is the coefficient of $2\log[(x+y)/2]$ in (6.13).

Comparison of (6.12) and (1.7) gives

$$\begin{aligned}\left[\frac{\partial}{\partial\nu}P_\nu^{-m}(x)\right]_{\nu=n}&=\log\left(\frac{1+x}{2}\right)P_n^{-m}(x)\\&\quad+\left(\frac{1-x^2}{4}\right)^{m/2}\sum_{s=1}^{n-m}\frac{(m-n)_s(n+m+1)_s}{s!(m+s)!}\\&\quad\cdot[\psi(n+m+1+s)-\psi(n+m+1)]\left(\frac{1-x}{2}\right)^s\\&\quad+\left(\frac{1-x}{1+x}\right)^{m/2}\sum_{s=1}^{n}\frac{(-n)_s(n+1)_s}{s!(m+s)!}[\psi(n+1+s)-\psi(n+1)]\left(\frac{1-x}{2}\right)^s,\end{aligned} \tag{6.14}$$

where m and $n-m$ are nonnegative integers and $-1<x\leqq 1$. The logarithmic term can be incorporated in the second sum by adding $\log[(1+x)/2]$ to the terms in square brackets and beginning the sum at $s=0$. Equation (6.14) generalizes a formula due to Schelkunoff [13, (114)], [12, p. 173] for the case $m=0$, in which the two sums are equal. Schelkunoff used his formula in the theory of conical antennas. If m is an integer, differentiation of

$$\frac{P_\nu^m(x)}{\Gamma(\nu+m+1)}=(-1)^m\frac{P_\nu^{-m}(x)}{\Gamma(\nu-m+1)},\qquad -1\leqq x\leqq 1 \tag{6.15}$$

gives

$$\begin{aligned}\left[\frac{\partial}{\partial\nu}P_\nu^m(x)\right]_{\nu=n}&=(-1)^m\frac{(n+m)!}{(n-m)!}\left[\frac{\partial}{\partial\nu}P_\nu^{-m}(x)\right]_{\nu=n}\\&\quad+[\psi(n+m+1)-\psi(n-m+1)]P_n^m(x),\qquad n\geqq m.\end{aligned} \tag{6.16}$$

If m is even, there is no branch point at $x=1$, and the last three equations then hold for $|\arg(1+x)|<\pi$.

Equation (6.14) requires $n\geqq m$, but a known formula [9, 8.762(2)] for the case $n=0$, $m=1$ can be generalized by putting $\beta=1/2+m$ in (6.10) to get

$$L_{-m}(1/2+m,1/2+m;x^2,y^2)=\left(\frac{x+y}{2}\right)^{-2m}L_0\left(1,m;\left(\frac{x+y}{2}\right)^2,xy\right). \tag{6.17}$$

If m is a nonnegative integer, the L_0 function has a terminating series given by (5.16). The result is

$$\begin{aligned}&L_{-m}(1/2+m,1/2+m;x^2,y^2)\\&=\left(\frac{2}{x-y}\right)^{2m}\log\frac{(x+y)^2}{4xy}+\left(\frac{2}{x+y}\right)^{2m}\\&\quad\cdot\left\{\log(xy)+\sum_{s=0}^{m-1}\binom{m}{s}[\psi(1+s)-\psi(1+m)]\left(\frac{2}{x-y}\right)^{2s}(xy)^s\right\},\end{aligned} \tag{6.18}$$

where x and y have positive real parts, $x\neq y$, and $m=0,1,2,\cdots$. From L_{-m} we can obtain also L_{-m-1} by using (2.11) and [3, Ex. 6.10-12]. Comparison of (6.18) and (1.7) shows that

$$\begin{aligned}m!\left[\frac{\partial}{\partial\nu}P_\nu^{-m}(x)\right]_{\nu=0}&=(-1)^m\left(\frac{1+x}{1-x}\right)^{m/2}\log\frac{1+x}{2}\\&\quad+\left(\frac{1-x}{1+x}\right)^{m/2}\sum_{s=0}^{m-1}\binom{m}{s}[\psi(1+s)-\psi(1+m)]\left(\frac{2}{x-1}\right)^s,\end{aligned} \tag{6.19}$$

where $-1<x<1$ and $m=0,1,2,\cdots$. The limit as $x\to 1$ is 0. If m is even, the condition $-1<x<1$ can be replaced by $|\arg(x+1)|<\pi$ and $x\neq 1$.

The infinite series given in [9, (8.761)] can be reproduced by using (1.7), (6.5) with the third form of both C and D, and (5.12). The second equality in (5.4) is used also. It suffices to assume $\mu\neq -1,-2,-3,\cdots$ instead of $\operatorname{Re}\mu>-1$.

7. Inequalities. We assume t real and b_i and x_i strictly positive for $i=1,\cdots,k$. The largest and smallest of the x_i are denoted by $x_{\max}$ and $x_{\min}$, respectively, and we assume $x_{\max}>x_{\min}$ to exclude trivialities. Writing L_t for $L_t(b,x)$ and R_t for $R_t(b,x)$, we see at once from (1.1) and (1.3) that

$$R_t\log x_{\min}<L_t<R_t\log x_{\max}. \tag{7.1}$$

Chebyshev's inequality for integrals [10, Thm. 236] implies

$$L_t > R_t L_0, \qquad t > 0, \tag{7.2}$$

with reversed inequality for $t < 0$. Since $R_0 = 1$, (7.1) and (7.2) are subsumed in the following theorem, which is used in [6, (3.33)].

THEOREM 7.1. *Both L_t and L_t/R_t are strictly increasing functions of t. The limit of L_t/R_t is* $\log x_{\max}$ *as $t \to +\infty$ and* $\log x_{\min}$ *as $t \to -\infty$.*

Proof. Since the integrand of (1.1) is log-convex in t, R_t is strictly log-convex in t by [3, Ap. B.6] or [1, Thm. 4]. Hence its derivative and logarithmic derivative are strictly increasing, which proves the first part of the theorem. To prove the second part we use [1, Thm. 3]:

$$\lim_{t \to +\infty} (R_t)^{1/t} = x_{\max}, \tag{7.3}$$

which implies that $R_t \to \infty$ if $x_{\max} > 1$ and $R_t \to 0$ if $x_{\max} < 1$. If $x_{\max} = 1$ then $x_i < 1$ for some i (since $x_{\max} > x_{\min}$), and $u \cdot x$ in (1.1) is less than unity except on the set of measure zero where $u_i = 0$. Hence $(u \cdot x)^t \to 0$ almost everywhere as $t \to +\infty$, and so $R_t \to 0$. Thus $|\log R_t| \to \infty$ in all cases, and we may use L'Hôpital's rule to conclude from (7.3) that

$$\log x_{\max} = \lim_{t \to +\infty} \frac{\log R_t}{t} = \lim_{t \to +\infty} \frac{L_t}{R_t}.$$

The proof for $t \to -\infty$ is similar.

Remark. If the assumptions of Theorem 7.1 are changed so that exactly one of the variables, say x_k, is 0, then R_t and L_t are well defined for $t > -b_1 - \cdots - b_{k-1}$. In this region L_t/R_t increases strictly with t and has limit $\log x_{\max}$ as $t \to +\infty$. This follows from (4.2) and two well-known facts: $\psi''(x) < 0$ for $x > 0$, implying $\psi'(c - b_k + t) - \psi'(c + t) > 0$; and $\psi(x) = \log x + O(1/x)$ as $x \to +\infty$, implying $\psi(c - b_k + t) - \psi(c + t) \to 0$ as $t \to +\infty$.

THEOREM 7.2. *Let t be real and $x_{\max} > x_{\min} > 0$. Let $cw = (cw_1, \cdots, cw_k)$, where $c > 0$ and the w's are positive weights with $\sum w_i = 1$. Then the limit of $L_t(cw, x)$ is $\sum w_i x_i^t \log x_i$ as $c \to 0+$ and $(\sum w_i x_i)^t \log (\sum w_i x_i)$ as $c \to +\infty$. Also, $L_0(cw, x)$ is strictly increasing and strictly concave in c, and $L_1(cw, x)$ is strictly decreasing and strictly convex in c.*

Proof. We assume for the moment that $x_{\max} < 1$, so that the series expansions (5.6) and (5.7) converge. The proof of the first part of the theorem is entirely similar to the proof of [1, Thm. 1], in which L is the L_0 of the present paper. The second part of the theorem follows from (5.8), (5.9), and [7, Thm. 5]. The case $x_{\max} \geqq 1$ reduces to the case $x_{\max} < 1$ by use of (2.5) and [1, Thm. 1].

Remark. Example 5 in § 1 is contained in Theorem 7.2. The underlying reason why L_0 and L_1 are exceptionally simple is that $x^t \log x$ is concave in x for all $x > 0$ only if $t = 0$ and convex for all $x > 0$ only if $t = 1$.

8. Special cases. Several cases of L_t with restrictions on the parameters have been evaluated in finite terms in (4.6), (5.15), (5.16), (5.17), (6.12), (6.13) and (6.18). Further cases can be evaluated from these by (2.11) and the relations between associated functions in § 3. In (5.15) and (5.17) the apparent restriction of a unit variable can be removed by (2.5). We mention here some other cases, starting with

$$L_0(\beta, 1 - 2\beta; 1, 0) = \psi(\beta) - \psi(1 - \beta) = -\pi \cot (\pi\beta), \tag{8.1}$$

where β is not an integer and $\operatorname{Re}\beta > 0$, and

$$(8.2)\qquad \begin{aligned} L_{2t}(\beta, \beta; x, -x) &= \frac{\pi^{1/2}\Gamma(\beta+1/2)}{2\Gamma(1/2-t)\Gamma(\beta+1/2+t)}(-x^2)^t \\ &\quad\cdot [\psi(1/2-t) - \psi(\beta+1/2+t) + \log(-x^2)], \end{aligned}$$

where $|\arg(\pm x)| < \pi$, $|\arg(-x^2)| < \pi$, and $\beta + 1/2 \neq 0, -1, -2, \cdots$. Equation (8.1) comes from (1.5) or (4.5), and (8.2) comes from differentiating the formula in [3, Ex. 8.3-9].

To evaluate an R-function or L-function of two variables with positive integral b-parameters, we raise the b-parameters from unity by successive differentiations using [3, (5.9-9)]:

$$(8.3)\qquad \begin{aligned} R_{t-1}(m, m'; x, y) &= \frac{(m+m'-1)!}{(m-1)!(m'-1)!(t)_{m+m'-2}} \\ &\quad\cdot D_x^{m-1} D_y^{m'-1} R_{t+m+m'-3}(1, 1; x, y). \end{aligned}$$

Use of [3, Ex. 5.9-13] and two applications of Leibnitz' rule lead to

$$(8.4)\qquad \begin{aligned} R_{t-1}(m, m'; x, y) &= (m)_{m'} x^{t-1} \sum_{s=0}^{m-1} \frac{(1-m)_s (m')_s}{(t)_{m'+s} s!} \left(\frac{x}{x-y}\right)^{m'+s} \\ &\quad + (m')_m y^{t-1} \sum_{s=0}^{m'-1} \frac{(1-m')_s (m)_s}{(t)_{m+s} s!} \left(\frac{y}{y-x}\right)^{m+s}, \end{aligned}$$

where m and m' are positive integers, $x, y \in \mathbb{C}_0$, $x \neq y$, and $(t)_{m+m'-1} \neq 0$. With the same conditions, differentiation with respect to t gives

$$(8.5)\qquad \begin{aligned} L_{t-1}(m, m'; x, y) &= (m)_{m'} x^{t-1} \sum_{s=0}^{m-1} \frac{(1-m)_s (m')_s}{(t)_{m'+s} s!} \\ &\quad\cdot [\psi(t) - \psi(t+m'+s) + \log x]\left(\frac{x}{x-y}\right)^{m'+s} \\ &\quad + (m')_m y^{t-1} \sum_{s=0}^{m'-1} \frac{(1-m')_s (m)_s}{(t)_{m+s} s!} \\ &\quad\cdot [\psi(t) - \psi(t+m+s) + \log y]\left(\frac{y}{y-x}\right)^{m+s}. \end{aligned}$$

Agreement with the special case (5.16) can be shown by a binomial expansion of $[1 + y/(x-y)]^m$.

Putting $n = 0$ and $\gamma = 1$ in (5.13), dividing by β, and letting $\beta \to 0$, we find a relation between L_0 and Euler's dilogarithm:

$$(8.6)\qquad -\lim_{\beta \to 0} \beta^{-1} L_0(\beta, 1-\beta; 1-x, 1) = x\,{}_3F_2(1, 1, 1; 2, 2; x) = \sum_{n=1}^{\infty} \frac{x^n}{n^2}.$$

By (2.3) the limit of the same quantity as β tends to 1 is

$$(8.7)\qquad -\lim_{\beta \to 1} \beta^{-1} L_0(\beta, 1-\beta; 1-x, 1) = -\log(1-x) = \sum_{n=1}^{\infty} \frac{x^n}{n}.$$

The conditions of validity for (8.5) exclude the function

$$(8.8)\qquad L_{-1}(1, 1; x, y) = \frac{(\log x)^2 - (\log y)^2}{2(x-y)}, \qquad x \neq y,$$

which is evaluated by putting $f(x)=(\log x)^2$ in [3, (5.5-3)]. Similarly putting $f(x)=x^t$ in [3, Ex. 5.5-1], we find

$$R_{t-2}(1,1,1;x,y,z)=\frac{2}{t(t-1)}\sum_{\text{cyc}}\frac{x^t}{(x-y)(x-z)}, \tag{8.9}$$

where $t\neq 0$, 1 and the sum extends over cyclic permutations of the distinct complex numbers x, y, z. We shall refrain from raising the b-parameters by differentiation with respect to x, y, z as in (8.3). Differentiation with respect to t gives

$$L_{t-2}(1,1,1;x,y,z)=\frac{2}{t(t-1)}\sum_{\text{cyc}}\frac{x^t}{(x-y)(x-z)}\left(\log x-\frac{1}{t}-\frac{1}{t-1}\right), \tag{8.10}$$

with the same conditions of validity as for (8.9). This is the function that occurs in (1.8). The exceptional cases with $t=0$ or 1 are evaluated by successively putting $f(x)=\log x$, $(\log x)^2$, $x\log x$, and $x(\log x)^2$ in [3, Ex. 5.5-1]:

$$R_{-2}(1,1,1;x,y,z)=-2\sum_{\text{cyc}}\frac{\log x}{(x-y)(x-z)}, \tag{8.11}$$

$$L_{-2}(1,1,1;x,y,z)=-\sum_{\text{cyc}}\frac{(\log x)^2+2\log x}{(x-y)(x-z)}, \tag{8.12}$$

$$R_{-1}(1,1,1;x,y,z)=2\sum_{\text{cyc}}\frac{x\log x}{(x-y)(x-z)}, \tag{8.13}$$

$$L_{-1}(1,1,1;x,y,z)=\sum_{\text{cyc}}\frac{x[(\log x)^2-2\log x]}{(x-y)(x-z)}. \tag{8.14}$$

Values for special cases of the L-function permit evaluation of some special cases of ${}_3F_2$ by using (5.13). For example,

$${}_3F_2(1,1,m+n+3/2;n+2,2m+n+2;x)$$

can be found from (5.13) and (6.12) if m and n are nonnegative integers. The case $n=0$ is

$$\begin{aligned}{}_3F_2(1,1,m+3/2;2,2m+2;x)&=(-4/x)\log y-(2/x)y^{-2m}\sum_{s=1}^{m}\binom{m}{s}\\&\quad\cdot[\psi(m+1+s)-\psi(m+1)](y-1)^{2s}(1-x)^{(m-s)/2},\end{aligned} \tag{8.15}$$

where $2y=1+(1-x)^{1/2}$, $|\arg(1-x)|<\pi$, and $m=0,1,2,\cdots$.

REFERENCES

[1] B. C. Carlson, *A hypergeometric mean value*, Proc. Amer. Math. Soc., 16 (1965), pp. 759-766.

[2] ———, *Invariance of an integral average of a logarithm*, Amer. Math. Monthly, 82 (1975), pp. 379-382.

[3] ———, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[4] ———, *Computing elliptic integrals by duplication*, Numer. Math., 33 (1979), pp. 1-16.

[5] ———, *The hypergeometric function and the R-function near their branch points*, Rend. Sem. Mat. Univ. Politec. Torino (1985), Fascicolo speciale, pp. 63-89.

[6] B. C. Carlson and J. L. Gustafson, *Asymptotic expansion of the first elliptic integral*, this Journal, 16 (1985), pp. 1072-1092.

[7] B. C. Carlson and M. D. Tobey, *A property of the hypergeometric mean value*, Proc. Amer. Math. Soc., 19 (1968), pp. 255-262.

[8] M. L. Glasser and J. I. Kaplan, *The surface of a neutron star in superstrong magnetic fields*, Astrophys. J., 199 (1975), pp. 208-219.

[9] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[10] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge Univ. Press, Cambridge, 1959.

[11] G. J. JONES AND R. K. TRIVEDI, *Lateral growth in solid-solid phase transitions*, J. Appl. Phys., 42 (1971), pp. 4299-4304.

[12] L. ROBIN, *Fonctions sphériques de Legendre et fonctions sphéroïdales*, Vol. 2, Gauthier-Villars, Paris, 1958.

[13] S. A. SCHELKUNOFF, *Theory of antennas of arbitrary size and shape*, Proc. IEEE, 29 (1941), pp. 493-521.

[14] A. SEEGER, *Diffusion problems associated with the growth of crystals from dilute solution*, Phil. Mag., (7), 44 (1953), pp. 1-13.

SIAM J. MATH. ANAL.
Vol. 18, No. 2, March 1987

SCHUR-OSTROWSKI THEOREMS FOR FUNCTIONALS ON $L_1(0, 1)$*

WAI CHAN†, FRANK PROSCHAN†‡ AND JAYARAM SETHURAMAN†§

Abstract. Hardy, Littlewood and Pólya [5] introduced the partial ordering of majorization among n-dimensional real vectors. Many well-known inequalities can be recast as the statement that certain functions are increasing with respect to this ordering. Such functions are said to be Schur-convex. An important result in the theory of majorization is the Schur-Ostrowski theorem, which characterizes Schur-convex functions. The concept of majorization has been extended to elements of $L_1(0, 1)$ by Ryff [10]. A functional on $L_1(0, 1)$ that is increasing with respect to the ordering of majorization is said to be Schur-convex. In this paper, we prove an analogue of the Schur-Ostrowski condition that characterizes Schur-convex functionals in terms of their Gâteaux differentials. We also introduce another partial ordering in $L_1(0, 1)$ called unrestricted majorization. This partial ordering is similar to majorization but does not involve the use of decreasing rearrangements. We establish a characterization of nondecreasing functionals on $L_1(0, 1)$ with respect to the partial ordering of unrestricted majorization through another analogue of the Schur-Ostrowski condition.

Key words. inequalities, majorization, Muirhead's theorem, peakedness in symmetric distributions, rearrangement, Schur functions, Schur-Ostrowski's theorem

AMS(MOS) subject classifications. 26D10, 60E15

1. Introduction. Hardy, Littlewood and Pólya [5] introduced the following partial order in n-dimensional Euclidean spaces: an n-vector $\mathbf{x}=(x_1, \cdots, x_n)$ majorizes $\mathbf{y}=(y_1, \cdots, y_n)$, ($\mathbf{x} \geqq^m \mathbf{y}$ in symbols), whenever

$$\sum_1^k x_i^* \geqq \sum_1^k y_i^*, \qquad k=1, \cdots, n-1$$

and

$$\sum_1^n x_i = \sum_1^n y_i,$$

where $\mathbf{x}^*$, $\mathbf{y}^*$ are the vectors obtained from $\mathbf{x}$ and $\mathbf{y}$ by rearranging their components in decreasing order.

This partial order has been extended to elements of $L_1(0, 1)$ by Ryff [10] and is given in Definition 1.2 below. Before giving this definition, we develop some notation to be used in defining a decreasing rearrangement of a function. Let x be a measurable, real valued function on (0, 1) and m be the Lebesgue measure. For each x, one can associate a function d_x on $(-\infty, \infty)$ defined by

$$d_x(s) = m(\{t: x(t) > s\}), \qquad -\infty < s < \infty.$$

This function d_x, called the distribution function of x, is nonincreasing and right continuous. Two functions x and y are said to be equivalent in distribution if $d_x = d_y$. The right continuous inverse of d_x, denoted x^*, is defined by

$$x^*(t) = \inf \{s: d_x(s) \leqq t\}.$$

* Received by the editors May 28, 1985; accepted for publication (in revised form) November 25, 1985.

† The Florida State University, Department of Statistics, Tallahassee, Florida 32306.

‡ The work of this author was supported by the Air Force Office of Scientific Research, System Command, United States Air Force, under grant number AFOSR 82-K-007.

§ The work of this author was supported by the U.S. Army Research Office under grant DAAG 29-82-K-0168.

The function x^*, which is nonincreasing, right continuous and has the same distribution function as x, is called the decreasing rearrangement of x. The functions x and x^* are simultaneously integrable (or nonintegrable), and their integrals are related by

$$\int_0^s x^*(t)\,dt \geqq \int_0^s x(t)\,dt, \qquad 0 \leqq s < 1$$

and

$$\int_0^1 x^*(t)\,dt = \int_0^1 x(t)\,dt.$$

The following theorem due to Ryff [12] shows that, by composing the decreasing rearrangement of a function with a measure preserving transformation, one can recover the original function.

THEOREM 1.1. *To each* $x \in L_1(0, 1)$, *there corresponds a measure preserving transformation* $\sigma: (0, 1) \to (0, 1)$ *such that* $x(t) = x^*[\sigma(t)]$, *where* σ *is defined by*

$$\sigma(s) = m\{t: x(t) > x(s)\} + m\{t \leqq s: x(t) = x(s)\}.$$

The definition of the partial ordering of majorization of elements in $L_1(0, 1)$, due to Ryff [10], is given below.

DEFINITION 1.2. Let $x, y \in L_1(0, 1)$. We say that x majorizes y, ($x \geqq^m y$ in symbols) if

$$\int_0^s x^*(t)\,dt \geqq \int_0^s y^*(t)\,dt, \qquad 0 \leqq s < 1,$$

and

$$\int_0^1 x(t)\,dt = \int_0^1 y(t)\,dt,$$

where x^* and y^* are the decreasing rearrangements of x and y, respectively.

Several authors (see, e.g., Day [4], Chong [3]) have obtained interesting results using this partial ordering. It is also related to the variability ordering of Ross [9].

By removing the rearrangement requirement in Definition 1.2, we obtain a different ordering called unrestricted majorization, as defined below.

DEFINITION 1.3. Let $x, y \in L_1(0, 1)$. We say that x dominates y in the ordering of unrestricted majorization, ($x \geqq^u y$ in symbols), if

$$\int_0^s x(t)\,dt \geqq \int_0^s y(t)\,dt, \qquad 0 \leqq s < 1,$$

and

$$\int_0^1 x(t)\,dt = \int_0^1 y(t)\,dt.$$

The ordering of unrestricted majorization as applied to the class of density functions leads to the usual stochastic ordering as seen below.

Let X and Y be random variables on $(0, 1)$ with densities f and g, respectively. If $f \geqq^u g$, then $\int_0^s f \geqq \int_0^s g$ for all $0 < s < 1$, or $P(X \leqq s) \geqq P(Y \leqq s)$. Thus the condition $X \leqq^{st} Y$ is equivalent to $f \geqq^u g$.

Many inequalities that arise from majorization in the finite dimensional case can be extended for elements of $L_p(0, 1)$. Ryff [11] proved the following analogue of Muirhead's inequality.

THEOREM 1.4. *Let x and y be bounded measurable functions on* $(0, 1)$. *If $x \geqq^m y$ and u is a positive function such that $u \in L_p(0, 1)$ for all p, $-\infty < p < \infty$, then*

$$\int_0^1 \log\left[\int_0^1 u(t)^{x(s)}\,dt\right] ds \geqq \int_0^1 \log\left[\int_0^1 u(t)^{y(s)}\,dt\right] ds.$$

Conversely, if the inequality holds for all such u, then $x \geqq^m y$.

In the discrete case, Muirhead's inequality can be reformulated by identifying an appropriate function which preserves the ordering of majorization. Such functions are said to be Schur-convex. Schur [13] and Ostrowski [7] gave necessary and sufficient conditions for a function to be Schur-convex in terms of their partial derivatives. We quote from Marshall and Olkin [6] about the importance of this result, "it is difficult to overemphasize the usefulness of the (Schur–Ostrowski) condition . . . , many or even most of the theorems giving Schur-convexity were first discovered by checking (the Schur–Ostrowski condition)." In the next section, we will present an analogue of this result for Schur-convex functionals on $L_\infty(0, 1)$. This result, given in Theorem 2.9, is then used to characterize Schur-convex functionals on $L_1(0, 1)$. We also characterize nondecreasing functionals on $L_1(0, 1)$ with respect to the partial ordering of unrestricted majorization through another analogue of the Schur–Ostrowski condition. These results will be used to prove the generalized Muirhead's theorem (Proschan and Sethuraman [8]) in § 3. An application to peakedness comparisons of distributions is discussed in § 3.

2. Main theorems. We first proceed with some definitions.

DEFINITION 2.1. A functional ϕ defined on a set $\mathscr{A} \subseteqq L_1(0, 1)$ is said to be Schur-convex on $\mathscr{A}$ if $y_1, y_2 \in \mathscr{A}$ and $y_1 \geqq^m y_2$ imply that $\phi(y_1) \geqq \phi(y_2)$.

A Schur-convex functional is necessarily constant over functions that are equivalent in distribution. Thus for a Schur-convex functional ϕ, the value $\phi(x)$ depends only on the distribution function of x. A set $\mathscr{A}$ is said to be invariant if $x \in \mathscr{A}$ and x and y are equivalent in distribution implying that $y \in \mathscr{A}$. Henceforth, we shall only consider Schur-convex functionals on an invariant set.

For a characterization of Schur-convex functionals, we need the following notion of directional derivative.

DEFINITION 2.2. Let ϕ be a functional defined on a convex set $\mathscr{A} \subseteqq L_1(0, 1)$. Let $y \in \mathscr{A}$ and h be such that $y + \theta h \in \mathscr{A}$ for all sufficiently small θ. The Gâteaux differential of ϕ at y in the direction of h is defined to be

$$\frac{\partial\phi}{\partial h}(y) = \lim_{\theta\to 0} \frac{\phi(y+\theta h) - \phi(y)}{\theta}$$

if the limit exists.

Note that $\partial\phi/\partial h(y)$ is simply the derivative, at $\theta = 0$, of the real valued function on $[0, 1]$ defined by $\psi(\theta) = \phi(y + \theta h)$.

Let $\mathscr{D}_1$ be the class of decreasing functions in $L_1(0, 1)$, let $\mathscr{D}_\infty$ be the class of decreasing functions in $L_\infty(0, 1)$. Let $\mathscr{T} = \{h\colon h = \lambda_1 I_{(a,b)} + \lambda_2 I_{(c,d)}$, where $0 \leqq a < b < c < d \leqq 1$, $\lambda_1 \geqq 0 \geqq \lambda_2$, $\lambda_1(b-a) + \lambda_2(d-c) = 0\}$. The class $\mathscr{T}$ consists of step functions h which take at most two nonzero values, are decreasing on their support and satisfy $\int_0^1 h(t)\,dt = 0$. Note that $h \in \mathscr{T}$ implies $h \geqq^m 0$.

Let $y \in \mathscr{D}_1$ and $h \in \mathscr{T}$. Then $y + h$ need not be decreasing. However, we have $y + h \geqq^m y$, as given in the next lemma.

LEMMA 2.3. *Let $y \in \mathscr{D}_1$ and $h \in \mathscr{T}$; then $y + h \geqq^m y$.*

Proof. Note that

$$\begin{aligned}\int_0^s (y+h)^* &\geqq \int_0^s (y+h)\\ &= \int_0^s y^* + \int_0^s h\\ &\geqq \int_0^s y^*, \qquad 0 \leqq s < 1\end{aligned}$$

and

$$\int_0^1 (y+h) = \int_0^1 y + \int_0^1 h = \int_0^1 y.$$

Hence, $y+h \geqq^m y$. □

In the following theorem, we give a necessary condition for functionals increasing in the ordering of unrestricted majorization.

THEOREM 2.4. *Let $\mathscr{A}$ be an open subset of $L_1(0,1)$. Let ϕ be a functional defined on $\mathscr{A}$ such that ϕ is nondecreasing with respect to the ordering of unrestricted majorization. Let $y \in \mathscr{A}$ and $h \in \mathscr{T}$. Suppose that the Gâteaux differential $\partial\phi/\partial h(y)$ exists. Then $\partial\phi/\partial h(y) \geqq 0$.*

Proof. Since $\mathscr{A}$ is open, $y+\theta h \in \mathscr{A}$ for all sufficiently small θ. Thus for all sufficiently small positive θ, $y+\theta h$ and y are elements of $\mathscr{A}$ and $y+\theta h \geqq^u y$. This implies that

$$\phi(y+\theta h) \geqq \phi(y)$$

and

$$\begin{aligned}\frac{\partial\phi}{\partial h}(y) &= \lim_{\theta\downarrow 0} \frac{1}{\theta}[\phi(y+\theta h) - \phi(y)]\\ &\geqq 0.\end{aligned}$$ □

Next, we consider Schur-convex functionals defined on an invariant set $\mathscr{A}$.

THEOREM 2.5. *Let $\mathscr{A}$ be an open invariant subset of $L_1(0,1)$. Let ϕ be a Schur-convex functional defined on $\mathscr{A}$. Let $y \in \mathscr{D}_\infty \cap \mathscr{A}$ and $h \in \mathscr{T}$. Suppose that the Gâteaux differential $\partial\phi/\partial h(y)$ exists. Then $\partial\phi/\partial h(y) \geqq 0$.*

Proof. Since $\mathscr{A}$ is open, $y+\theta h \in \mathscr{A}$ for all sufficiently small θ. Furthermore, for sufficiently small positive θ, $y+\theta h \geqq^m y$ from Lemma 2.3. Hence $\phi(y+\theta h) \geqq \phi(y)$ and

$$\begin{aligned}\frac{\partial\phi}{\partial h}(y) &= \lim_{\theta\downarrow 0} \frac{1}{\theta}[\phi(y+\theta h) - \phi(y)]\\ &\geqq 0.\end{aligned}$$ □

To show that this condition is also sufficient, which is the content of the main theorem, Theorem 2.10 of this section, we need the following lemmas.

LEMMA 2.6. *Let $\mathscr{A}$ be a convex subset of $L_1(0,1)$. Let ϕ be a functional defined on an open set containing $\mathscr{A}$. Let $\partial\phi/\partial h(y) \geqq 0$ for $y \in \mathscr{A}$ and $h \in \mathscr{T}$. Then $y_1, y_2 \in \mathscr{A}$ and $y_2 - y_1 \in \mathscr{T}$ imply that $\phi(y_2) \geqq \phi(y_1)$.*

Proof. Let $h = y_2 - y_1 \in \mathscr{T}$. For $\theta \in [0,1]$, define

$$y_\theta = y_1 + \theta(y_2 - y_1) = \theta y_2 + (1-\theta) y_1$$

and

$$\psi(\theta)=\phi(y_\theta).$$

Note that y_θ is in $\mathscr{A}$. Now,

$$\begin{aligned}\frac{d}{d\theta}\psi(\theta)&=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[\psi(\theta+\varepsilon)-\psi(\theta)]\\&=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[\phi(y_\theta+\varepsilon h)-\phi(y_\theta)]\\&=\frac{\partial\phi}{\partial h}(y_\theta)\geqq 0\quad\text{for } 0\leqq\theta\leqq 1.\end{aligned}$$

Hence,

$$\begin{aligned}\phi(y_2)-\phi(y_1)&\geqq\int_0^1\frac{d}{d\theta}\psi(\theta)\,d\theta\\&=\int_0^1\frac{\partial\phi}{\partial h}(y_\theta)\,d\theta\geqq 0.\end{aligned}$$ □

In the next lemma, we show that if y_1, y_2 are step functions such that $y_2\geqq^u y_1$, then y_2-y_1 can be written as the sum of functions in $\mathscr{T}$.

LEMMA 2.7. *Let y_1, y_2 be step functions on $(0,1)$ such that $y_2\geqq^u y_1$. Then there exist $h_1,\cdots,h_N$ in $\mathscr{T}$ such that*

$$y_2=y_1+\sum_{i=1}^N h_i. \tag{2.1}$$

Proof. There is nothing to prove if $y_1=y_2$.

Let $y_1\neq y_2$. Since y_1 and y_2 are step functions, there is an integer $n\geqq 2$, such that

$$y_2(t)-y_1(t)=\sum_{i=1}^n a_i I_{(c_i,d_i)}(t)\quad\text{where } a_i\neq 0,$$

(c_i,d_i) are disjoint intervals and $0\leqq c_1<d_1<\cdots<c_n<d_n\leqq 1$. Note that $y_2\geqq^u y_1$ implies that $a_1>0$ and $a_n<0$. We will prove that (2.1) holds with $N\leqq n-1$, by an induction on n.

Note that the lemma is immediate when $n=2$. Assume that the lemma is true for $n=2,\cdots,k-1$. We will now prove that the lemma holds for $n=k$. Let a_j be the first negative term such that either $a_{j+1}>0$ or $j=k$. Define a function in $\mathscr{T}$ by

$$h=a_1 I_{(c_1,c_1')}+a_j I_{(d_j',d_j)},$$

where $c_1'\leqq d_1$ and $d_j'\geqq c_j$ are chosen so that $a_1(c_1'-c_1)+a_j(d_j=d_j')=0$ and one of the following holds:

1) $c_1'=d_1$ and $d_j'=c_j$ if $a_1(d_1-c_1)+a_j(d_j-c_j)=0$,
2) $c_1'<d_1$ and $d_j'=c_j$ if $a_1(d_1-c_1)+a_j(d_j-c_j)>0$,
3) $c_1'=d_1$ and $d_j'>c_j$ if $a_1(d_1-c_1)+a_j(d_j-c_j)<0$.

We will now establish that $y_2\geqq^u y_1+h$ by showing that $\int_0^s(y_2-y_1-h)\geqq 0$ for all $0<s\leqq 1$. Note that $h=0$ on the interval $(d_j,1)$.

Let $s>d_j$, then $\int_0^s h=\int_0^1 h=0$. Thus,

$$\int_0^s(y_2-y_1-h)=\int_0^s(y_2-y_1)\geqq 0.$$

Let $0 < s \leqq d_j$. Then either $y_2(t) - y_1(t) \geqq 0$ for all $0 < t < s$ or $y_2(t) - y_1(t) \leqq 0$ for all $s < t \leqq d_j$, since there is only one sign change among $a_1, \cdots, a_j$ and the sign changes from positive to negative. Note that h agrees with $y_2 - y_1$ on the intervals (c_1, c_1') and (d_j', d_j), and that h is identically zero outside these intervals.

If $y_2(t) - y_1(t) \geqq 0$ for all $0 < t < s$, then $y_2 - y_1 \geqq h \geqq 0$ on the interval $(0, s)$. This implies that

$$\int_0^s (y_2 - y_1 - h) \geqq 0.$$

If $y_2(t) - y_1(t) \leqq 0$ for all $s < t < d_j$, then $y_2 - y_1 \leqq h \leqq 0$ on the interval (s, d_j). This implies that

$$\begin{aligned}\int_0^s (y_2 - y_1 - h) &\leqq \int_0^{d_j} (y_2 - y_1 - h)\\ &= \int_0^{d_j} (y_2 - y_1)\\ &\geqq 0.\end{aligned}$$

Hence we have $y_2 \geqq^u y_1 + h$. Since $y_2 - y_1 - h$ is a step function which takes at most $k-1$ nonzero values, it follows from the induction hypothesis that $y_2 - y_1 - h = \sum_{i=1}^N h_i'$ where $h_i' \in \mathcal{T}$ for $i = 1, \cdots, N$, and $N \leqq k-2$. This completes the proof. □

In Lemma 2.7, if we assume that y_1, y_2 are decreasing step functions, then the condition $y_2 \geqq^u y_1$ is equivalent to $y_2 \geqq^m y_1$. In addition, we can choose $y_1 + h_1, y_1 + h_1 + h_2, \cdots, y_1 + \sum_{i=1}^{N-1} h_i$ to be decreasing functions as shown in the following lemma.

LEMMA 2.8. *Let y_1, y_2 be decreasing step functions on $(0, 1)$ such that $y_2 \geqq^m y_1$. Then there exist $h_1, \cdots, h_N$ in $\mathcal{T}$ such that*

(i) $y_2 = y_1 + \sum_{i=1}^{N} h_i$, *and*

(ii) $y_1 + h_1, \cdots, y_1 + \sum_{i=1}^{N-1} h_i$ *are decreasing functions.*

Proof. Define $h = a_1 I_{(c_1, c_1')} + a_j I_{(d_j', d_j)}$ as in the proof of Lemma 2.7. We need to show that $y_1 + h$ is decreasing. Note that

$$y_1(t) + h(t) = \begin{cases} y_2(t) & \text{if } 0 < t < c_1', \\ y_1(t) & \text{if } c_1' \leqq t \leqq d_j', \\ y_2(t) & \text{if } d_j' < t \leqq d_j, \\ y_1(t) & \text{if } d_j \leqq t < 1. \end{cases}$$

Since $a_1 > 0$, $y_1 + h$ is decreasing on a neighborhood of c_1'. Similarly, $a_j < 0$ implies that $y_1 + h$ is decreasing on a neighborhood of d_j'. Suppose that $d_j < 1$, then the choice of a_j indicates that for $\varepsilon > 0$ sufficiently small, $y_2 - y_1 \geqq 0$ on $(d_j, d_j + \varepsilon)$. Since $y_1 + h = y_2$ on $(d_j', d_j]$, it follows that $y_1 + h$ is decreasing on the interval $(d_j', d_j + \varepsilon)$. Thus $y_1 + h$ is decreasing on the interval $(0, 1)$.

Note that $h \in \mathcal{T}$ implies $y_1 + h \geqq^u y_1$. Since $y_1 + h$, y_1 are decreasing functions, this is equivalent to $y_1 + h \geqq^m y_1$. Following the same induction argument as in Lemma 2.7, we conclude that there exist $h_1, \cdots, h_N$ in $\mathcal{T}$ such that $y_2 = y_1 + h + \sum_{i=1}^N h_i$ and that $y_1 + h + h_1, \cdots, y_1 + h + \sum_{i=1}^{N-1} h_i$ are decreasing functions. This proves the lemma. □

In the next theorem, we give a sufficient condition for a functional of $L_\infty(0,1)$ to be Schur-convex.

THEOREM 2.9. *Let $\mathcal{A}$ be an invariant open convex subset of $L_\infty(0,1)$. Let ϕ be a continuous functional defined on $\mathcal{A}$ such that ϕ is constant over functions that are equivalent in distribution. If the Gâteaux differential $\partial\phi/\partial h(y) \geqq 0$ for each $y \in \mathcal{D}_\infty \cap \mathcal{A}$ and $h \in \mathcal{T}$, then ϕ is Schur-convex on $\mathcal{A}$.*

Proof. Since ϕ is constant over functions that are equivalent in distribution, it suffices to prove that ϕ is Schur-convex on $\mathcal{D}_\infty \cap \mathcal{A}$.

Let $y_1, y_2 \in \mathcal{D}_\infty \cap \mathcal{A}$ be right continuous and $y_2 \geqq^m y_1$. Let $\varepsilon > 0$ be arbitrary. Then for $i = 1, 2$, the sets $\{t: y_i(t^-) - y_i(t) > \varepsilon\}$ are finite, where $y_i(t_0^-) = \inf_{t<t_0} y_i(t)$. Hence there exists a partition $0 < a_1 < \cdots < a_n < 1$ such that

$$y_i(a_k) - y_i(a_{k+1}^-) < \varepsilon, \quad i = 1, 2; \quad k = 1, \cdots, n-1.$$

Define

$$y_{i\varepsilon}(t) = \frac{1}{a_1}\left[\int_0^{a_1} y_i(s)\,ds\right] I_{(0,a_1)}(t) + \sum_{k=1}^{n-1} \frac{1}{a_{k+1}-a_k}\left[\int_{a_k}^{a_{k+1}} y_i(s)\,ds\right] I_{[a_k,a_{k+1})}(t)$$
$$+ \frac{1}{1-a_n}\left[\int_{a_n}^{1} y_i(s)\,ds\right] I_{[a_n,1)}(t), \qquad i = 1, 2.$$

Then $y_{1\varepsilon}$, $y_{2\varepsilon}$ are decreasing step functions satisfying $\int_0^{a_k} y_{i\varepsilon}(s)\,ds = \int_0^{a_k} y_i(s)\,ds$ for $k = 1, \cdots, n$. This implies $y_{2\varepsilon} \geqq^m y_{1\varepsilon}$. Since $\mathcal{A}$ is open and $\|y_i - y_{i\varepsilon}\|_\infty < \varepsilon$, for sufficiently small positive ε, $y_{1\varepsilon}$, $y_{2\varepsilon}$ are in $\mathcal{A}$.

By Lemma 2.8, $y_{2\varepsilon} - y_{1\varepsilon} = \sum_{i=1}^N h_i$ for some $\{h_1, \cdots, h_N\} \subseteq \mathcal{T}$, where $y_{1\varepsilon} + h_1, \cdots, y_{1\varepsilon} + \sum_{i=1}^{N-1} h_i$ are decreasing functions. The functions $y_{1\varepsilon} + h_1, \cdots, y_{1\varepsilon} + \sum_{i=1}^{N-1} h_i$ need not be elements of $\mathcal{A}$. Since $\mathcal{A}$ is open, for sufficiently small positive θ, $y_{1\varepsilon} + \theta h_1, \ldots, y_{1\varepsilon} + \theta \sum_{i=1}^N h_i$ are decreasing functions in $\mathcal{A}$ satisfying

$$y_{1\varepsilon} + \theta \sum_{i=1}^N h_i \geqq^m y_{1\varepsilon} + \theta \sum_{i=1}^{N-1} h_i \geqq^m \cdots \geqq^m y_{1\varepsilon}.$$

It now follows from Lemma 2.6 that

$$\phi\left(y_{1\varepsilon} + \theta \sum_{i=1}^N h_i\right) \geqq \phi\left(y_{1\varepsilon} + \theta \sum_{i=1}^{N-1} h_i\right) \geqq \cdots \geqq \phi(y_{1\varepsilon}).$$

Next, we shall show that this implies $\phi(y_{2\varepsilon}) \geqq \phi(y_{1\varepsilon})$. Note that we have just demonstrated that the set

$$\Theta = \left\{0 \leqq \theta \leqq 1: \phi\left(y_{1\varepsilon} + \theta \sum_{i=1}^N h_i\right) \geqq \phi(y_{1\varepsilon})\right\}$$

is nonempty. Let $\theta_0 = \sup\{\theta: \theta \in \Theta\}$. Since ϕ is continuous, we have $\phi(y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_i) \geqq \phi(y_{1\varepsilon})$, which shows that $\theta_0 \in \Theta$. Now suppose $\theta_0 < 1$. The preceding arguments show that for sufficiently small positive r,

$$y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_i + rh_1, \cdots, y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_1 + r \sum_{i=1}^N h_i$$

are decreasing functions in $\mathcal{A}$ and satisfy

$$\phi\left(y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_i\right) \leqq \phi\left(y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_i + rh_1\right) \leqq \cdots \leqq \phi\left(y_{1\varepsilon} + \theta_0 \sum_{i=1}^N h_i + r \sum_{i=1}^N h_i\right).$$

Thus $\theta_0 + r \in \Theta$, which provides a contradiction to the assumption that $\theta_0 < 1$. We therefore conclude that $\theta_0 = 1$ and thus, $\phi(y_{2\varepsilon}) \geqq \phi(y_{1\varepsilon})$.

Since the functional is continuous with respect to L_∞-norm, we conclude that $\phi(y_2) \geqq \phi(y_1)$ by letting $\varepsilon \to 0$. □

We now use this theorem to establish a sufficient condition for Schur-convex functional of $L_1(0, 1)$.

THEOREM 2.10. *Let $\mathcal{A}$ be an invariant open convex subset of $L_1(0, 1)$. Let ϕ be a continuous functional defined on $\mathcal{A}$ such that ϕ is constant over functions that are equivalent in distribution. If the Gâteaux differential $\partial\phi/\partial h(y) \geqq 0$ for each $y \in \mathcal{D}_\infty \cap \mathcal{A}$ and $h \in \mathcal{T}$, then ϕ is Schur-convex on $\mathcal{A}$.*

Proof. Since ϕ is constant over functions that are equivalent in distribution, it suffices to prove that ϕ is Schur-convex on $\mathcal{D}_1 \cap \mathcal{A}$.

Let $y_1, y_2 \in \mathcal{D}_1 \cap \mathcal{A}$ be right continuous and $y_2 \geqq^m y_1$. Let $\varepsilon > 0$ be arbitrary, then there exists $\delta > 0$ such that

$$\int_0^\delta |y_i(t)|\, dt < \frac{\varepsilon}{4} \quad \text{and} \quad \int_{1-\delta}^1 |y_i(t)|\, dt < \frac{\varepsilon}{4}, \quad i = 1, 2.$$

Since the y_i's are in $\mathcal{D}_1$, they are bounded on the interval $[\delta, 1-\delta]$. Define

$$y_{i\varepsilon}(t) = \frac{1}{\delta}\left[\int_0^\delta y_i(s)\, ds\right] I_{(0,\delta)}(t) + y_i(t) I_{[\delta, 1-\delta]}(t)$$

$$+ \frac{1}{\delta}\left[\int_{1-\delta}^1 y_i(s)\, ds\right] I_{(1-\delta,1)}(t), \qquad i = 1, 2.$$

Then $y_{i\varepsilon} \in \mathcal{D}_\infty$, $i = 1$, 2 and $y_{2\varepsilon} \geqq^m y_{1\varepsilon}$. We also have

$$\|y_{i\varepsilon} - y_i\|_1 = \int_0^\delta |y_{i\varepsilon} - y_i| + \int_{1-\delta}^1 |y_{i\varepsilon} - y_i| \leqq \varepsilon, \qquad i = 1, 2.$$

Hence, for sufficiently small ε, $y_{i\varepsilon} \in \mathcal{D}_\infty \cap \mathcal{A}$. Since ϕ is also L_∞-continuous on $\mathcal{A} \cap L_\infty(0, 1)$, it now follows from Theorem 2.9 that $\phi(y_{2\varepsilon}) \geqq \phi(y_{1\varepsilon})$. Since ϕ is a continuous functional, we obtain that $\phi(y_2) \geqq \phi(y_1)$ by letting $\varepsilon \to 0$. This completes the proof. □

The following lemma is used to prove Theorem 2.12, which is an analogue of Theorem 2.10 for functionals on $L_1(0, 1)$ which are nondecreasing with respect to the ordering of unrestricted majorization.

LEMMA 2.11. *Let $y_1, y_2 \in L_1(0, 1)$ such that $y_2 \geqq^u y_1$. For each $\varepsilon > 0$, there exists a partition $0 < a_1 < \cdots < a_n < 1$ such that the step functions defined by*

$$y_{i\varepsilon}(t) = \frac{1}{a_1}\int_0^{a_1} [y_i(s)\, ds] I_{(0,a_1)}(t) + \sum_{k=1}^{n-1} \frac{1}{a_{k+1} - a_k}\left[\int_{a_k}^{a_{k+1}} y_i(s)\, ds\right] I_{[a_k, a_{k+1})}(t)$$

$$+ \frac{1}{1-a_n}\left[\int_{a_n}^1 y_i(s)\, ds\right] I_{[a_n, 1)}(t), \qquad i = 1, 2,$$

satisfy the following:

(i) $\|y_{i\varepsilon} - y_i\|_1 < \varepsilon$, $i = 1, 2$, *and*

(ii) $y_{2\varepsilon} \geqq^u y_{1\varepsilon}$.

Proof. Note that if y_1, y_2 are continuous functions on the interval $[0, 1]$, then (i) follows from the uniform continuity of y_1 and y_2. If y_i's are not continuous on $[0, 1]$, we first approximate y_i's by continuous functions x_i on $[0, 1]$ such that $\|y_i - x_i\|_1 < \varepsilon/3$,

$i=1, 2$. Next, we find a partition $0<a_1<\cdots<a_n<1$ such that the step functions defined by

$$x_{i\varepsilon}(t)=\frac{1}{a_1}\left[\int_0^{a_1} x_i(s)\,ds\right]I_{(0,a_1)}(t)+\sum_{k=1}^{n-1}\frac{1}{a_{k+1}-a_k}\left[\int_{a_k}^{a_{k+1}} x_i(s)\,ds\right]I_{[a_k,a_{k+1})}(t)$$
$$+\frac{1}{1-a_n}\left[\int_{a_n}^{1} x_i(s)\,ds\right]I_{[a_n,1)}(t), \qquad i=1,2$$

satisfy $\|x_{i\varepsilon}-x_i\|_1<\varepsilon/3$.

Now, define the step functions $y_{1\varepsilon}$, $y_{2\varepsilon}$ by

$$y_{i\varepsilon}(t)=\frac{1}{a_1}\left[\int_0^{a_1} y_i(s)\,ds\right]I_{(0,a_1)}(t)+\sum_{k=1}^{n-1}\frac{1}{a_{k+1}-a_k}\left[\int_{a_k}^{a_{k+1}} y_i(s)\,ds\right]I_{[a_k,a_{k+1})}(t)$$
$$+\frac{1}{1-a_n}\left[\int_{a_n}^{1} y_i(s)\,ds\right]I_{[a_n,1)}(t), \qquad i=1,2.$$

Then

$$\|y_{i\varepsilon}-x_{i\varepsilon}\|_1=\left|\int_0^{a_1}[y_i(s)-x_i(s)]\,ds\right|+\sum_{k=1}^{n-1}\left|\int_{a_k}^{a_{k+1}}[y_i(s)-x_i(s)]\,ds\right|$$
$$+\left|\int_{a_n}^{1}[y_i(s)-x_i(s)]\,ds\right|$$
$$\leqq \|x_i-y_i\|_1, \qquad i=1,2.$$

Thus

$$\|y_i-y_{i\varepsilon}\|_1\leqq\|y_i-x_i\|_1+\|x_i-x_{i\varepsilon}\|_I+\|x_{i\varepsilon}-y_{i\varepsilon}\|_1$$
$$\leqq\frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon, \qquad i=1,2.$$

This proves the first part of the lemma.

Let $y_2\geqq^u y_1$. Then for any partition $0<a_1<\cdots<a_n<1$, the step functions $y_{i\varepsilon}$, $y_{2\varepsilon}$ satisfy that $y_{2\varepsilon}\geqq^u y_{1\varepsilon}$. This proves (ii). □

THEOREM 2.12. *Let $\mathscr{A}$ be an open convex subset of $L_1(0,1)$ and let ϕ be a continuous functional on $\mathscr{A}$ such that the Gâteaux differentials $\partial\phi/\partial h(y)\geqq 0$ for $y\in\mathscr{A}$ and $h\in\mathscr{T}$. Then $y_1, y_2\in\mathscr{A}$ and $y_2\geqq^u y_1$ imply that $\phi(y_2)\geqq\phi(y_1)$.*

Proof. We shall first prove the theorem for step functions. Let y_1, y_2 be step functions in $\mathscr{A}$ and $y_2\geqq^u y_1$. Then by Lemma 2.7, $y_2-y_1=\sum_{i=1}^N h_i$ for some $\{h_1,\cdots,h_N\}\subseteq\mathscr{T}$. Since $\mathscr{A}$ is open, for sufficiently small positive θ, $y_1+\theta h_1,\cdots,y_1+\theta\sum_{i=1}^N h_i$ are functions in $\mathscr{A}$ and satisfy

$$y_1+\theta\sum_{i=1}^N h_i\overset{u}{\geqq} y_1+\theta\sum_{i=1}^{N-1} h_i\overset{u}{\geqq}\cdots\overset{u}{\geqq} y_1+\theta h_1\geqq y_1.$$

It now follows from Lemma 2.6 that $\phi(y_1+\theta\sum_{i=1}^N h_i)\geqq\cdots\geqq\phi(y_1+\theta h_1)\geqq\phi(y_1)$. Define $\Theta=\{0\leqq\theta\leqq 1\colon \phi(y_1+\theta\sum_{i=1}^N h_i)\geqq\phi(y_i)\}$ and $\theta_0=\sup\{\theta\colon\theta\in\Theta\}$. Following the same argument as in the proof of Theorem 2.9, we can show that $\theta_0=1$. Hence,

$$\phi(y_2)=\phi\left(y_1+\sum_{i=1}^N h_i\right)\geqq\phi(y_1).$$

In general, let $y_1, y_2 \in L_1(0,1)$ and $y_2 \geqq^u y_1$. Let $\varepsilon > 0$. By Lemma 2.11 there exist step functions $y_{1\varepsilon}$, $y_{2\varepsilon}$ such that $\|y_{i\varepsilon} - y_i\|_1 < \varepsilon$ for $i = 1, 2$ and $y_{2\varepsilon} \geqq^u y_{1\varepsilon}$. Since $\mathscr{A}$ is open, for sufficiently small ε, $y_{1\varepsilon}$, $y_{2\varepsilon}$ are functions in $\mathscr{A}$. Thus $\phi(y_{2\varepsilon}) \geqq \phi(y_{1\varepsilon})$. Since ϕ is continuous, we conclude that $\phi(y_2) \geqq \phi(y_1)$ by letting $\varepsilon \to 0$. □

3. Applications. The inequality given in Theorem 1.4 can be reformulated as the statement that the functional defined by

$$\phi(x) = \int_0^1 \log \left[\int_0^1 u(t)^{x(s)}\, dt \right] ds$$

is Schur-convex. By Theorem 2.10, this is equivalent to the condition $\partial\phi/\partial h(x) \geqq 0$ for all $x \in \mathscr{D}_\infty$, $\forall h \in \mathscr{T}$. This condition can be verified as follows.

Using Holder's inequality, we note that the function $M(\alpha) = \log \|u\|_\alpha^\alpha$ is convex in α, and thus

$$M'(\alpha) = \frac{\int_0^1 u(t)^\alpha \log u(t)\, dt}{\int_0^1 u(t)^\alpha\, dt}$$

is increasing in α. Let $x \in \mathscr{D}_\infty$ and $h \in \mathscr{T}$; then both x and h are functions decreasing on their supports, and $\int_0^1 h(t)\, dt = 0$. This implies that

$$\frac{\partial\phi}{\partial h}(x) = \int_0^1 \left[\frac{\int_0^1 u(t)^{x(s)} \log u(t)\, dt}{\int_0^1 u(t)^{x(s)}\, dt} \right] h(s)\, ds \geqq 0.$$

More generally, we can replace the function $u(t)^{x(s)}$ by functions of the form $\psi(t, z)$ which are log convex in z for fixed t. This is the result of Proschan and Sethuraman [8], which we will state below.

THEOREM 3.1. *Let the function* $\psi(t, z)$ *on* $(0,1) \times (-\infty, \infty)$ *be a log convex function in* z *for fixed* t, *and the partial derivative* $\psi_2(t, z) = \partial/\partial z \psi(t, z)$ *exists. Also let* $\sup_{|z| \leqq k} \psi(t, z)$ *belong to* $L_1(0,1)$ *for each* $k < \infty$. *For any bounded measurable function* x *on* $(0,1)$, *define*

$$M_\psi(x) = \int_0^1 \log \left[\int_0^1 \psi(t, x(s))\, dt \right] ds.$$

Then M_ψ *is Schur-convex.*

Proof. It follows from Artin's theorem [1] that the positive linear combination $\int_0^1 \psi(t, z)\, dt$ is log convex in z. Let $x \in \mathscr{D}_\infty$ and $h \in \mathscr{T}$, then

$$\frac{\partial M_\psi}{\partial h}(x) = \int_0^1 \left[\frac{\int_0^1 \psi_2(t, x(s))\, dt}{\int_0^1 \psi(t, x(s))\, dt} \right] h(s)\, ds \geqq 0,$$

which implies that M_ψ is Schur-convex. □

Next, we shall study an application of unrestricted majorization to peakedness ordering of symmetric distributions.

Let X and Y be random variables possessing densities symmetric about zero. According to the definition of peakedness given by Birnbaum [2], X is more peaked than Y, ($X \geqq^p Y$ in symbols), if $P(X \leqq t) \geqq P(Y \leqq t)$ for all $t \geqq 0$. Let f and g be the densities of X and Y respectively. Then the condition $X \geqq^p Y$ is equivalent to $f >^u g$ on the interval $(0, \infty)$.

Birnbaum [2] showed that under appropriate conditions, $X_1 \geqq^p Y_1$ and $X_2 \geqq^P Y_2$ imply that $X_1 + X_2 \geqq^P Y_1 + Y_2$. This result can be obtained by considering certain order preserving functionals. We first introduce some simplifying notations.

For $s > 0$, define

$$\chi_s(x) = I(|x| < s)$$

and, for a symmetric function h, define

$$\begin{aligned} h(s, x) &= (h * \chi_s)(x) \\ &= \int h(x-y)\chi_s(y)\,dy \end{aligned}$$

and

$$\begin{aligned} h(s, 0) &= \int h(-y)\chi_s(y)\,dy \\ &= \int h(y)\chi_s(y)\,dy. \end{aligned}$$

Note that

$$h(s, x) = \begin{cases} \frac{1}{2}[h(x+s, 0) - h(x-s, 0)] & \text{if } s < x, \\ \frac{1}{2}[h(x+s, 0) + h(-x+s, 0)] & \text{if } -s \leqq x \leqq s, \\ \frac{1}{2}[h(-x+s, 0) - h(-x-s, 0)] & \text{if } x < -s. \end{cases}$$

We need the following lemma.

LEMMA 3.2. *Let $\mathscr{C} = \{h\colon h$ symmetric and $h(s, 0) \geqq 0$ for all $s > 0\}$. Let g be symmetric and decreasing on $(0, \infty)$. Then $h * g \in \mathscr{C}$ for all $h \in \mathscr{C}$, i.e., $(h * g * \chi_s)(0) \geqq 0$ for $h \in \mathscr{C}$ and $s > 0$.*

Proof. Let $h \in \mathscr{C}$ and $s > 0$. Then

$$\begin{aligned} (h * g * \chi_s)(0) &= \int (h * \chi_s)(x) g(-x)\,dx \\ &= \int h(s, x) g(-x)\,dx \\ &= \frac{1}{2}\left\{\int_{x>s} [h(x+s, 0) - h(x-s, 0)] g(x)\,dx \right. \\ &\qquad + \int_{-s\leqq x\leqq s} [h(x+s, 0) - h(-x+s, 0)] g(x)\,dx \\ &\qquad \left. + \int_{x<-s} [h(-x+s, 0) - h(-x-s, 0)] g(x)\,dx \right\} \\ &= \frac{1}{2}\left[\int_{x>-s} h(x+s, 0) g(x)\,dx + \int_{x<s} h(-x+s, 0) g(x)\,dx \right. \\ &\qquad \left. - \int_{x<-s} h(-x-s, 0) g(x)\,dx - \int_{x>s} h(x-s, 0) g(x)\,dx \right]. \end{aligned}$$

Let $y = x + s$ in the first integral, $y = -x + s$ in the second integral, $y = -x - s$ in the third integral and $y = x - s$ in the fourth integral. We get

$$\begin{aligned} (h * g * \chi_s)(0) = \frac{1}{2}\left[\int_{y>0} h(y, 0) g(y-s)\,dy + \int_{y>0} h(y, 0) g(-y+s)\,dy \right. \\ \left. - \int_{y>0} h(y, 0) g(-y-s)\,dy - \int_{y>0} h(y, 0) g(y+s)\,dy \right]. \end{aligned}$$

By the symmetry of g,

$$(h * g * \chi_s)(0) = \int_{y>0} h(y, 0)[g(y-s) - g(y+s)]\, dy.$$

Since $h(y, 0) \geqq 0$ for all $y > 0$, and $g(y-s) - g(y+s) \geqq 0$ for $y > 0$ and $s > 0$, we conclude that $(h * g * \chi_s)(0) \geqq 0$. □

We may now obtain the following result.

THEOREM 3.3. *Let* X_1, X_2, Y_1, Y_2 *be independent symmetric random variables on* $(-1, 1)$ *with densities* f_1, f_2, g_1, g_2, *respectively. Let* f_1, g_2 *be nonincreasing on* $(0, 1)$. *Let* $X_i \geqq^P Y_i$ *for* $i = 1, 2$. *Then* $X_1 + X_2 \geqq^P Y_1 + Y_2$.

Proof. We will first establish that $X_1 + X_2 \geqq^P X_1 + Y_2$.

Fix f_1. For each $f \in L_1(0, 1)$, let f_S be the symmetric function on $(-1, 1)$ defined by

$$f_S(t) = f(|t|).$$

For each $s > 0$, define a functional on $L_1(0, 1)$ by

$$\phi_s(f) = \iint I(|x_1 + x_2| \leqq s) f_1(x_1) f_S(x_2)\, dx_1\, dx_2.$$

Let $T(0, 1)$ be the class of nonnegative functions u on $(0, 1)$ with $\int_0^1 u(t)\, dt = \frac{1}{2}$. Note that for $f \in T(0, 1)$,

$$\phi_s(f) = P(|X_1 + Z| \leqq s),$$

where X_1, Z are independent random variables with densities f_1, f_S respectively.

We shall show that for each $s > 0$, ϕ_s is nondecreasing with respect to the ordering of unrestricted majorization on $T(0, 1)$. Let $s > 0$. Let $f \in T(0, 1)$ and $h \in \mathscr{T}$. Then,

$$\begin{aligned}\frac{\partial \phi_s}{\partial h}(f) &= \lim_{\theta \to 0} \frac{1}{\theta} \iint I(|x_1 + x_2| \leqq s) \\ &\quad \cdot \{f_1(x_1)[f_S(x_2) + \theta h_S(x_2)] - f_1(x_1) f_S(x_2)\}\, dx_1\, dx_2 \\ &= \iint I(|x_1 + x_2| \leqq s) f_1(x_1) h_S(x_2)\, dx_1\, dx_2 \\ &= (f_1 * h_S * \chi_s)(0).\end{aligned}$$

Since $h \in \mathscr{T}$, $h_S(t, 0) \geqq 0$ for all $t > 0$. By Lemma 3.2, $(f_1 * h_S * \chi_s)(0) \geqq 0$. It now follows from Theorem 2.12 that ϕ_s is nondecreasing with respect to the ordering of unrestricted majorization on $T(0, 1)$.

Note that $X_2 \geqq^P Y_2$ implies that $f_2 \geqq^u g_2$ when these are considered as elements of $T(0, 1)$. We now have

$$\begin{aligned}P(|X_1 + X_2| \leqq s) &= \phi_s(f_2) \\ &\geqq \phi_s(g_2) \\ &= P(|X_1 + Y_2| \leqq s) \quad \text{for all } s > 0.\end{aligned}$$

Thus $X_1 + X_2 \geqq^P X_1 + Y_2$. Similarly, we can establish that $X_1 + Y_2 \geqq^P Y_1 + Y_2$. Hence $X_1 + X_2 \geqq^P Y_1 + Y_2$. □

REFERENCES

[1] E. ARTIN (1931), *Einfürhring in die Theorie der Gammafunktion*, Hambuger Mathematische Einzelschriften, II. *The Gamma Function*, translated by M. Butler, Holt, Rinehart and Winston, 1964. (English translation.)

[2] Z. W. BIRNBAUM (1948), *On random variables with comparable peakedness*, Ann. Math. Statist., 19, pp. 76-81.

[3] K. M. CHONG (1976), *Doubly stochastic operators and rearrangement theorems*, J. Math. Anal. Appl., 56, pp. 309-316.

[4] P. W. DAY (1973), *Decreasing rearrangements and doubly stochastic operators*, Trans. Amer. Math. Soc., 178, pp. 383-392.

[5] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA (1934), *Inequalities*, Cambridge University Press, Cambridge.

[6] A. W. MARSHALL AND I. OLKIN (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

[7] A. M. OSTROWSKI (1952), *Sur quelques applications des fonctions convexes et concaves au sens de I. Schur*, J. Math. Pures Appl., 31, pp. 253-292.

[8] F. PROSCHAN AND J. SETHURAMAN (1977), *Two generalizations of Muirhead's theorem*, Bull. Calcutta Math. Soc., 69, pp. 341-344.

[9] S. M. ROSS (1982), *Stochastic Processes*, John Wiley, New York.

[10] J. V. RYFF (1963), *On the representation of doubly stochastic operators*, Pacific J. Math., 13, pp. 1379-1386.

[11] ——— (1967), *On Muirhead's theorem*, Pacific J. Math., 21, pp. 567-576.

[12] ———, (1970), *Measure preserving transformations and rearrangements*, J. Math. Anal. Appl., 31, pp. 449-458.

[13] I. SCHUR (1923), *Über eine Klasse von Mittelbeildunger mit Anwendungen die Determinanten—Theorie Sitzungsber*, Berlin Math. Gesellschaft, 22, pp. 9-20.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

NONRESONANT BIFURCATIONS WITH SYMMETRY*

JACK CARR†, JAN A. SANDERS‡ AND STEPHAN A. VAN GILS‡

Abstract. We prove uniqueness of the limit cycle for generic perturbations of a planar integrable vectorfield which arises in the unfolding of a system of two linear uncoupled nonresonant oscillators.

Key words. Abelian integrals, limit cycle, planar vectorfields

AMS(MOS) subject classifications. 34C05, 34C25, 34C29

1. Introduction. Consider a system of ordinary differential equations in $\mathbb{R}^n$, $n \geqq 2$, and suppose that the origin is a nonhyperbolic equilibrium. If the linear part of the vectorfield is doubly degenerate then, after reduction to a center manifold, it takes one of the forms [2], [6], [8]

$$A_1=\begin{pmatrix}0&1\\0&0\end{pmatrix},\quad A_2=\begin{pmatrix}0&1&0\\-1&0&0\\0&0&0\end{pmatrix},\quad A_3=\begin{pmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&\omega\\0&0&-\omega&0\end{pmatrix}.$$

The unfolding of A_1 has been given by Arnold [1], Bogdanov [3] and Takens [12]. The unfoldings of A_2 and A_3 are still far from complete. Partial results have been obtained by Langford [10] and Iooss and Langford [9]. Their methods are restricted to the analysis of planar vectorfields. In these planar vectorfields a global Hopf bifurcation occurs (see Guckenheimer and Holmes [8]). The main mathematical difficulty is to prove the uniqueness of the limit cycle for generic perturbations.

In the case of the singularity A_2 this has been done by Carr, Chow and Hale [5], Sanders and Cushman [11] and Zholondek [13]. Only the last paper solves the problem without any restriction.

The singularity A_3 leads to the study of generic perturbations of the planar Hamiltonian vectorfield with Hamiltonian function $H(x, y)=x^p y^q(1-x-y)$. Most important, one wants to prove the uniqueness of limit cycles for these perturbed vectorfields.

The method of proof was taken from Carr, Chow and Hale [5]; the proof in this paper may seem to be rather ad hoc at first sight, but we have extended it in a more general setting, which as a bonus has produced a sharper result in the original problem [7].

On the one hand, our method seems to work only in the case $p=q$ where the family of curves is hyperelliptic. On the other hand, at least part of this work depends only on that fact and consequently will be useful in other bifurcation problems, for instance in problems with more than one parameter, too.

Zholondek [14] proves the uniqueness of limit cycles by a priori estimates without restriction on p and q.

We will not attempt to relate the results to the behaviour of the original four-dimensional system, because the expected homoclinic phenomenon leads to quite complex behaviour (see [4]).

* Received by the editors January 27, 1985; accepted for publication (in revised form) January 22, 1986.

† Heriot-Watt University, Edinburgh, Scotland.

‡ Free University, Amsterdam, The Netherlands.

In § 2 we show how to obtain the relevant equations in the plane and we discuss the unperturbed vectorfield in the symmetric case. The proof of the uniqueness of periodic orbits reduces to proving monotonicity of a certain function, which is the ratio of two Abelian integrals. In this section we also give a sketch of the proof of the uniqueness result.

In § 4 we give all the details of the proof.

2. Formulation of the problem. Consider a singularity of type A_3. In order to classify the behaviour of nearby vectorfields we look at the vectorfield in $\mathbb{R}^4$ given by

$$\dot{v} = \tilde{A}v + G(v), \tag{2.1}$$

where

$$\tilde{A} = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ -1 & \alpha & 0 & 0 \\ 0 & 0 & \beta & \omega \\ 0 & 0 & -\omega & \beta \end{pmatrix},$$

$\alpha, \beta \in \mathbb{R}$ are small parameters, $\omega \in \mathbb{R}$, G is smooth and of higher order in v. We will assume the nonresonance condition: $\omega \neq 0, \pm 1, \pm 2, \pm 3$. If one introduces polar coordinates

$$v_1 = r_1 \cos \theta_1, \qquad v_3 = r_2 \cos \theta_2,$$

$$v_2 = -r_1 \sin \theta_1, \qquad v_4 = -r_2 \sin \theta_2,$$

then (2.1) is equivalent to

$$\begin{aligned} \dot{r}_1 &= \alpha r_1 + R_1(\theta_1, \theta_2, r_1, r_2), \\ \dot{r}_2 &= \beta r_2 + R_2(\theta_1, \theta_2, r_2, r_2), \\ \dot{\theta}_1 &= 1 + \Theta_1(\theta_1, \theta_2, r_1, r_2), \\ \dot{\theta}_2 &= \omega + \Theta_2(\theta_1, \theta_2, r_1, r_2) \end{aligned} \tag{2.2}$$

where each function is 2π-periodic in θ_1 and θ_2. After third order averaging one obtains the vectorfield

$$\begin{aligned} \dot{r}_1 &= r_1(\alpha + cr_1^2 + dr_2^2) + \cdots, \\ \dot{r}_2 &= r_2(\beta + er_1^2 + fr_2^2) + \cdots, \\ \dot{\theta}_1 &= 1 + \cdots, \\ \dot{\theta}_2 &= \omega + \cdots, \end{aligned} \tag{2.3}$$

where the dots indicate higher order terms and c, d, e, f are real constants. We will ignore the higher order terms and analyze the planar autonomous system

$$\dot{r}_1 = r_1(\alpha + cr_1^2 + dr_2^2), \qquad \dot{r}_2 = r_2(\beta + er_1^2 + fr_2^2). \tag{2.4}$$

Substituting $r_1^2 = x$ and $r_2^2 = y$ and rescaling time we obtain the polynomial vectorfield

$$\dot{x} = x(\alpha + cx + dy), \qquad \dot{y} = y(\beta + ex + fy). \tag{2.5}$$

First we look for those values of the constants for which the system becomes integrable. More precisely, if we use the rescalings

$$t=\bar{t}p/\beta,\quad y=-\beta/f\bar{y},\quad x=\frac{-\beta(p+1)}{ep}\bar{x},$$

and we replace $\bar{t}, \bar{y}, \bar{x}$ by t, y, x, respectively, then (2.5) becomes

$$\dot{x}=px(\tilde{a}+\tilde{c}x+\tilde{d}y),\qquad \dot{y}=py\left(1-\frac{p+1}{p}x-y\right),\tag{2.6}$$

for some constants $\tilde{a}$, $\tilde{c}$ and $\tilde{d}$. If we use the integration factor $|x|^{p-1}|y|^{q-1}$, (2.6) is integrable with first integral

$$H=|x|^p|y|^q(1-x-y),\tag{2.7}$$

provided we choose

$$\tilde{a}=-\frac{q}{p},\quad \tilde{c}=\frac{q}{p},\quad \tilde{d}=\frac{q+1}{p}.$$

In all of what follows we will assume (for purely technical reasons) that $p=q$. Under this symmetry hypothesis the integrable system (2.6) becomes

$$X_p:\quad \begin{aligned}\dot{x}&=px\left(-1+x+\frac{p+1}{p}y\right),\\ \dot{y}&=py\left(1-\frac{p+1}{p}x-y\right).\end{aligned}\tag{2.8}$$

Figure 1 gives the various phase portraits of the one-parameter family of vectorfields X_p depending on the value of p.

For $p\in(-\frac{1}{2},\infty)\backslash\{0\}$ we see that X_p has periodic orbits.

For computational convenience we introduce new coordinates u and v defined by

$$x=u+v,\qquad y=u-v.$$

The induced Hamiltonian vectorfield X_H is

$$\dot{u}=pv(-1+2u),\qquad \dot{v}=pu\left(-1+\frac{2p+1}{p}u\right)-v^2\tag{2.9}$$

with Hamiltonian function

$$H=(u^2-v^2)^p(1-2u).\tag{2.10}$$

For ε_1, ε_2 small, consider the vectorfield Y_ε, given by

$$\dot{u}=pv(-1+2u),\qquad \dot{v}=pu\left(-1+\frac{2p+1}{p}u\right)-v^2+\varepsilon_1v+\varepsilon_2u^2v,\tag{2.11}$$

which is a non-Hamiltonian perturbation of X_H. Any other cubic perturbation can by partial integration be shown to lead to a fractional linear transformation of the quotient $\tilde{Q}$ (to be defined in the sequel) and therefore preserves monotonicity results. The derivative of H along the integral curves of Y_ε is

$$\begin{aligned}\dot{H}=\frac{\partial H}{\partial u}\dot{u}+\frac{\partial H}{\partial v}\dot{v}&=2((\varepsilon_1v+\varepsilon_2u^2v-\dot{v})\dot{u}+\dot{u}\dot{v})(u^2-v^2)^{p-1}\\ &=2(\varepsilon_1v+\varepsilon_2u^2v)(u^2-v^2)^{p-1}\dot{u}.\end{aligned}$$

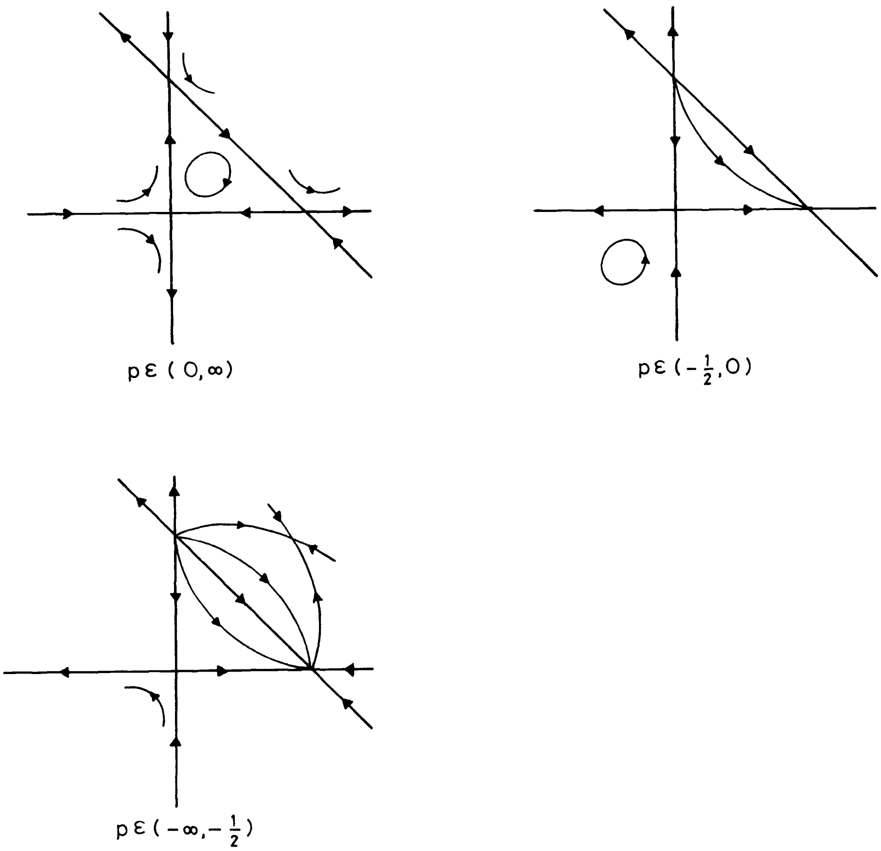


FIG. 1. *Phase portraits of* X_p.

It is not difficult to show (Carr, Chow and Hale [5], Chow and Hale [6]) that a necessary and sufficient condition for an orbit $\Gamma(\varepsilon_1, \varepsilon_2)$ of (2.11) to be periodic is that

$$\int_\Gamma \dot{H}\, dt = 0;$$

in other words

$$\varepsilon_1 \int_\Gamma v(u^2 - v^2)^{p-1}\, du + \varepsilon_2 \int_\Gamma u^2 v(u^2 - v^2)^{p-1}\, du = 0.$$

Next it is shown that if $\varepsilon_1/\varepsilon_2$ is in the range of the strictly monotone function $\tilde{Q}(h)$, (2.11) has, for ε_1, ε_2 small enough, precisely one periodic orbit, where

$$\tilde{Q}(h) = \frac{\int_{\gamma(h)} u^2 v(u^2 - v^2)^{p-1}\, du}{\int_{\gamma(h)} v(u^2 - v^2)^{p-1}\, du}. \tag{2.12}$$

Here $\gamma(h)$ is a periodic orbit of the unperturbed vectorfield. More precisely $\gamma(h)$ is a smooth compact connected component of the level set $H^{-1}(h)$. When $p > 0$, h takes values in the set $(0, h_H)$ and when $-\frac{1}{2} < p < 0$, h takes values in the set (h_H, ∞) where

$$h_H = \left[\left(\frac{p}{2p+1}\right)^2\right]^p \frac{1}{2p+1}. \tag{2.13}$$

In other words, h_H is the value of H where the periodic orbit arises from the center. $\gamma(h)$ is sketched in Fig. 2.

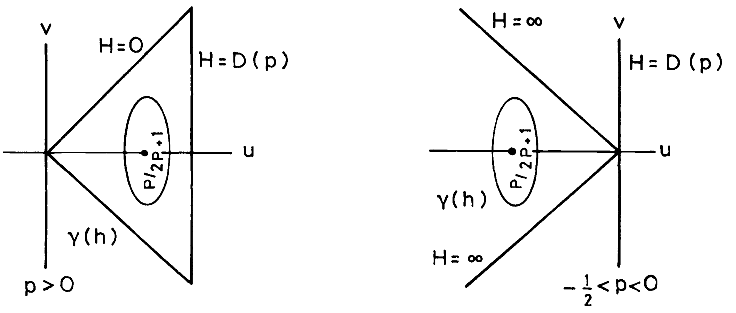


FIG. 2. *Sketch of the level sets of* $H^{-1}(h)$.

In § 4 we will prove the monotonicity of $\tilde{Q}(h)$. This result is a first step in verifying the bifurcation diagrams given in Guckenheimer and Holmes [8].

DEFINITION 2.1. For $-\frac{1}{2}<p<0$ ($p>0$) let I_p denote the interval $[h_H, \infty)$ ($[0, h_H]$).

The following result is the main theorem of this paper.

THEOREM 2.1. *For* $p\in(-\frac{1}{2},\infty)\backslash\{0\}$, $\tilde{Q}(h)$ *is strictly monotone.*

Sketch of the proof. For $n\in\mathbb{N}\cup\{0\}$ and $m\in\mathbb{N}\cup\{-1,0\}$ let

$$\delta^n_m(h)=\int_{\gamma(h)}\frac{u^n v^m}{(1-2u)^{p-1/p}}\,du.$$

It is clear from (2.10), (2.12) that

$$\tilde{Q}(h)=\delta^2_1/\delta^0_1.$$

In Lemma 4.3 we show that δ^0_1 and δ^1_1 are related by $\delta^1_1=p/(2p+1)\delta^0_1$. Therefore in place of $\tilde{Q}$ we may equally well consider

$$Q(h)=\delta^2_1/\delta^1_1=\frac{2p+1}{p}\tilde{Q}(h).$$

Lemma 4.4 states that δ^2_1, δ^1_1 satisfy an inhomogeneous nonautonomous linear differential equation:

$$2ph\dot{\delta}^1_1=\delta^1_1-\delta^3_{-1},\qquad 2ph\dot{\delta}^2_1=-\frac{2p}{2p+1}\delta^1_1+\frac{6p+2}{2p+1}\delta^2_1-\frac{p}{2p+1}\delta^3_{-1}.$$

These equations fit into the framework of the next section. We can apply Theorem 2.1 for $p\in(-\frac{1}{3},\infty)\backslash\{0\}$ directly to obtain a priori bounds for $Q(h)$. For $-\frac{1}{2}<p\leqq-\frac{1}{3}$ we will have to add some more arguments to obtain:

$$\left.\begin{matrix}p\in(0,\infty)\\ h\in[0,h_H]\end{matrix}\right\}\Rightarrow\frac{3p}{6p+2}=Q(0)\geqq Q(h)\geqq Q(h_H)=\frac{p}{2p+1},$$

$$\left.\begin{matrix}p\in(-\frac{1}{3},0)\\ h\in[h_H,\infty)\end{matrix}\right\}\Rightarrow\frac{p}{2p+1}=Q(h_H)\geqq Q(h)\geqq\frac{3p}{6p+2}=Q(\infty),$$

$$\left.\begin{matrix}p\in(-\frac{1}{2},-\frac{1}{3}]\\ h\in[h_H,\infty)\end{matrix}\right\}\Rightarrow\frac{p}{2p+1}=Q(h_H)\geqq Q(h)>-\infty=Q(\infty).$$

Using these bounds on Q we prove that the second derivative of Q is of one sign at critical points of Q. More precisely, in Theorem 3.3 we prove that $\dot{Q}=0$ implies that

$$\frac{h\delta^1_1}{p/(2p+1)-Q}\bar{Q}=h\bar{\delta}^1_1+\left(1-\frac{h\dot{\delta}^1_1}{\delta^1_1}\right)\dot{\delta}^1_1.$$

The hardest part of the argument is Lemma 4.6, where we prove the inequality $\operatorname{sign}(p)\{h\ddot{\delta}_1^1+(1-(h\dot{\delta}_1^1/\delta_1^1))\dot{\delta}_1^1\}<0$ at critical points. Consequently, $\operatorname{sign}(p)\ddot{Q}>0$ at points where $\dot{Q}$ vanishes. This shows that Q is monotone. □

3. A useful framework. In this section we explain some of the structure behind the computations in the paper of Carr, Chow and Hale [5]. We formulate two theorems which apply to both the problem considered in [5] as well as to ours.

Assume that α_0, α_1, q and R are real valued smooth functions, defined on $I=[h_0, h_1)$, $h_0\in\mathbb{R}$, $h_0<h_1\leqq\infty$. Furthermore suppose that on I the differential equations

$$q\dot{\alpha}_0=a\alpha_0+b\alpha_1+eR, \qquad q\dot{\alpha}_1=c\alpha_0+d\alpha_1+fR \tag{3.1}$$

where

$$a, b, c, d, e, f\in\mathbb{R}$$

hold. Let

$$Q=\alpha_1/\alpha_0, \quad \Delta_1=af-ec, \quad \Delta_2=ed-bf, \quad \Delta_3=ad-bc.$$

Then we have

THEOREM 3.1. *Suppose that the following conditions hold*:

(i) $\quad e\neq 0, \quad \Delta_2\neq 0, \quad \dfrac{f}{e}-\dfrac{\Delta_1}{\Delta_2}\neq 0, \quad R(h_0)\neq 0,$

(ii) $\quad \alpha_1(h_0)=\alpha_0(h_0)=0,$

(iii) $\quad \lim\limits_{h\to h_1} Q(h)=\Delta_1/\Delta_2,$

(iv) $\quad \operatorname{sign}\left\{q(h)\dfrac{\dot{\alpha}_0(h)}{\alpha_0(h)}\right\}\neq\operatorname{sign}\dfrac{\Delta_2}{e}$ *on* I,

(v) $\quad Q(h)$ *is bounded on* I,

then,

(α) $\quad Q(h_0)=f/e,$

(β) $\quad \forall h\in I\colon \ \min(f/e), \Delta_1/\Delta_2)\leqq Q(h)\leqq\max(f/e, \Delta_1/\Delta_2).$

Proof. To prove (α) observe that

$$Q(h_0)=\frac{\alpha_1(h_0)}{\alpha_0(h_0)}=\frac{\dot{\alpha}_1(h_0)}{\dot{\alpha}_0(h_0)}=\frac{q(h_0)\dot{\alpha}_1(h_0)}{q(h_0)\dot{\alpha}_0(h_0)}=\frac{f}{e}.$$

Suppose that for some $\bar{h}\in I$, $\dot{Q}(\bar{h})=0$. Then at $h=\bar{h}$

$$q\frac{\dot{\alpha}_0}{\alpha_0}\left(Q-\frac{f}{e}\right)=\frac{\Delta_2}{e}\left(Q-\frac{\Delta_1}{\Delta_2}\right), \tag{3.2}$$

which follows from

$$q\frac{\dot{\alpha}_0}{\alpha_0}Q=q\frac{\dot{\alpha}_1}{\alpha_0}=c+dQ+\frac{f}{e}\left(q\frac{\dot{\alpha}_0}{\alpha_0}-a-bQ\right).$$

Combining (3.2) with (iii) – (v) gives (β). □

Remark. The third hypothesis in the theorem is equivalent to

$$\lim_{h\to h_1} q(h)\dot{\alpha}_i(h)=0, \qquad i=1,2.$$

It will be useful to have an expression for $\dot{Q}(h_0)$ in terms of the coefficients in the right-hand side of (3.1). This is given in the next lemma. We omit the proof.

LEMMA 3.2. *Suppose that $q(h_0) \neq 0$, $e \neq 0$, then*

$$2q(h_0)\dot{Q}(h_0) = -b\left(\frac{f}{e}\right)^2 + (d-a)\frac{f}{e} + c.$$

The next theorem gives an expression for the second derivative of Q at a critical point of Q.

THEOREM 3.3. *Suppose that for some $\bar{h} \in I$, $\dot{Q}(\bar{h}) = 0$. Furthermore suppose that $e \neq 0, f/e \neq \Delta_1/\Delta_2$. Then at $h = \bar{h}$:*

$$\frac{q\alpha_0}{f/e - Q}\ddot{Q} = q\ddot{\alpha}_0 + \left(\dot{q} - \frac{q\dot{\alpha}_0}{\alpha_0}\right)\dot{\alpha}_0.$$

Proof. We will assume that all relations given above are evaluated at $h = \bar{h}$. To keep the notation simple we will suppress the argument $\bar{h}$. From (3.1) we infer that

$$q\ddot{\alpha}_0 = (a - \dot{q})\dot{\alpha}_0 + \dot{b}\alpha_1 + e\dot{R}, \qquad q\ddot{\alpha}_1 = c\dot{\alpha}_0 + (d - \dot{q})\dot{\alpha}_1 + f\dot{R}.$$

Thus after eliminating $\dot{R}$ we obtain

$$q\ddot{\alpha}_1 = c\dot{\alpha}_0 + (d - \dot{q})\dot{\alpha}_1 + \frac{f}{e}(q\ddot{\alpha}_0 - (a - \dot{q})\dot{\alpha}_0 - b\dot{\alpha}_1).$$

Because $\dot{Q} = 0$ both the relations

$$\dot{\alpha}_0 Q = \dot{\alpha}_1, \qquad \ddot{\alpha}_0 Q + \alpha_0 \ddot{Q} = \ddot{\alpha}_1,$$

hold. Straightforward calculation gives:

$$\begin{aligned} q\alpha_0\ddot{Q} &= -q\ddot{\alpha}_0 Q + q\ddot{\alpha}_1 \\ &= -q\ddot{\alpha}_0 Q + c\dot{\alpha}_0 + (d - \dot{q})\dot{\alpha}_0 Q + \frac{f}{e}q\ddot{\alpha}_0 - \frac{af}{e}\dot{\alpha}_0 + \frac{f}{e}\dot{q}\dot{\alpha}_0 - \frac{bf}{e}\dot{\alpha}_0 Q \\ &= q\left(\frac{f}{e} - Q\right)\ddot{\alpha}_0 + \frac{\dot{\alpha}_0}{e}(-\Delta_1 + f\dot{q} + (\Delta_2 - \dot{q}e)Q). \end{aligned}$$

Dividing by $f/e - Q$ gives

$$\frac{q\alpha_0}{f/e - Q}\ddot{Q} = q\ddot{\alpha}_0 + \dot{\alpha}_0\left\{\frac{f\dot{q} - \Delta_1 + (\Delta_2 - \dot{q}e)Q}{f/e - Q}\right\}. \tag{3.3}$$

If we define $\beta = q\dot{\alpha}_0/\alpha_0$ then (3.2) reads $Q(e\beta - \Delta_2) = \beta f - \Delta_1$. Plugging this relation into the right-hand side of (3.3) gives the desired result. □

4. Monotonicity. In this section we will fill in the details of the sketch of the proof of Theorem 2.1. To do this we derive a differential equation for δ_1^2, δ_1^1 which fits into the framework of the previous section.

For $n \in \mathbb{N} \cup \{0\}$ and $m \in \mathbb{N} \cup \{-1, 0\}$ let

$$\delta_m^n(h) = \int_{\gamma(h)} \frac{u^n v^m}{(1-2u)^{p-1/p}}\,du, \tag{4.1}$$

where $\gamma(h)$ is as in (2.12). From (2.10) and (2.12) we have

$$\tilde{Q}(h) = \delta_1^2/\delta_1^0. \tag{4.2}$$

The next two lemmas give a recursion relation and a differential equation which are satisfied by the functions δ_m^n.

LEMMA 4.1. *Let* $(m+3)/2 \in \mathbb{N}$ *and* $n \in \mathbb{N} \cup \{0\}$, *then*

$$R(m,n): \quad m(2p+1)\delta_{m-2}^{n+2} = mp\delta_{m-2}^{n+1} + (m-2pn-2)\delta_m^n + pn\delta_m^{n-1}.$$

Proof. Differentiate (2.10) with respect to u to obtain

$$0 = pu - (2p+1)u^2 + v^2 - pv\frac{\partial v}{\partial u}(1-2u). \tag{4.3}$$

Multiplying this identity with $u^n v^{m-2}/(1-2u)^{p-1/p}$ and integrating along $\gamma(h)$ yields

$$0 = p\delta_{m-2}^{n+1} - (2p+1)\delta_{m-2}^{n+2} + \delta_m^n - p\int (1-2u)^{1/p} u^n v^{m-1}\, dv.$$

Integrating the last integral by parts gives

$$\begin{aligned}
-\frac{p}{m}\int (1-2u)^{1/p}u^n\, dv^m &= \frac{p}{m}\int \frac{d}{du}\{(1-2u)^{1/p}u^n\}v^m\, du \\
&= -\frac{2}{m}\int (1-2u)^{1/p-1}u^n v^m\, du + \frac{pn}{m}\int (1-2u)^{1/p}u^{n-1}v^m\, du \\
&= -\frac{2}{m}\int \frac{u^n v^m}{(1-2u)^{p-1/p}}\, du + \frac{pn}{m}\int \frac{(1-2u)u^{n-1}v^m}{(1-2u)^{p-1/p}}\, du.
\end{aligned}$$

Therefore, by definition of δ_m^n

$$0 = p\delta_{m-2}^{n+1} - (2p+1)\delta_{m-2}^{n+2} + \delta_m^n - \frac{2}{m}\,\delta_m^n + \frac{pn}{m}\,\delta_m^{n-1}.$$

Multiplying this result by m gives the desired result. □

LEMMA 4.2. *Let* $m, n \in \mathbb{N} \cup \{0\}$, *then*

$$2ph\frac{d}{dh}\,\delta_m^n = m(\delta_m^n - \delta_{m-2}^{n+2}).$$

Proof. Introducing the value of H in (2.10) as a variable and differentiating (2.10) with respect to h gives

$$\frac{\partial v}{\partial h} = \frac{v^2 - u^2}{2pvh}. \tag{4.4}$$

Therefore

$$\begin{aligned}
\frac{d}{dh}\,\delta_m^n &= \frac{d}{dh}\int_{\gamma(h)} \frac{u^n v^m}{(1-2u)^{p-1/p}}\, du \\
&= m\int_{\gamma(h)} \frac{u^n v^{m-1}}{(1-2u)^{p-1/p}}\frac{\partial v}{\partial h}\, du = \frac{m}{2ph}\int_{\gamma(h)} \frac{u^n v^{m-2}}{(1-2u)^{p-1/p}}(v^2-u^2)\, du \\
&= \frac{m}{2ph}(\delta_m^n - \delta_{m-2}^{n+2}).
\end{aligned}$$

□

LEMMA 4.3. $p/(2p+1)\delta_1^0 = \delta_1^1$.

Proof. Combining Lemma 4.2 with $R(1,1)$ gives

$$2ph\frac{d}{dh}\delta_1^0=\delta_1^0-\delta_{-1}^2,$$

$$2ph\frac{d}{dh}\delta_1^1=-\frac{p}{2p+1}\,\delta_1^0+2\delta_1^1-\frac{p}{2p+1}\,\delta_{-1}^2.$$

Therefore,

$$2ph\frac{d}{dh}\left(\frac{p}{2p+1}\,\delta_1^0-\delta_1^1\right)=2\left(\frac{p}{2p+1}\delta_1^0-\delta_1^1\right).$$

At $h=h_H$ both δ_1^0 and δ_1^1 vanish. This implies the result. □

LEMMA 4.4. *Let $R=R(h)=-\delta_{-1}^3$. Then*

$$(4.5)\qquad \begin{aligned}2ph\frac{d}{dh}\,\delta_1^1&=\delta_1^1+R,\\ 2ph\frac{d}{dh}\,\delta_1^2&=-\frac{2p}{2p+1}\,\delta_1^1+\frac{6p+2}{2p+1}\,\delta_1^2+\frac{p}{2p+1}\,R.\end{aligned}$$

Proof. Combine Lemma 4.2 with $R(2,1)$, $R(1,1)$. □

To simplify notation we will write

$$\alpha_0=\delta_1^1,\qquad \alpha_1=\delta_1^2.$$

Note that the differential equation (4.5), which α_0, α_1 satisfy, fits into the framework of the previous section with $q(h)=2ph$. As in § 2 the quotient of α_1 and α_0 will be denoted by Q.

The next lemma gives some useful properties of α_0 and Q.

LEMMA 4.5. (See Definition 2.1 for I_p).

(i) *For $h\in I_p$, $\operatorname{sign}(p)\alpha_0(h)\geqq 0$;*

(ii) *for $h\in I_p$, $\dot\alpha_0(h)\leqq 0$;*

(iii) $\operatorname{sign}(p)\dfrac{d}{dh}Q(h_H)>0$;

(iv) *for each $p\in(0,\infty)$ and $h\in[0,h_H]$:*

$$\frac{3p}{6p+2}=Q(0)\geqq Q(h)\geqq Q(h_H)=\frac{p}{2p+1};$$

(v) *for each $p\in(-\frac{1}{3},0)$ and $h\in[h_H,\infty)$:*

$$\frac{p}{2p+1}=Q(h_H)\geqq Q(h)\geqq\frac{3p}{6p+2}=Q(\infty);$$

(vi) *for each $p\in(-\frac{1}{2},-\frac{1}{3}]$ and $h\in[h_H,\infty)$:*

$$\frac{p}{2p+1}=Q(h_H)\geqq Q(h)\geqq-\infty=Q(\infty);$$

(vii) *suppose that $\dot Q(\bar h)=0$ for some $\bar h\in I_p$, then at $h=\bar h$:*

$$(4.6)\qquad \frac{\bar h\alpha_0}{(p/2p+1)-Q}\,\ddot Q=\bar h\ddot\alpha_0+\left(1-\frac{\bar h\dot\alpha_0}{\alpha_0}\right)\dot\alpha_0.$$

Proof. Recall the definition of α_0:

$$\alpha_0 = \int_{\gamma(h)} \frac{uv}{(1-2u)^{p-1/p}}\, du,$$

where $\gamma(h)$ is the level curve $H^{-1}(h), h \in I_p$ (see Fig. 2). Observe that $\operatorname{sign}(u) = \operatorname{sign}(p), 1-2u \geqq 0$. This gives (i). By Lemma 4.2

$$2ph\dot{\alpha}_0(h) = \int_{\gamma(h)} \frac{u(v^2-u^2)}{v(1-2u)^{p-1/p}}\, du.$$

On $\gamma(h)$, $v^2-u^2 \leqq 0$. This gives (ii). Combining Lemma 3.2 with Lemma 4.4 gives (iii).

When $p>0$,

$$\delta_1^n(0) = 2\int_0^{1/2} \frac{u^{n+1}}{(1-2u)^{p-1/p}}\, du = \left(\frac{1}{2}\right)^{n+1} \int_0^1 u^{n+1} \frac{1}{(1-u)^{p-1/p}}\, du$$
$$= \left(\frac{1}{2}\right)^{n+1} B(n+2, 1/p).$$

Therefore $Q(0) = 3p/(6p+2) = \Delta_1/\Delta_2$. The fact that the value of Q at the Hopf point h_H equals $p/(2p+1)$ is a consequence of Stokes' theorem. Consequently when $p>0$ all the hypotheses of Theorem 3.1 are satisfied and therefore (iv) holds.

Suppose that $-\frac{1}{2}<p<0$ and let $u_\pm(h)$ be the two real roots of the equation $u^2(1-2u)^{1/p} = h^{1/p}$. Then

$$\delta_1^n(h) = 2\int_{u_-(h)}^{u_+(h)} \frac{u^n\sqrt{u^2-(h^{1/p}/(1-2u)^{1/p})}}{(1-2u)^{p-1/p}}\, du.$$

Observe that $\int_{-\infty}^0 u^n(1-2u)^{1/p} = (\frac{1}{2})^{n+1}(-1)^n B(n+1), -1/p-n-1)$, which is convergent for $-1/p > n+1 > 0$. Furthermore, $u_-(h)$, $u_+(h)$ goes to $-\infty, 0$, respectively, as h goes to infinity. Thus we conclude that

$$\lim_{h\to\infty} Q(h) = \begin{cases} -\infty, & -\frac{1}{2}<p\leqq -\frac{1}{3}, \\ \dfrac{3p}{6p+2} = \dfrac{\Delta_1}{\Delta_2}, & -\frac{1}{3}<p<0. \end{cases}$$

So (v) is a consequence of Theorem 3.1. For the values of p between $-\frac{1}{2}$ and $-\frac{1}{3}$ we cannot directly apply Theorem 3.1. In this case we argue as follows. For $-\frac{1}{2}<p<-\frac{1}{3}$ we know that $\dot{Q}(h)=0$ implies that (compare (3.2))

$$(*) \qquad (Q-p/2p+1) = \frac{6p+2}{2p+1}\frac{\alpha_0}{2ph\dot{\alpha}_0}\left(Q-\frac{3p}{6p+2}\right)$$

where $((6p+2)/(2p+1))(\alpha_0/2ph\dot{\alpha}_0) > 0$. Suppose that $Q(\bar{h}) > p/2p+1$ for some $\bar{h} > h_H$. Then as $Q(\infty) = -\infty$ and $Q \leqq 0$ there exists a $\bar{h} > h_H$ such that $0 \geqq Q(\bar{h}) \geqq p/2p+1$ and $\dot{Q}(\bar{h}) = 0$. But this contradicts $(*)$. If $p = -\frac{1}{3}$ then (3.2) reads: $\dot{Q}(\bar{h}) = 0$ implies that

$$(Q-p/2p+1) = -\frac{\alpha_0}{2ph\dot{\alpha}_0}\frac{3p}{2p+1}.$$

Thus the same reasoning as above leads to the desired result. This completes (vi). The last assertion is a direct consequence of Theorem 3.3. □

It is our goal to show that the right-hand side of (4.6) is sign definite. To reach this goal we introduce a function $J(h)$ which serves a twofold purpose. First, $J(h)$ gives an h-dependent estimate of the quotient $h\dot{\alpha}_0/\alpha_0$; and second, $J(h)$ gives rise to an integral representation for $\ddot{\alpha}_0$.

LEMMA 4.6. *Suppose that* $\dot{Q}(\bar{h})=0$ *for some* $\bar{h}\in I$. *Then, at* $h=\bar{h}$, $\operatorname{sign}(p)\{\bar{h}\ddot{\alpha}_0+(1-(\bar{h}\dot{\alpha}_0/\alpha_0))\dot{\alpha}_0\}<0$.

COROLLARY 4.7. *Suppose that* $\bar{h}$ *is defined as in the lemma. Then*

$$\operatorname{sign}(p)\ddot{Q}(\bar{h})>0.$$

Proof of the corollary. Combine Lemma 4.6 with (4.6). □

Therefore, when Lemma 4.6 is proved, all the details in the proof of Theorem 2.1 have been given. Only very interested readers will appreciate the lengthy argument needed to establish Lemma 4.6.

Proof of Lemma 4.6.

Step 1. Let

$$J(h)=\int_{\gamma(h)}\{v^2(1-2u)^{1/p}-X(h)\}\frac{uv}{(1-2u)^{p-1/p}}\,du, \tag{4.7}$$

where

$$X(h)=\max_{\gamma(h)} v^2(1-2u)^{1/p}. \tag{4.8}$$

J has some useful properties.

(i) $\operatorname{sign}(p)J(h)\leqq 0,$

(ii) $\dot{J}(h)\geqq 0,$

(iii) $\dot{J}(h)=-(h^{1/p}/2ph)\alpha_0-X(h)\dot{\alpha}_0.$

(i) is clear from the definition of J. By construction of J the factor inside the curly brackets is independent of h. More precisely, from (2.10) and (4.8) it follows that

$$v^2(1-2u)^{1/p}-X(h)=u^2(1-2u)^{1/p}-C(p) \tag{4.9}$$

where

$$C(p)=\left(\frac{p}{2p+1}\right)^2\left(\frac{1}{2p+1}\right)^{1/p}=\max_{u\in\gamma(h)} u^2(1-2u)^{1/p}. \tag{4.10}$$

Therefore the derivative of J is

$$\dot{J}=\frac{1}{2ph}\int_{\gamma(h)}\{v^2(1-2u)^{1/p}-X(h)\}\frac{u(v^2-u^2)}{v(1-2u)^{p-1/p}}\,du, \tag{4.11}$$

and so (ii) is clear. To obtain (iii) we replace $(v^2-u^2)(1-2u)^{1/p}$ by $-h^{1/p}$ in the first term of the right-hand side of (4.11). As a consequence we get

$$\operatorname{sign}(p)\left\{h\ddot{\alpha}_0+\left(1-\frac{h\dot{\alpha}_0}{\alpha_0}\right)\dot{\alpha}_0\right\}\leqq\operatorname{sign}(p)\left\{h\ddot{\alpha}_0+\left(1+\frac{h^{1/p}}{2pX(h)}\right)\dot{\alpha}_0\right\}. \tag{4.12}$$

Step 2. Before we compute the second derivative of J to obtain an integral representation for $\ddot{\alpha}_0$ we integrate J by parts as follows: From (4.3), (4.7) and (4.9) we may write

$$J=\int\frac{u^2(1-2u)^{1/p}-C}{((2p+1)u^2-pu)(1-2u)^{p-1/p}}\,uv^2\left(v-pv\frac{\partial v}{\partial u}(1-2u)\right)du. \tag{4.13}$$

Let

$$\varphi(u)=\frac{u^2(1-2u)^{1/p}-C}{((2p+1)u-pu)(1-2u)^{p-1/p}}.$$

Then integrating (4.13) by parts gives

$$J=\int \mathcal{F}(u)v^3\,du, \tag{4.14}$$

where

$$\mathcal{F}(u)=(1-2p/3)\varphi(u)+p/3\dot\varphi(u)(1-2u). \tag{4.15}$$

The integral representation (4.14) differs in an essential way from the one given by (4.7). Because v is raised to the power three in (4.14), we may differentiate J twice without running into troubles with divergent integrals.

Step 3. We will prove that

$$\begin{aligned} h\ddot\alpha_0+\left(1+\frac{h^{1/p}}{2pX(h)}\right)\dot\alpha_0=\frac{1}{4p^2hX(h)}&\left\{\int_{\gamma(h)}\frac{(v^2-u^2)u}{v(1-2u)((2p+1)u^2-pu)^2}\right\}\\ &\cdot\left[2Cu^2(1-2u)^{1/p}((2p+1)u-p)^2\right.\\ &\left.+3h^{1/p}\frac{\mathcal{F}(u)}{u}((2p+1)u^2-pu)^2(1-2u)^{p-1/p}\right]du. \end{aligned} \tag{4.16}$$

From the definition of J we obtain

$$2phX(h)\ddot\alpha_0=(1-1/p)\frac{h^{1/p}}{h}\alpha_0+h^{1/p}\dot\alpha_0-2ph\ddot J.$$

Consequently,

$$\begin{aligned} h\ddot\alpha_0+\left(1+\frac{h^{1/p}}{2pX(h)}\right)\dot\alpha_0&=\frac{1}{2pX(h)}\left\{(1-1/p)\frac{h^{1/p}}{h}\alpha_0+(2pX(h)+2h^{1/p})\dot\alpha_0-2ph\ddot J\right\}\\ &=\frac{1}{2pX(h)}\{(2X(h)+2h^{1/p})\dot\alpha_0+2\dot J-2p\dot J-2ph\ddot J\}. \end{aligned}$$

In the last step we have expressed α_0 in terms of $\dot\alpha_0$ and $\dot J$ using property (iii) of J (Step 1). Differentiating (4.13) twice and inserting the result in the above identity gives (4.16). This completes Step 3.

Step 4. We will prove that the integrand in (4.16) is sign definite. Let

$$P_3(u)=((2p+1)u-p)^2,\quad W(u)=(1-2u)^{1/p},\quad V(u)=u^2W, \tag{4.17}$$

and let

$$I=2CVP_3+3h^{1/p}\mathcal{F}(u)uP_3W^{p-1}; \tag{4.18}$$

then (4.16) reads

$$h\ddot\alpha_0+\left(1+\frac{h^{1/p}}{2pX(h)}\right)\dot\alpha_0=\frac{1}{4p^2X(h)}\int_{\gamma(h)}\frac{(v^2-u^2)u}{v(1-2u)((2p+1)u^2-pu)^2}I(u)\,du. \tag{4.19}$$

Observe that on $\gamma(h)$

$$\operatorname{sign}\left\{\frac{(v^2-u^2)u}{v(1-2u)((2p+1)u^2-pu)^2}\right\}=-\operatorname{sign}(p). \tag{4.20}$$

If we let

$$P_1(u) = (-2p-1)u^2 + (2p^2+p)u - p^2,$$
$$P_2(u) = (-8p^2-6p-1)u^2 + (6p^2-3p)u - p^2,$$

then by a straightforward computation it follows that

$$3\mathscr{F}(u)uP_3W^{p-1} = VP_1 + CP_2.$$

Consequently we may rewrite (4.18) as

$$I = 2CVP_3 + h^{1/p}VP_1 + h^{1/p}CP_2.$$

We can still simplify this expression by observing that the relation $P_1 + P_2 = -2P_3$ holds, which enables us to rewrite I as

$$I = 2CP_3(V - h^{1/p}) + h^{1/p}P_1(V - C(p)). \tag{4.21}$$

This relation is nice because on $\gamma(h)$ both $V - h^{1/p}$ and $C - V$ are positive (see (2.10), (4.10)). We will finish the last step by proving:

PROPOSITION. *For $p > -\frac{1}{2}$ and $u \in (-\infty, \frac{1}{2}]$, $P_1(u) \leqq 0$.*

Recall that

$$P_1(u) = (-2p-1)u^2 + (2p^2+p)u - p^2.$$

Thus for $p > -\frac{1}{2}, p_1$ has a global maximum at $u = p/2$, and $P_1(p/2) = (p^2/4(2p+1))(p+\frac{1}{2})(4p-6)$. Therefore, for $p \in (-\frac{1}{2}, \frac{2}{3})$, P_1 is nonpositive on $\mathbb{R}$. $P_1(\frac{1}{2}) = -\frac{1}{4}$ and for $p > \frac{2}{3}$ the maximum of P_1 lies to the right of $\frac{1}{2}$. This proves the proposition.

Combining this with (4.21) finishes the last step. □

Note. After the completion of this manuscript, the authors received a paper of Zholondek [14] which deals with the same problem. His methods differ completely from ours and he solves the problem without any restriction.

REFERENCES

[1] V. I. ARNOLD, *Lectures on bifurcations in versal families*, Russian Math. Surveys, 27 (1972), pp. 54-123.

[2] ———, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[3] R. I. BOGDANOV, *Versal deformation of a singularity of a vector field on the plane in the case of zero eigenvalues*, Selecta Math. Soviet., 1 (1981), pp. 389-421.

[4] H. BROER AND G. VEGTER, *Subordinate Sil'nikov bifurcations near some singularities of vector fields having low codimension*, Ergodic Theory and Dynamical Systems, 4 (1984), pp. 509-525.

[5] J. CARR, S-N. CHOW AND J. K. HALE, *Abelian integrals and bifurcation theory*, J. Differential Equations, 59 (1985), pp. 413-436.

[6] S-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Grundlehren 251, New York, 1982.

[7] S. A. VAN GILS, *A note on: "Abelian integrals and bifurcation theory"*, J. Differential Equations, 59 (1985) pp. 437-441.

[8] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear oscillations, dynamical systems and bifurcation of vector fields*, in Applied Math. Sciences, Springer-Verlag, New York, 1983.

[9] G. IOOSS AND W. F. LANGFORD, *Conjecture on the routes to turbulence via bifurcation*, in Nonlinear Dynamics, R. H. G. Helleman, ed., New York Academy of Sciences, New York, 1980, pp. 489-505.

[10] W. F. LANGFORD, *Periodic and steady mode interactions lead to tori*, SIAM J. Appl. Math., 37 (1979), pp. 22-48.

[11] J. A. SANDERS AND R. H. CUSHMAN, *A codimension two bifurcation with third order Picard-Fuchs equation*, J. Differential Equations, 59 (1985), pp. 243-256.

[12] F. TAKENS, *Forced oscillations and bifurcations*, Communications of the Mathematical Institute RUU, 3 (1974), pp. 1-59.

[13] K. ZHOLONDEK, *On the versality of a family of symmetric vector-fields in the plane*, Math. USSR-Sb. 48 (1984), pp. 463-492.

[14] ———, *Bifurcation of a certain family of vectorfields on the plane tangent to the axes*, preprint.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

002

PERIODIC ORBITS IN SLOWLY VARYING OSCILLATORS*

STEPHEN WIGGINS† AND PHILIP HOLMES‡

Abstract. We develop a global perturbation technique for the study of periodic orbits in three-dimensional, time dependent and independent, perturbations of planar Hamiltonian differential equations. We give existence, stability and bifurcation theorems and illustrate our results with examples that exhibit saddle-node and Hopf bifurcations of periodic orbits.

Key words. bifurcation, Hamiltonian system, periodic orbit, Melnikov method, perturbation theory

AMS(MOS) classification numbers. 34CXX, 58F14, 70KXX

1. Introduction. In this and the following paper we develop tools for the study of periodic and homoclinic orbits occurring in a class of third order systems that arise in mechanical and other applications. We call these systems *slowly varying oscillators*; they take the general form

$$\left.\begin{aligned} \dot{x} &= f_1(x, y, z) + \varepsilon g_1(x, y, z, t; \boldsymbol{\mu}) \\ \dot{y} &= f_2(x, y, z) + \varepsilon g_2(x, y, z, t; \boldsymbol{\mu}) \\ \dot{z} &= \varepsilon g_3(x, y, z, t; \boldsymbol{\mu}) \end{aligned}\right\} \quad \text{or} \quad \dot{\mathbf{q}} = \mathbf{f}(\mathbf{q}) + \varepsilon \mathbf{g}_{\boldsymbol{\mu}}(\mathbf{q}, t) \tag{1.1}$$

where the g_i are T-periodic in t and depend upon parameter(s) $\boldsymbol{\mu} \in \mathbb{R}^k$. Additional structural assumptions on the unperturbed and perturbed phase spaces are given in § 2.

Earlier work on systems with slowly varying parameters includes that of Baker et al. [1971] and Marzec and Spiegel [1980], who pointed out that "strange attractors" similar to the Hénon attractor (Hénon [1976]) apparently arise from the Poincaré maps of such systems. In particular, Marzec and Spiegel [1980] studied two systems with fourth order potentials, similar to (1.2), below. Robinson [1983] and Robbins [1979] also worked on three-dimensional systems with a slow dependent variable; in fact the famous Lorenz model (Lorenz [1963]) can be put into such a form for high Rayleigh numbers (also see Sparrow [1982]). However, our immediate reason for studying such systems is that they occur as models of simple nonlinear elastic structures subject to feedback control when there is a nonnegligible time constant in the control process. (In a "rigid" control system the parameter ε would be very large, so that a singularly perturbed system would result.) See Sparrow [1981], Holmes [1983], [1985], and Moon and Rand [1984] for examples.

An example of such a system is the Duffing equation with slowly varying stiffness and weak linear dissipation:

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - x^3 - z - \varepsilon\delta y, \\ \dot{z} &= \varepsilon(-\alpha z + g(x, y)), \end{aligned} \tag{1.2}$$

* Received by the editors October 10, 1984; accepted for publication (in revised form) April 22, 1986. This research was supported in part by the Air Force Office of Scientific Research under AFOSR 84-0051.

† Department of Applied Mathematics, California Institute of Techriology, Pasadena, California 91125.

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

where g is a feedback control function. F. Moon's numerical integration of a fifth order system similar in form to (1.2) revealed chaotic solutions that appear to wind back and forth among a set of unstable periodic orbits, the shapes of which approximate those of certain orbits of the corresponding unperturbed Hamiltonian system. Baker et al. [1971] had made similar observations considerably earlier.

Our main analytical tool is a three-dimensional generalization of the computational method of Melnikov [1963]. The method utilizes the integrable structure of the unperturbed phase space as a framework on which to construct tools for the analysis of the perturbed phase space. See Greenspan and Holmes [1983] and Guckenheimer and Holmes [1983, Chap. 4] for details of the two-dimensional theory. Generalizations to $2n$-dimensional Hamiltonian systems have been made (Holmes and Marsden [1982a], [1982b], [1983]), as well as to homoclinic orbits for general n-dimensional systems (Gruendler [1985]).

The paper is organized as follows: in § 2 we describe the geometrical structure of the phase space and thus motivate our construction of the Melnikov functions. In § 3 we show that zeros of Melnikov functions imply the existence of periodic orbits; in §§ 4 and 5 we give stability and bifurcation results. In § 6 we illustrate our results with two examples. In the companion paper (Wiggins and Holmes [1987]) we consider the related problem of homoclinic orbits in slowly varying oscillators. Although, for simplicity, we restrict ourselves to three-dimensional systems in these papers, our methods generalize to arbitrary perturbations of $2n$-dimensional slowing varying Hamiltonian systems (Wiggins [1986]).

2. The phase space of slowly varying oscillators. We consider systems of the form

$$\left.\begin{aligned} \dot{x} &= f_1(x, y, z) + \varepsilon g_1(x, y, z, t; \boldsymbol{\mu}) \\ \dot{y} &= f_2(x, y, z) + \varepsilon g_2(x, y, z, t; \boldsymbol{\mu}) \\ \dot{z} &= \varepsilon g_3(x, y, z, t; \boldsymbol{\mu}) \end{aligned}\right\} \quad \text{or} \quad \dot{\mathbf{q}} = \mathbf{f}(\mathbf{q}) + \varepsilon \mathbf{g}(\mathbf{q}, t; \boldsymbol{\mu}), \tag{2.1}$$

with $0 < \varepsilon \ll 1$, $\mathbf{f}$ and $\mathbf{g}$ sufficiently smooth (C^r, $r \geqq 4$), $\mathbf{g}$ periodic in t with period T and $\boldsymbol{\mu} \in R^k$ a vector of parameters. We write $\mathbf{g}(\mathbf{q}, t; \boldsymbol{\mu}) = \mathbf{g}_{\boldsymbol{\mu}}(\mathbf{q}, t)$ and frequently drop the explicit dependence on $\boldsymbol{\mu}$. We make the following assumptions on the unperturbed system.

(A1) For $\varepsilon = 0$, (2.1) reduces to a one parameter family of planar Hamiltonian systems with Hamiltonian $H(x, y; z)$,

$$\begin{aligned} \dot{x} &= f_1(x, y, z) = \frac{\partial H}{\partial y}, \\ \dot{y} &= f_2(x, y, z) = -\frac{\partial H}{\partial x}, \\ &(\dot{z} = 0). \end{aligned} \tag{2.2}$$

(A2) For each value of z in some open interval $J \subset \mathbb{R}$ the "planar" system (2.2) possesses a one parameter family of periodic orbits, $q^{\alpha,z}(t-\theta)$, $\alpha \in L(z)$, where $L(z)$ is an open interval in $\mathbb{R}$ and θ denotes the "phase" or starting point of the orbit. We denote the period of $q^{\alpha,z}(t-\theta)$ by $T(\alpha, z)$ and assume that it is a differentiable function of α and z. Thus, when viewed in the full x, y, z phase, system (2.2) possesses a smooth family of invariant cylinders. (See Fig. 1.)

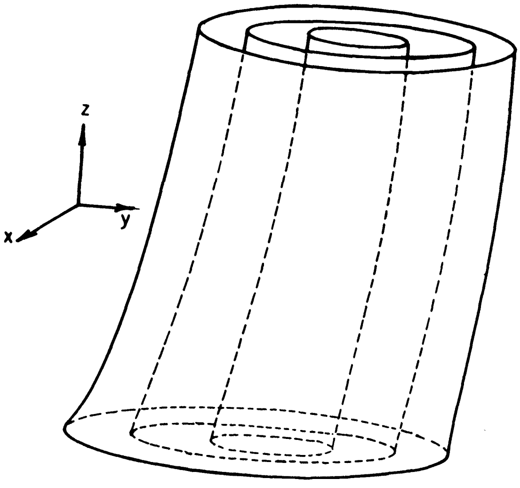


FIG. 1. *Structure of the unperturbed phase space.*

It will be convenient to think of (2.1) as an autonomous differential equation. This is accomplished by defining the function $\phi(t)=t$, mod T; by T-periodicity of the g_i we then have

$$
\begin{aligned}
&\dot{x}=f_1(x, y, z)+\varepsilon g_1(x, y, z, \phi), \\
&\dot{y}=f_2(x, y, z)+\varepsilon g_2(x, y, z, \phi), \qquad (x, y, z, \phi)\in R^3\times S^1, \\
&\dot{z}=\varepsilon g_3(x, y, z, \phi), \\
&\dot{\phi}=1.
\end{aligned} \tag{2.3}
$$

We remark that this suspension makes sense even when the g_i are independent of ϕ, although it then becomes trivial. The main point to note is that all the results developed below in the context of ϕ-dependent perturbations also hold good for ϕ-independent (i.e., autonomous) perturbations, although there are sometimes differences in interpretation.

The following perturbation results will be useful. Consider a subset of the two parameter family of periodic orbits whose period is uniformly bounded above. Let $\tilde{L}(z)\subset L(z)$ denote the set of α such that on a fixed $z=\text{constant}$ plane the periods $T(\alpha, z)$ of the periodic orbits are uniformly bounded above, say by a constant K.

PROPOSITION 2.1. *Let $\mathbf{q}_0^{\alpha,z}(t-\theta_0)$ be a periodic orbit of the unperturbed system with period $T(\alpha, z)<K$. Then there exists a perturbed orbit $\mathbf{q}_\varepsilon^{\alpha,z}(t, \theta)$, not necessarily periodic, which can be expressed as*

$$\mathbf{q}_\varepsilon^{\alpha,z}(t, \theta)=\mathbf{q}_0^{\alpha,z}(t-\theta)+\varepsilon\mathbf{q}_1^{\alpha,z}(t, \theta)+\mathcal{O}(\varepsilon^2) \tag{2.4}$$

uniformly in $t\in[t_0, t_0+T(\alpha, z)]$ for ε sufficiently small and all $\alpha\in\tilde{L}(z)$.

Remarks. This result follows directly from regular perturbation theory and Gronwall estimates (e.g., Hartman [1964]). The restriction to $\tilde{L}(z)$ avoids problems that arise when periodic orbits limit on homoclinic orbits and the period becomes unbounded. Furthermore, $\mathbf{q}_0^{\alpha,z}(t, \theta)$ may be found by solving the first variational equation

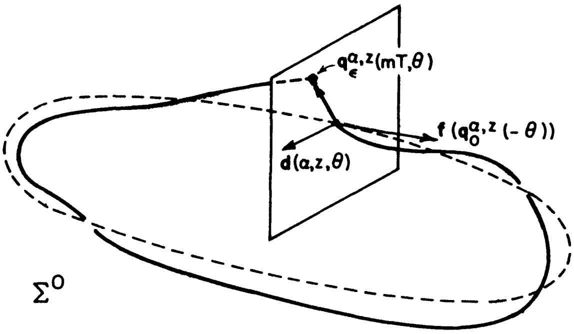
$$\dot{\mathbf{q}}_1^{\alpha,z}=\mathbf{Df}(\mathbf{q}_0^{\alpha,z})\mathbf{q}_1^{\alpha,z}+\mathbf{g}(\mathbf{q}_0^{\alpha,z}, t). \tag{2.5}$$

The main question which we ask is the following: Do any of the two parameter family of periodic orbits in the unperturbed system remain in the perturbed system? In order to answer this question we will reduce the study of the four-dimensional problem (2.3) to a three-dimensional Poincaré map. We define a global cross section transverse to the vector field (2, 3)

$$\Sigma^0 = \{(x, y, z, \phi) | \phi = 0\}$$

and define a mapping of Σ^0 into itself by letting orbits which start on Σ^0 evolve for time T until they return to Σ^0. Thus fixed points of the map correspond to periodic orbits of the vector field. (Note: the construction of the cross section and associated Poincaré map is different when the perturbation is autonomous, for then the problem is three-, not four-dimensional. However, we will see that formally the calculations are identical in both cases.) Although our Poincaré map is three-dimensional we will see that extending an idea due to Melnikov [1963] enables us to eliminate one component.

We now describe the geometric and intuitive notions which motivate the construction of the Melnikov functions. The structure of the unperturbed phase space will form the framework for the analysis of the system. Since the unperturbed system is autonomous, on each cross section, Σ^0, we have identical copies of the unperturbed phase space. In addition, the system possesses two constants of the motion, $H(x, y; z)$ and z itself. These constants are used to construct a moving system of "orbit coordinates" along the unperturbed orbits on the cross section Σ^0. Taking the gradient of the two constants of the motion we obtain the vectors $(\partial H/\partial x, \partial H/\partial y, 0)$ and $(0, 0, 1)$, which span a plane, Π. These vectors are evaluated on the unperturbed orbits on the cross section Σ^0, denoted $\mathbf{q}_0^{\alpha,z}(-\theta)$. Thus varying θ moves our two-dimensional plane Π around the unperturbed orbits on the cross section Σ^0 (see Fig. 2).



$q_\varepsilon^{\alpha,z}(mT,\theta)$
$f(q_0^{\alpha,z}(-\theta))$
$d(\alpha,z,\theta)$
Σ^0

FIG. 2. *The orbit coordinate system on* Σ^0.

In order to study periodic orbits in resonance with a time-periodic perturbation, we introduce the following two component vector valued function:

$$\mathbf{d}(\alpha, \theta, z) = \frac{\left([q_\varepsilon^{\alpha,z}(mT, \theta) - q_\varepsilon^{\alpha;z}(0, \theta)] \cdot \left(\frac{\partial H}{\partial x}, \frac{\partial H}{\partial y}, 0\right), [q_\varepsilon^{\alpha,z}(mT, \theta) - q_\varepsilon^{\alpha,z}(mT, \theta)] \cdot (0, 0, 1)\right)}{\|f(q_0^{\alpha,z}(-\theta))\|}$$

$$(2.6) \qquad = \frac{\varepsilon\left([q_1^{\alpha,z}(mT, \theta) - q_1^{\alpha,z}(0, \theta)] \cdot \left(\frac{\partial H}{\partial x}, \frac{\partial H}{\partial y}, 0\right), [q_1^{\alpha,z}(mT, \theta) - q_1^{\alpha,z}(mT, \theta)] \cdot (0, 0, 1)\right)}{\|f(q_0^{\alpha,z}(-\theta))\|} + \mathcal{O}(\varepsilon^2)$$

$$\overset{\text{def}}{=} \varepsilon \mathbf{M}^{m/n}(\alpha, \theta, z) + \mathcal{O}(\varepsilon^2),$$

where $\mathbf{M}^{m/n}(\alpha, \theta, z) \equiv (M_1^{m/n}(\alpha, \theta, z), M_3^{m/n}(\alpha, \theta, z))$ is defined to be the *subharmonic Melnikov function*, "$\| \ \|$" denotes the Euclidean norm, "$\cdot$" is the vector dot product, and $mT = nT(\alpha, z)$, where $T(\alpha, z)$ is the period of the unperturbed periodic orbit $\mathbf{q}_0^{\alpha,z}(t-\theta)$ and m, n are relatively prime integers.

The intuition behind the definition of $\mathbf{d}(\alpha, \theta, z)$ is the following: on the cross section Σ^0 we have the plane Π normal to the vector $\mathbf{f}(\mathbf{q}_0^{\alpha,z}(-\theta))$ at the point $\mathbf{q}_0^{\alpha,z}(-\theta)$. If a perturbed orbit has its initial value on this plane and evolves for time mT, after which it has returned to within $\mathcal{O}(\varepsilon)$ of the plane Π, $M^{m/n}$ is a measure of the "push" or "shear" the orbit undergoes due to the perturbation in directions transverse to the unperturbed vector field. Thus we expect that if the "push" is zero at some point, an orbit of period mT/n will be preserved in the perturbed flow. These heuristic notions are made precise in § 3, where we show that nondegenerate zeros of $\mathbf{M}^{m/n}$ correspond to isolated fixed points of a Poincaré mapping of $\Sigma^0 \to \Sigma^0$. Another important consequence of the definition of $\mathbf{M}^{m/n}$ is the fact that an explicit, computable expression may be obtained for $\mathbf{M}^{m/n}$.

$$\mathbf{M}^{m/n}(\alpha, \theta, z) = \frac{1}{\|\mathbf{f}(\mathbf{q}_0^{\alpha,z}(-\theta))\|} \left[\int_0^{mT} \left(\frac{\partial H}{\partial x} g_1 + \frac{\partial H}{\partial y} g_2 + \frac{\partial H}{\partial z} g_3 \right) (\mathbf{q}_0^{\alpha,z}(t), t+\theta) \, dt \right.$$

$$(2.7) \qquad - \frac{\partial H}{\partial z} \left(\mathbf{q}_0^{\alpha,z} \left(\frac{mT}{2} \right) \right) \int_0^{mT} g_3(\mathbf{q}_0^{\alpha,z}(t), t+\theta) \, dt,$$

$$\left. \int_0^{mT} g_3(\mathbf{q}_0^{\alpha,z}(t), t+\theta) \, dt \right],$$

where $q_0^{\alpha,z}(t-\theta)$ is an unperturbed periodic orbit of period $T(\alpha, z) = mT/n$. In the next section we derive this expression.

3. Periodic orbits. Here we study the two parameter family of periodic orbits and show how the subharmonic Melnikov function constructed in § 2 is related to a Poincaré map constructed from the perturbed vector field. We will restrict ourselves to a region where the periods are uniformly bounded above by a constant K (so Proposition 2.1 applies). This allows us to use the Hamiltonian structure of the unperturbed system to transform system (2.3) to action-angle variables (see Goldstein [1980] or Arnold [1978]):

$$(x(I, \theta), y(I, \theta), z) \to (I(x, y, z), \theta(x, y, z), z).$$

Under the transformation, (2.3) becomes

$$\begin{aligned}
\dot{I} &= \varepsilon \left(\frac{\partial I}{\partial x} g_1 + \frac{\partial I}{\partial y} g_2 + \frac{\partial I}{\partial z} g_3 \right) \equiv \varepsilon F(I, \theta, z, \phi), \\
\dot{\theta} &= \Omega(I, z) + \varepsilon \left(\frac{\partial \theta}{\partial x} g_1 + \frac{\partial \theta}{\partial y} g_2 + \frac{\partial \theta}{\partial z} g_3 \right) \equiv \Omega(I, z) + \varepsilon G(I, \theta, z), \\
\dot{z} &= \varepsilon g_3, \\
\dot{\phi} &= 1,
\end{aligned} \qquad (3.1)$$

with F, G, g_3 T-periodic in ϕ and where $\Omega(I, z) \equiv \partial H / \partial I$ is the angular frequency of the closed orbit in the unperturbed system on the $z = \text{constant}$ plane with action I and energy $H(I, z)$. So in the action-angle coordinate system the action I plays the role of the parameter α in our more general notation. We note that for $\varepsilon=0$ the solution to (3.1) is given by

$$\begin{aligned} I&=I_0,\\ \theta&=\Omega(I_0,z_0)(t-t_0)+\theta_0,\\ z&=z_0. \end{aligned}\tag{3.2}$$

Now we will construct an approximation to the Poincaré map associated with system (3.1). The cross section to the flow defined by (3.1) is

$$\Sigma^0=\{(I,\theta,z,\phi)\,|\,\phi=0\},\tag{3.3}$$

and the mth iterate of the Poincaré map, P_ε^m, is

$$\begin{aligned} P_\varepsilon^m\colon\ &(I_\varepsilon(0,0,I_0,\theta_0,z_0),\theta_\varepsilon(0,0,I_0,\theta_0,z_0),z_\varepsilon(0,0,I_0,\theta_0,z_0))\\ &\to(I_\varepsilon(mT,0,I_0,\theta_0,z_0),\theta_\varepsilon(mT,0,I_0,\theta_0,z_0),z_\varepsilon(mT,0,I_0,\theta_0,z_0)), \end{aligned}\tag{3.4}$$

where the initial conditions are chosen such that

$$\begin{aligned} I_\varepsilon(0,0,I_0,\theta_0,z_0)&=I_0,\\ \theta_\varepsilon(0,0,I_0,\theta_0,z_0)&=\theta_0,\\ z_\varepsilon(0,0,I_0,\theta_0,z_0)&=z_0. \end{aligned}\tag{3.5}$$

Using Proposition 2.1 and the solutions to the unperturbed problem given by (3.2), we can approximate the Poincaré map using regular perturbation theory.

The solution of (3.1) can be written as

$$\begin{aligned} P_\varepsilon^m\colon\ &(I_0,\theta_0,z_0)\to(I_0,\theta_0,z_0)+(0,mT\Omega(I_0,z_0),0)\\ &+\varepsilon(I_1(mT,0,I_0,\theta_0,z_0),\theta_1(mT,0,I_0,\theta_0,z_0),z_1(mT,0,I_0,\theta_0,z_0))+\mathcal{O}(\varepsilon^2). \end{aligned}\tag{3.6}$$

Recall that this approximation is uniformly valid for one period of an unperturbed orbit, $T(I,z)=mT$. For the case of ultrasubharmonics, $T(I,z)=mT/n$, $n\geqq 2$, m, n relatively prime, the approximation is not uniformly valid since ε must shrink to zero as n increases.

We now compute I_1, θ_1, and z_1 by solving the first variational equation. Using (3.2) we obtain

$$\begin{pmatrix}\dot I_1\\ \dot\theta_1\\ \dot z_1\end{pmatrix}=\begin{pmatrix}0&0&0\\ \partial\Omega/\partial I_0|_{(I_0,z_0)}&0&\partial\Omega/\partial z_0|_{(I_0,z_0)}\\ 0&0&0\end{pmatrix}\begin{pmatrix}I_1\\ \theta_1\\ z_1\end{pmatrix}+\begin{pmatrix}F(I_0,\Omega(I_0,z_0)t+\theta_0,z_0,t)\\ G(I_0,\Omega(I_0,z_0)t+\theta_0,z_0,t)\\ g_3(I_0,\Omega(I_0,z_0)t+\theta_0,z_0,t)\end{pmatrix},\tag{3.7}$$

and consequently

$$I_1(mT,0,I_0,\theta_0,z_0)=\int_0^{mT}F(I_0,\Omega(I_0,z_0)t+\theta_0,z_0,t)\,dt\equiv\bar M_1^{m/n}(I_0,\theta_0,z_0),$$

$$\begin{aligned} \theta_1(mT,0,I_0,\theta_0,z_0)=&\left.\frac{\partial\Omega}{\partial I_0}\right|_{(I_0,z_0)}\int_0^{mT}\int_0^t F(I_0,\Omega(I_0,z_0)\xi+\theta_0,z_0,\xi)\,d\xi\,dt\\ &+\int_0^{mT}G(I_0,\Omega(I_0,z_0)t+\theta_0,z_0,t)\,dt\\ &+\left.\frac{\partial\Omega}{\partial z_0}\right|_{(I_0,z_0)}\int_0^{mT}\int_0^t g_3(I_0,\Omega(I_0,z_0)\xi+\theta_0,z_0,\xi)\,d\xi\,dt\\ \equiv&\ \bar M_2^{m/n}(I_0,\theta_0,z_0), \end{aligned}\tag{3.8}$$

$$z_1(mT, 0, I_0, \theta_0, z_0) = \int_0^{mT} g_3(I_0, \Omega(I_0, z_0)t + \theta_0, t)\, dt \equiv \bar{M}_3^{m/n}(I_0, \theta_0, z_0).$$

The Poincaré map P_ε^m becomes

$$\begin{aligned} P_\varepsilon^m\colon (I_0, \theta_0, z_0) &\to (I_0, \theta_0, z_0) + (0, \Omega(I_0, z_0)mT, 0) \\ &\quad + \varepsilon(\bar{M}_1^{m/n}(I_0, \theta_0, z_0), \bar{M}_2^{m/n}(I_0, \theta_0, z_0), \bar{M}_3^{m/n}(I_0, \theta_0, z_0)) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{3.9}$$

We will subsequently show that the first and third components of the $\mathcal{O}(\varepsilon)$ part of the Poincaré map make up the subharmonic Melnikov function which we defined in § 2, i.e., $\bar{M}_i^{m/n} = M_i^{m/n}$, $i = 1, 3$.

We define the vector $\bar{\mathbf{M}}^{m/n}$ as

$$\bar{\mathbf{M}}^{m/n}(I_0, \theta_0, z_0) = (\bar{M}_1^{m/n}(I_0, \theta_0, z_0), \bar{M}_2^{m/n}(I_0, \theta_0, z_0), \bar{M}_3^{m/n}(I_0, \theta_0, z_0)). \tag{3.10}$$

We remark that the superscript m/n denotes our search for periodic orbits (which we will frequently omit) which satisfy the resonance relation $T(I, z) = mT/n$, m, n relatively prime integers.

We now state our main theorem.

THEOREM 3.1. *Suppose* $(I_0^*, \theta_0^*, z_0^*)$ *is a point where* $T(I_0^*, z_0^*) = mT/n$ *and the following condition is satisfied*:

$$\left.\frac{\partial \Omega}{\partial I_0}\right|_{(I_0^*, z_0^*)} \neq 0 \quad \text{or} \quad \left.\frac{\partial \Omega}{\partial z_0}\right|_{(I_0^*, z_0^*)} \neq 0,$$

$$\left.\left[\frac{\partial \Omega}{\partial I_0}\left(\frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial z_0} - \frac{\partial \bar{M}_1}{\partial z_0}\frac{\partial \bar{M}_3}{\partial \theta_0}\right) + \frac{\partial \Omega}{\partial z_0}\left(\frac{\partial \bar{M}_1}{\partial I_0}\frac{\partial \bar{M}_3}{\partial \theta_0} - \frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial I_0}\right)\right]\right|_{(I_0^*, \theta_0^*, z_0^*)} \neq 0$$

and

$$\bar{M}_1(I_0^*, \theta_0^*, z_0^*) = \bar{M}_3(I_0^*, \theta_0^*, z_0^*) = 0.$$

Then for $0 < \varepsilon \leqq \varepsilon(n)$ *the Poincaré map*, P_ε^m, *has a fixed point of period m. If* $n = 1$ *the result is uniformly valid in* $0 < \varepsilon \leqq \varepsilon(1)$.

We remark that the conditions

$$\frac{\partial \Omega}{\partial I_0}\left(\frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial z_0} - \frac{\partial \bar{M}_1}{\partial z_0}\frac{\partial \bar{M}_3}{\partial \theta_0}\right) + \frac{\partial \Omega}{\partial z_0}\left(\frac{\partial \bar{M}_1}{\partial I_0}\frac{\partial \bar{M}_3}{\partial \theta_0} - \frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial I_0}\right) \neq 0$$

is a sufficient condition for the Poincaré map minus the identity map to be invertible in each case and for the fixed point to be isolated. We also call attention to the fact that no knowledge of $\bar{M}_2^{m/n}$ is needed. Thus evaluation of the double integrals of (3.8) is unnecessary.

Proof. We note that by hypothesis the resonance relation, $T(I_0, z_0) = mT/n$, is satisfied so that $mT\Omega(I_0^*, z_0^*) = 2\pi n = 0$ since $mT\Omega$ is an angular variable. For definiteness we assume $\partial\Omega/\partial z_0|_{(I_0^*, z_0^*)} \neq 0$ (the case with $\partial\Omega/\partial I_0|_{(I_0^*, z_0^*)} \neq 0$ is proved similarly). Let us perturb the point $(I_0^*, \theta_0^*, z_0^*)$ to $(I_0^*, \theta_0^*, z_0^* + \Delta z)$. Then at this point the Poincaré map takes the form

$$\begin{aligned} &\mathbf{P}_\varepsilon^m(I_0^*, \theta_0^*, z_0^* + \Delta z) - (I_0^*, \theta_0^*, z_0^* + \Delta z) \\ &\quad = \left(0, mT\left.\frac{\partial \Omega}{\partial z_0}\right|_{(I_0^*, z_0)} \Delta z + \varepsilon \bar{M}_2(I_0^*, \theta_0^*, z_0^*) + \mathcal{O}(\Delta z^2) + \mathcal{O}(\Delta z \varepsilon), 0\right) + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Now if we choose

$$\Delta z = -\varepsilon \frac{\bar{M}_2(I_0^*, \theta_0^*, z_0^*)}{mT\left.\dfrac{\partial \Omega}{\partial z_0}\right|_{(I_0^*, z_0^*)}},$$

we have

$$\mathbf{P}_\varepsilon^m(I_0^*, \theta_0^*, z_0^*+\Delta z) - (I_0^*, \theta_0^*, z_0^*+\Delta z) = \mathcal{O}(\varepsilon^2),$$

$$\det |\mathbf{DP}_\varepsilon^m - \mathbf{id}|\bigg|_{(I_0^*,\, \theta_0^*,\, z_0^*+\Delta z)} = \varepsilon^2 \left[\frac{\partial \Omega}{\partial I_0}\left(\frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial z_0} - \frac{\partial \bar{M}_1}{\partial z_0}\frac{\partial \bar{M}_3}{\partial \theta_0}\right) + \frac{\partial \Omega}{\partial z_0}\left(\frac{\partial \bar{M}_1}{\partial I_0}\frac{\partial \bar{M}_3}{\partial \theta_0} - \frac{\partial \bar{M}_1}{\partial \theta_0}\frac{\partial \bar{M}_3}{\partial I_0}\right)\right]\bigg|_{(I_0^*,\, \theta_0^*,\, z_0^*)} + \mathcal{O}(\varepsilon^3) \neq 0 \text{ by hypothesis.}$$

Thus, since $\|\mathbf{DP}_\varepsilon^m - \mathbf{id}\| = \mathcal{O}(1)$, the implicit function theorem allows us to conclude that the Poincaré map has a fixed point near $(I_0^*, \theta_0^*, z_0^*+\Delta z)$ and thus that the differential equation has an isolated periodic orbit of period mT/n near this point. □

We make the following remarks:

(1) Although P_ε^m is a diffeomorphism of $\mathbb{R}^3$, in order to determine whether or not it has fixed points we need only check the first and third components of the map. This is a result of the nonzero twist condition ($\partial\Omega/\partial I_0$ or $\partial\Omega/\partial z_0 \neq 0$), which insures locally that we return to the correct section after time mT.

(2) The advantage of using action-angle variables is that they enable us to explicitly relate the subharmonic Melnikov function to a Poincaré map by allowing us to easily solve the first variational equation (3.7), which might otherwise be intractable analytically. In §§ 4 and 5 this relationship is exploited by utilizing existing theorems concerning stability and bifurcations of maps in order to get similar theorems expressed entirely in terms of the subharmonic Melnikov function.

(3) Theorem 3.1 does not apply to the case of autonomous vector fields. In such cases it suffices to study a diffeomorphism of R^2 obtained by fixing $\theta = \theta_0$ and allowing the I and z variables with initial values at $\theta = \theta_0$ to evolve in time until they return to $\theta = \theta_0$. Hereafter we will delete the subscript 0 on I, θ, and z when there is no possibility of confusion. We have the following theorem.

THEOREM 3.2. *Suppose there exists a point* (I^*, z^*) *such that*

(a) $\bar{M}_1(I^*, z^*) = \bar{M}_3(I^*, z^*) = 0,$

(b) $\dfrac{\partial(\bar{M}_1, \bar{M}_3)}{\partial(I, z)}\bigg|_{(I^*, z^*)} \neq 0.$

Then $(I^*, z^*) + \mathcal{O}(\varepsilon)$ *is an isolated fixed point for the Poincaré map that corresponds to an isolated periodic orbit for the three-dimensional flow.*

We remark that M_1 and M_3 are defined exactly as in the nonautonomous case except that the limits of integration now become $0 \to T(I, z)$; hence we drop the superscript m/n.

Proof. The proof is very similar to that of Theorem 3.1. The details can be found in Wiggins [1985]. □

Finally we want to show that the subharmonic Melnikov function derived in this section using action-angle variables is identical to the expression that we gave in § 2 for an arbitrary coordinate system. We can easily see that $\bar{M}_3^{m/n}$ derived in action-angle variables is identical to the corresponding $M_3^{m/n}$ of § 2 derived using the orbit coordinate system, since the functions under the integrand are the same and both are evaluated on an unperturbed periodic orbit. It makes no difference whether or not the unperturbed orbit is expressed in (I, θ, z) coordinates or (x, y, z) coordinates, since the Jacobian of the transformation between the two coordinate systems is identically one and hence does not affect the integral. The fact that the $M_1^{m/n}$ are the same is not as obvious, but is nevertheless true. In action angle variables $M_1^{m/n}$ is defined as

$$\bar{M}_1^{m/n}(I,\theta,z)=\int_0^{mT}\left(\frac{\partial I}{\partial x}g_1+\frac{\partial I}{\partial y}g_2+\frac{\partial I}{\partial z}g_3\right)dt, \tag{3.11}$$

where the integrand is evaluated on an unperturbed periodic orbit expressed in action-angle variables. (Note: By $\partial I/\partial x$ we mean the partial derivative of I with y and z held fixed; we denote this by $\partial I/\partial x|_{y,z}$, similarly for $\partial I/\partial y$ and $\partial I/\partial z$. This will be important in the following.) Now using the action angle transformation we can write

$$H=H(I,z), \tag{3.12}$$

and since we have assumed that we are in a region where $\partial H/\partial I|_z$ is nonzero we can invert (3.12) to obtain

$$I=I(H,z). \tag{3.13}$$

Differentiating (3.13) we obtain

$$\begin{aligned}\left.\frac{\partial I}{\partial x}\right|_{y,z}&=\left.\frac{\partial I}{\partial H}\right|_z\left.\frac{\partial I}{\partial H}\right|_{y,z},\\ \left.\frac{\partial I}{\partial y}\right|_{x,z}&=\left.\frac{\partial I}{\partial H}\right|_z\left.\frac{\partial H}{\partial y}\right|_{x,z},\\ \left.\frac{\partial I}{\partial z}\right|_{x,y}&=\left.\frac{\partial I}{\partial H}\right|_z\left.\frac{\partial H}{\partial z}\right|_{x,y}+\left.\frac{\partial I}{\partial z}\right|_H,\end{aligned} \tag{3.14}$$

and by definition of the action-angle transformation we have

$$\left.\frac{\partial H}{\partial I}\right|_z=\Omega(I,z). \tag{3.15}$$

Substituting (3.14) and (3.15) into (3.11) we get

$$\bar{M}_1^{m/n}=\frac{1}{\Omega(I,z)}\int_0^{mT}\left(\frac{\partial H}{\partial x}g_1+\frac{\partial H}{\partial y}g_2+\frac{\partial H}{\partial z}g_3\right)dt+\left.\frac{\partial I}{\partial z}\right|_H\int_0^{mT}g_3\,dt, \tag{3.16}$$

where we have pulled the $\partial I/\partial z|_H$ term out of the integrand since it is constant on unperturbed periodic orbits. Recall that on an unperturbed periodic orbit we have

$$I=I(H,z)=\text{constant}. \tag{3.17}$$

So differentiating along this orbit we get

$$0=\left.\frac{\partial I}{\partial H}\right|_z\left.\frac{\partial H}{\partial z}\right|_I+\left.\frac{\partial I}{\partial z}\right|_H. \tag{3.18}$$

Using (3.18), (3.16) becomes

$$\bar{M}_1^{m/n}(I,\theta,z)=\frac{1}{\Omega(I,z)}\left[\int_0^{mT}\left(\frac{\partial H}{\partial x}g_1+\frac{\partial H}{\partial y}g_2+\frac{\partial H}{\partial y}g_3\right)dt-\left.\frac{\partial H}{\partial z}\right|_I\int_0^{mT}g_3\,dt\right], \tag{3.19}$$

where the integrand is evaluated on an unperturbed periodic orbit (note: as in the case of $M_3^{m/n}$, it does not matter whether this periodic orbit is expressed in (I,θ,z) coordinates or (x,y,z) coordinates). Finally, we note that in the action-angle coordinate system (see (3.1)) $\|\mathbf{f}(\mathbf{q}_0^{\alpha,z}(-\theta))\|=\Omega(I,z)$, so we see that the expression which we have derived here in action-angle coordinates (i.e., in a region where the periods of the unperturbed orbits are uniformly bounded) is identical to that obtained from orbit coordinates in § 2, so hereafter we will drop the overbars on $\bar{M}$, and $\bar{M}_3$.

4. Stability. We have shown that the study of system (2.1) can be reduced to the study of a three-dimensional Poincaré map in the nonautonomous case

$$P_\varepsilon^m(q)=q+(0, mT\Omega(I, z), 0)+\varepsilon(M_1^{m/n}(q), \bar{M}_2^{m/n}(q), M_3^{m/n}(q))+\mathcal{O}(\varepsilon^2), \quad q=(I, \theta, z) \tag{4.1}$$

and a two-dimensional Poincaré map in the autonomous case

$$P_\varepsilon^m(q)=q+\varepsilon(M_1(q), M_3(q))+\mathcal{O}(\varepsilon^2), \qquad q=(I, z), \tag{4.2}$$

and that nondegenerate fixed points of these maps correspond to isolated periodic orbits in the ordinary differential equations. We can compute the stability of these fixed points in the obvious way, namely linearize the map about the fixed point and examine the eigenvalues. After some computation, we find that the eigenvalues of (4.1) (the nonautonomous case) are given by the following cases.

Case 1. $\Delta_2\neq 0, \Delta_4\neq 0$.

$$\begin{aligned}\lambda_{1,2}&=1\pm\varepsilon^{1/2}\sqrt{\Delta_2}+\varepsilon/2\left(\Delta_1-\frac{\Delta_4}{\Delta_2}\right)+\mathcal{O}(\varepsilon^{3/2}),\\ \lambda_3&=1+\varepsilon\frac{\Delta_4}{\Delta_2}+\mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.3}$$

Case 2. $\Delta_2=0, \Delta_4\neq 0$.

$$\begin{aligned}\lambda_1&=1+\varepsilon^{2/3}(-\Delta_4)^{1/3}+\varepsilon\frac{\Delta_1}{3}+\frac{\varepsilon^{4/3}}{(-\Delta_4)^{1/3}}\left[\frac{5}{9}\Delta_1^2-\Delta_3\right]+\mathcal{O}(\varepsilon^{5/3}),\\ \lambda_2&=1+\varepsilon^{2/3}e^{4\pi i/3}(-\Delta_4)^{1/3}+\varepsilon\frac{\Delta_1}{3}+\frac{\varepsilon^{4/3}e^{4\pi i/3}}{(-\Delta_4)^{1/3}}\left[\frac{5}{9}\Delta_1^2-\Delta_3\right]+\mathcal{O}(\varepsilon^{5/3}),\\ \lambda_3&=1+\varepsilon^{2/3}e^{8\pi i/3}(-\Delta_4)^{1/3}+\frac{\varepsilon\Delta_1}{3}+\frac{\varepsilon^{4/3}e^{8\pi i/3}}{(-\Delta_4)^{1/3}}\left[\frac{5}{9}\Delta_1^2-\Delta_3\right]+\mathcal{O}(\varepsilon^{5/3}),\end{aligned} \tag{4.4}$$

where

$$\begin{aligned}\Delta_1&=\frac{\partial M_1}{\partial I}+\frac{\partial \bar{M}_2}{\partial \theta}+\frac{\partial M_3}{\partial z}=\text{trace}\,[\mathbf{DM}],\\ \Delta_2&=mT\frac{\partial\Omega}{\partial I}\frac{\partial M_1}{\partial\theta}+mT\frac{\partial\Omega}{\partial z}\frac{\partial M_3}{\partial\theta},\\ \Delta_3&=\frac{\partial(M_1, \bar{M}_2)}{\partial(I, \theta)}+\frac{\partial(M_1, M_3)}{\partial(I, z)}+\frac{\partial(\bar{M}_2, M_3)}{\partial(\theta, z)},\\ \Delta_4&=mT\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\theta, z)}+mT\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(I, \theta)},\\ \Delta_5&=\frac{\partial\bar{M}_2}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\theta, z)}+\frac{\partial\bar{M}_2}{\partial\theta}\frac{\partial(M_1, M_3)}{\partial(z, I)}+\frac{\partial\bar{M}_2}{\partial z}\frac{\partial(M_1, M_3)}{\partial(I, \theta)}=-\det\,[\mathbf{DM}].\end{aligned} \tag{4.5}$$

Here all partial derivatives are evaluated at the zero of M_1, M_3, and

$$\frac{\partial(M_1, \bar{M}_2)}{\partial(I, \theta)}=\frac{\partial M_1}{\partial I}\frac{\partial\bar{M}_2}{\partial\theta}-\frac{\partial M_1}{\partial\theta}\frac{\partial\bar{M}_2}{\partial I}, \text{ etc.}$$

denote Jacobian determinants. These two cases give all possible forms for the eigenvalues for the Poincaré map with $\partial\Omega/\partial I \neq 0$ or $\partial\Omega/\partial z \neq 0$. If $\Delta_1 = 0$, the eigenvalues still maintain the above forms and if $\Delta_4 = 0$ our existence theorem cannot be applied (we are at a bifurcation point). (Note: the computation of these expressions for the eigenvalues are not trivial; they involve some results from algebraic function theory along with the use of Newton diagrams (for complete details see Wiggins [1985]).) Finally, note that the eigenvalues are completely determined by M_1 and M_3 at the lowest order in ε (this is a consequence of the nonzero twist condition).

In the autonomous case the situation is simpler, the eigenvalues of (4.2) are given by

$$\lambda_{1,2} = 1 + \frac{\varepsilon}{2}\operatorname{tr} DM \pm \frac{\varepsilon}{2}\sqrt{(\operatorname{tr} DM)^2 - 4\det DM} + \theta(\varepsilon^2) \tag{4.6}$$

where

$$\operatorname{tr} DM = \frac{\partial M_1}{\partial I} + \frac{\partial M_3}{\partial z}, \qquad \det DM = \frac{\partial M_1}{\partial I}\frac{\partial M_3}{\partial z} - \frac{\partial M_1}{\partial z}\frac{\partial M_3}{\partial I}$$

and the partial derivatives are evaluated at a zero of (M_1, M_3).

Once the eigenvalues are in hand, stability follows from the standard linearization theory.

5. Bifurcations. We now address the question of bifurcations for parameterized families of systems and restrict ourselves to codimension one bifurcations (Guckenheimer and Holmes [1983], Arnold [1982]). Thus, if the parameter is multidimensional, we fix all but one component and vary that alone. We begin with a theorem concerning saddle-node bifurcations.

THEOREM 5.1. *Consider the parametrized Poincaré map with the parameter $\mu \in \mathbb{R}$. Suppose there exists a point $(I^*, \theta^*, z^*, \mu^*) = \mathbf{q}^*$ such that $mT\Omega(I^*, z^*) = 2\pi n$ and*

$$M_1^{m/n}(\mathbf{q}^*) = M_3^{m/n}(\mathbf{q}^*) = 0, \qquad \left[\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z, \theta)} + \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta, I)}\right]\Bigg|_{\mathbf{q}^*} = 0,$$

and one of the following holds:

(a) $$\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta, \mu)}\Bigg|_{\mathbf{q}^*} \neq 0, \qquad \frac{d}{dI}\left[\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z, \theta)} + \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta, I)}\right]\Bigg|_{\mathbf{q}^*} \neq 0;$$

(b) $$\left[\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\mu, z)} + \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(I, \mu)}\right]\Bigg|_{\mathbf{q}^*} \neq 0,$$

$$\frac{d}{d\theta}\left[\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z, \theta)} + \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta, I)}\right]\Bigg|_{\mathbf{q}^*} \neq 0;$$

(c) $$\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\mu, \theta)}\Bigg|_{\mathbf{q}^*} \neq 0, \qquad \frac{d}{dz}\left[\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z, \theta)} + \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta, I)}\right]\Bigg|_{\mathbf{q}^*} \neq 0.$$

Then near $\mathbf{q}^*$ there is a bifurcation point at which saddle-nodes of periodic orbits occur.

Proof. The equations

$$\begin{aligned} &M_1(I, \theta, z, \mu) + \mathcal{O}(\varepsilon) = 0, \\ &mT\Omega(I, z) - 2\pi n + \varepsilon\bar{M}_2(I, \theta, z, \mu) + \mathcal{O}(\varepsilon^2) = 0, \\ &M_3(I, \theta, z, \mu) + \mathcal{O}(\varepsilon) = 0 \end{aligned} \tag{5.1}$$

represent a curve in (I, θ, z, μ) space corresponding to fixed points of the Poincaré map P_ε^m. The tangent to this curve can be computed:

$$\begin{aligned}\mathbf{T}=\Bigg(&\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta,\mu)}+\mathcal{O}(\varepsilon), \frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(I,\mu)}+\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\mu,z)}+\mathcal{O}(\varepsilon),\\ &\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\mu,\theta)}+\mathcal{O}(\varepsilon), \frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z,\theta)}+\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta,I)}+\mathcal{O}(\varepsilon)\Bigg)^T.\end{aligned}\tag{5.2}$$

We must show that there exists a point near $\mathbf{q}^*=(I^*, \theta^*, z^*, \mu^*)$ such that

$$\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(z,\theta)}+\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta,I)}+\mathcal{O}(\varepsilon)=0$$

at this point and the rate of change of the μ component of the tangent to the curve at this point is nonzero. This will show that the curve of fixed points is locally parabolic in the μ direction about this point and takes one of two possible forms for nondegenerate saddle-node bifurcations (Fig. 3).

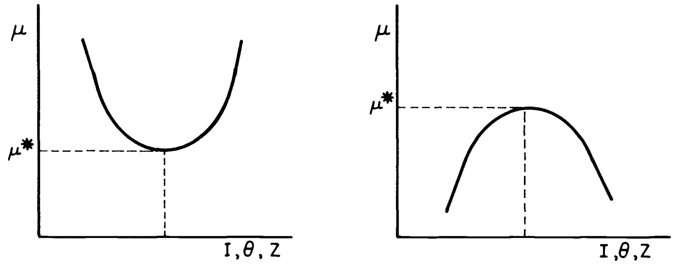


FIG. 3. *Locally parabolic bifurcation curves.*

Let us suppose condition (a) of FP1 holds. Then since $(\partial\Omega/\partial z)(\partial(M_1, M_3)/\partial(\theta, \mu))$ is nonzero at $\mathbf{q}^*$, the I component of the tangent to the curve is nonzero at this point. So by the implicit function theorem we can parametrize the curve of equilibria locally in terms of I. To show that there exists a nearby point such that

$$\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\theta,\mu)}+\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta,I)}+\mathcal{O}(\varepsilon)=0$$

we need only show that

$$\frac{d}{dI}\left(\frac{\partial\Omega}{\partial I}\frac{\partial(M_1, M_3)}{\partial(\theta,\mu)}+\frac{\partial\Omega}{\partial z}\frac{\partial(M_1, M_3)}{\partial(\theta,I)}\right)\Bigg|_{\mathbf{q}^*}\neq 0.$$

This follows by hypothesis and also implies that the rate of change of the μ component of the tangent to the curve of fixed points is nonzero at this point. This completes the proof of case (a). Cases (b) and (c) follow from similar arguments. □

The corresponding result for the autonomous case is the following.

THEOREM 5.2. *Consider the parametrized Poincaré map* (3.1) *with the parameter* $\mu\in\mathbb{R}^1$. *Suppose that there exists a point* $q^*=(I^*, z^*, \mu^*)$ *such that*

$$\mathbf{M}(\mathbf{q}^*)=0,\quad \det \mathbf{DM}|_{\mathbf{q}^*}=\frac{\partial(M_1, M_3)}{\partial(I,z)}\bigg|_{\mathbf{q}^*}=0,\quad \operatorname{tr}\mathbf{DM}|_{\mathbf{q}^*}\neq 0,$$

with at least one of the following conditions satisfied

$$(1)\qquad \left.\frac{\partial(M_1, M_3)}{\partial(I, \mu)}\right|_{\mathbf{q}^*} \neq 0, \qquad \frac{d}{dz}(\det \mathbf{DM})|_{\mathbf{q}^*} \neq 0;$$

$$(2)\qquad \left.\frac{\partial(M_1, M_3)}{\partial(z, \mu)}\right|_{\mathbf{q}^*} \neq 0, \qquad \frac{d}{dI}(\det \mathbf{DM})|_{\mathbf{q}^*} \neq 0.$$

Then $\mathbf{q} = \mathbf{q}^* + \mathcal{O}(\varepsilon)$ *is a bifurcation point at which saddle-nodes of periodic orbits occur.*

Proof. The proof is very similar to the proof of Theorem 5.1. □

We next prove a theorem concerning Hopf bifurcations for the Poincaré map. We utilize the notation given in (3.17).

THEOREM 5.3. *Let* $\mathbf{q}(\mu)$ *be a smooth curve of fixed points for the Poincaré map,* $\mu \in K$, *where* K *is some open interval in* $\mathbb{R}$. *Suppose there exists a* $\mu_0 \in K$ *such that*

$$\left.\frac{\partial \Omega}{\partial I}\right|_{\mathbf{q}(\mu_0)} \neq 0 \quad \mathit{or} \quad \left.\frac{\partial \Omega}{\partial z}\right|_{\mathbf{q}(\mu_0)} \neq 0$$

and

$$(1)\qquad \Delta_2(\mathbf{q}(\mu_0)) < 0,$$

$$(2)\qquad \left.\left(\Delta_1 - \frac{\Delta_4}{\Delta_2} - \Delta_2\right)\right|_{\mathbf{q}(\mu_0)} = 0,$$

$$(3)\qquad \left.\frac{d}{d\mu}\left(\Delta_1 - \frac{\Delta_4}{\Delta_2} - \Delta_2\right)\right|_{\mathbf{q}(\mu_0)} \neq 0,$$

$$(4)\qquad \Delta_4(\mathbf{q}(\mu_0)) \neq 0.$$

Then near μ_0 *there is a bifurcation value for the Poincaré map* (4.1) *at which invariant circles occur.*

Proof. The proof consists of showing that the hypotheses of this theorem imply the hypotheses of the Hopf bifurcation theorem for diffeomorphisms (see Iooss [1979] or Marsden and McCracken [1976]). It is a routine calculation using the expressions for the eigenvalues given in (4.3). □

The analogous theorem in the autonomous case is the following.

THEOREM 5.4. *Let* $\mathbf{q}(\mu)$ *be a smooth curve of zeros for* $\mathbf{M}$, $\mu \in I$, *where* I *is an open interval in* $\mathbb{R}$. *Suppose there exists* $\mu_0 \in I$ *such that*

$$(1)\qquad \operatorname{tr} \mathbf{DM}|_{\mathbf{q}(\mu_0)} = 0,$$

$$(2)\qquad \frac{d}{d\mu}(\operatorname{tr} \mathbf{DM}|_{\mathbf{q}(\mu_0)}) \neq 0,$$

$$(3)\qquad \det \mathbf{DM}|_{\mathbf{q}(\mu_0)} > 0.$$

(*We assume these three quantities to be* $O(1)$.) *Then, for* ε *sufficiently small* (*but not zero*), $\hat{\mu} = \mu_0 + O(\varepsilon)$ *is a bifurcation value for the Poincaré map* (3.1) *at which invariant circles occur.*

Proof. Again a routine calculation using the eigenvalues of (4.6) enables one to verify the hypotheses of the (two-dimensional) Hopf bifurcation theorem. □

The stability of the invariant circles (and direction of bifurcation) can be determined by computation of the quantity

$$\begin{aligned} a = &-\frac{\varepsilon}{32\sqrt{\det \mathbf{DM}}}\{(\bar{f}_{uu}+\bar{f}_{vv})(g_{uu}-g_{vv}+2\bar{f}_{uv})+(g_{uu}+g_{vv})(\bar{f}_{uu}-\bar{f}_{vv}-2g_{uv})\} \\ &+\frac{\varepsilon}{16}\{g_{uuu}+g_{uvv}+\bar{f}_{uuv}+\bar{f}_{vvv}\}, \end{aligned} \tag{5.3}$$

where $\bar{f}=\frac{1}{\sqrt{\det \mathbf{DM}}}[M_{1,I}g-M_{3,I}f]$ and f and g are given by

$$\begin{aligned} f(h, k) &= \tfrac{1}{2}[M_{1,II}h^2+2M_{1,Iz}h^2k+M_{1,zz}k^2] \\ &\quad +\tfrac{1}{6}[M_{1,III}h^3+3M_{1,IIz}h^2k+3M_{1,Izz}hk^2+M_{1,zzz}k^3]+O(4), \\ g(h, k) &= \tfrac{1}{2}[M_{3,II}h^2+2M_{3,Iz}hk+M_{3,zz}k^2] \\ &\quad +\tfrac{1}{6}[M_{3,III}h^3+3M_{3,IIz}h^2k+3M_{3,Izz}hk^2+M_{3,zzz}k^3]+O(4), \end{aligned}$$

where all partial derivatives are evaluated at $\mathbf{q}(\mu_0)$. Here h and k are related to u and v via

$$\begin{pmatrix} h \\ k \end{pmatrix} = \begin{pmatrix} \dfrac{M_{1,I}}{M_{3,I}} & -\dfrac{\sqrt{\det \mathbf{DM}}}{M_{3,I}} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}.$$

If $a>0$ (resp. <0) bifurcating circles are unstable (resp. stable). These dreadful formulae follow from application of the standard stability formula of Iooss [1979] or Guckenheimer and Holmes [1983, § 3.5].

6. Examples. In this section we give two examples, the first illustrating the autonomous case and the second illustrating the nonautonomous case.

Our first example not only illustrates how the Hopf and saddle-node bifurcations for diffeomorphisms can interact, but also how the periodic orbits detected by Melnikov theory can be connected to those created in sub- and supercritical Hopf bifurcation from an equilibrium point for the three-dimensional flow. Thus we are able to demonstrate the relationship between the global bifurcation results developed here and conventional local bifurcation methods.

We consider a modification of the van der Pol equation:

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x+z+\varepsilon(x^2y-\delta y), \\ \dot{z} &= \varepsilon(\gamma-z+y^2), \end{aligned} \tag{6.1}$$

with parameters $(\gamma, \delta)\in\mathbb{R}^2$. For $\varepsilon=0$, the system is Hamiltonian with energy

$$H(x, y; z)=\frac{y^2}{2}+\frac{x^2}{2}-zx=h. \tag{6.2}$$

The family of unperturbed periodic orbits in action angle variables is given by:

$$\begin{aligned} x &= z+\sqrt{2I}\sin\theta, \\ y &= \sqrt{2I}\cos\theta, \\ z &= z. \end{aligned} \tag{6.3}$$

Computation of the Melnikov functions by substitution into (2.7) yields

$$\mathbf{M}=\begin{pmatrix} M_1 \\ M_3 \end{pmatrix}=2\pi\begin{bmatrix} I(z^2+I/2-\delta) \\ \gamma-z+I \end{bmatrix}. \tag{6.4}$$

We now have the following.

THEOREM 6.1. *For $\varepsilon>0$ sufficiently small the bifurcation set of* (6.1) *is as in Fig.* 4; *i.e., there exists a parabola $\gamma=\delta^2$ on which Hopf bifurcations to periodic orbits from the unique fixed point $(\gamma, 0, \gamma)$ occur, yielding stable orbits in the region* COAD *and unstable orbits in* DAE, *and there exists two curves within $O(\varepsilon)$ of* AOB $(16\delta=-(1+8\gamma)$; $\gamma<-1/4)$ *and* CO$(\delta=1+(\gamma+2)^2$; $\gamma<-9/4)$ *on which saddle-nodes of periodic orbits and Hopf bifurcations to unstable invariant tori occur. The tori exist above and near* CO (*in* COF). AOB *and* CO *are tangent at* $O(-9/4, 17/16)$ *and* AOB *and* DAE *are tangent at* A $(-1/4, 1/16)$.

Proof. The behavior of the periodic orbits is derived by verification of the hypotheses of Theorems 5.2 and 5.4, using the Melnikov function (6.4), and computation of the stability coefficient a of (5.3). The local Hopf bifurcations from $(\gamma, 0, \gamma)$ follow from elementary linear computations and conventional center manifold and Hopf bifurcation analysis for flows, as in Carr [1981] or Guckenheimer and Holmes [1983, Chap. 3]. We remark that the parabola $\gamma=\delta^2$ also emerges from a study of the Melnikov function, the zeros of which are given by

$$\begin{aligned} I&=\tfrac{1}{2}\{-(2\gamma+\tfrac{1}{2})\pm\sqrt{2\gamma+\tfrac{1}{4}+4\delta}], \\ z&=I+\gamma. \end{aligned} \tag{6.5}$$

For $\delta>\gamma^2$ $\mathbf{M}$ has only one zero in $I>0$ while for $\delta<\gamma^2$, $\gamma<-1/4$ and $16\delta>-(1+8\gamma)$ $\mathbf{M}$ has two such zeros. □

In Fig. 5 the results of numerical integrations of (6.1) are shown for parameter values in various regions of Fig. 4. These results clearly illustrate the periodic orbits and the invariant torus for the flow predicted by the Melnikov analysis.

To illustrate the nonautonomous theory we consider the equation governing a pendulum subject to weak damping and variable torque. The torque is supplied by a servo-motor or some other device, the dynamics of which is modeled by a first order

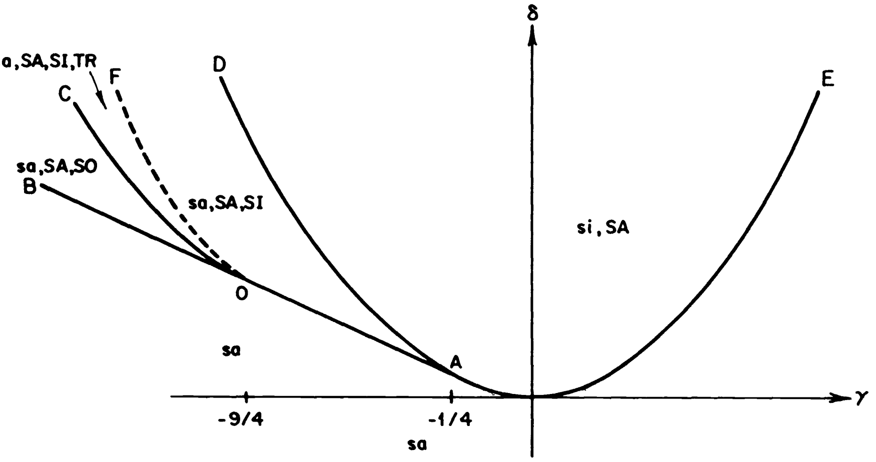


FIG. 4. *The bifurcation set for* (6.1): si = *sink*, sa = *saddle*; SA = *saddle type periodic orbit*; SI = *attracting periodic orbit*; SO = *repelling periodic orbit*; TR = *repelling invariant torus.*

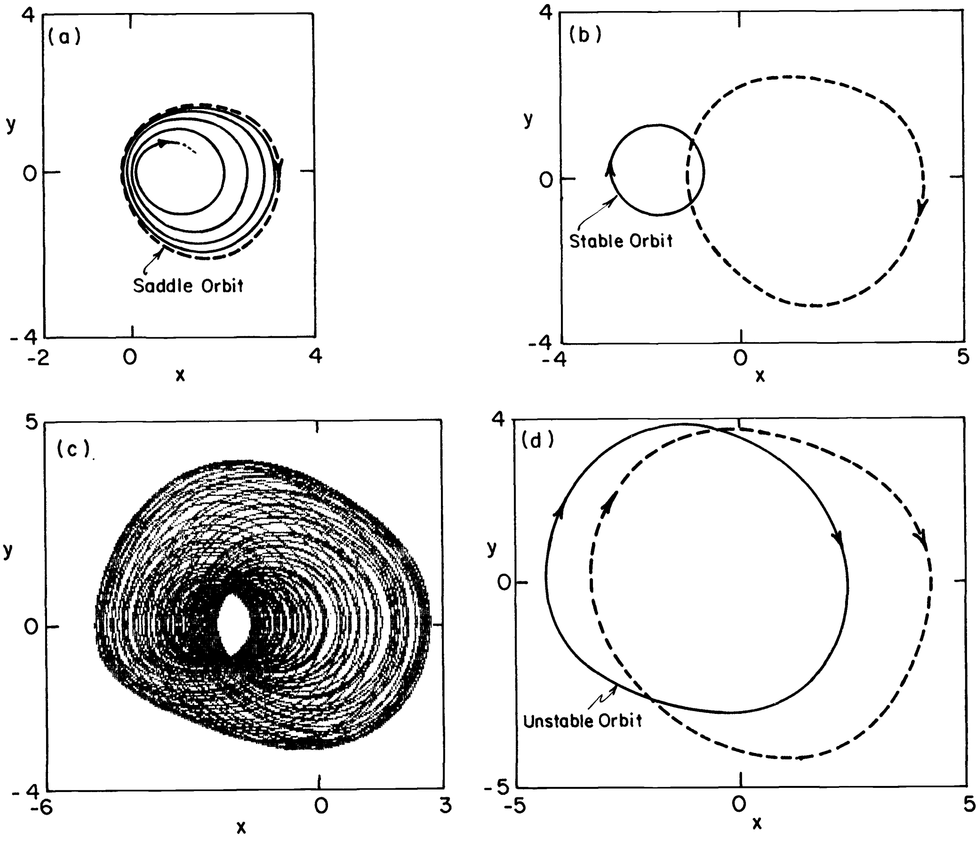


FIG. 5. *Numerical integrations of* (6.1).

equation. The servo-motor is also driven by an external periodic perturbation. The equations of motion are

$$\dot{x} = y,$$
$$\dot{y} = -\sin x + \varepsilon(z - \delta y), \tag{6.6}$$
$$\dot{z} = \varepsilon(-\gamma z + y \cos t),$$

so that the unperturbed system is Hamiltonian with energy

$$H(x, y) = \frac{y^2}{2} + (1 - \cos x). \tag{6.7}$$

We will consider the effect of the perturbation on rotational orbits (overswinging of the pendulum). Note that here H does not depend on z and so the unperturbed system is identical on each $z = z_0$ constant slice. This is unimportant in what follows but does simplify the computations somewhat. The unperturbed orbits are given by

$$x(t) = 2 \sin^{-1}\left(\text{sn}\left[\frac{K(k)}{\pi}(t + t_0)\right]\right),$$
$$y(t) = \frac{2}{k}\,\text{dn}\left[\frac{K(k)}{\pi}(t + t_0)\right], \qquad k \in (0, 1), \tag{6.8}$$
$$z(t) = z_0,$$

where sn and dn are the Jacobi elliptic functions and K is the complete elliptic integral of the first kind with modulus k. The resonance relation is given by

$$T(I)=T(k)=\frac{2\pi m}{n}=\frac{2\pi}{\Omega(I)}=2kK(k), \tag{6.9}$$

where $2kK(k)$ is the period of the unperturbed orbit and 2π the period of the forcing function. Note that the modulus k plays essentially the same role as the action I or energy H, since $I(k)$ and $H(k)$ are both monotonic functions (see below).

Using the remark at the end of § 3, we now compute the Melnikov functions $M_1^{m/n}$ and $M_3^{m/n}$. Dropping the superscripts, we have

$$M_1=\frac{1}{\Omega(I)}\int_0^{2\pi m}(yz-\delta y^2)\,dt, \tag{6.10a}$$

$$M_3=\int_0^{2\pi m}(-\gamma z+y\cos t)\,dt, \tag{6.10b}$$

where $\Omega(I)$ and y depend implicitly on k, which is selected to satisfy the resonance relation (6.9). The computations are lengthy and make use of Fourier expansions of the elliptic function dn (cf. Byrd and Friedman [1971]). For the case $n=1$ we obtain

$$\begin{aligned} M_1&=2m\left[\pi z_0-\frac{4\delta E(k)}{k}\right],\\ M_3&=\pi m\left[-2\gamma z_0+\frac{1}{k}\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right)\cos m\theta_0\right], \end{aligned} \tag{6.11}$$

where $E(k)$ is the complete elliptic integral of the second kind and $K'(k)=K(\sqrt{1-k^2})$ is the complementary complete elliptic integral of the first kind. In transforming from the initial starting time $t_0\in[0,2\pi m/n]$, which appears in (6.8), to the phase θ_0 we use the relationship $t_0=m\theta_0/n$. When $n\neq 1$ we obtain

$$M_3=-2\pi m\gamma z_0 \tag{6.12}$$

and M_1 remains as in (6.10a). Thus, at first order, only "pure" $m/1$ subharmonics are excited (cf. Greenspan and Holmes [1983], [1984]). Henceforth we drop the subscript "0."

Since the frequency of the unperturbed orbits is

$$\Omega(k)=\frac{2\pi}{T(k)}=\frac{\pi}{kK(k)}, \tag{6.13}$$

we have $\partial\Omega/\partial z\equiv 0$ and

$$\frac{\partial\Omega}{\partial I}=\frac{\partial\Omega}{\partial k}\frac{\partial k}{\partial I}=\left(\frac{-\pi E(k)}{k^2(1-k)^2K(k)}\right)\left(\frac{-\pi k^2}{4K(k)}\right)=\frac{\pi^2E(k)}{4(1-k)^2K^2(k)}>0\quad\text{for } k\in(0,1), \tag{6.14}$$

where we have used the fact that the action $I(k)$ is given by the monotonic function

$$I(k)=\frac{4}{\pi k}E(k). \tag{6.15}$$

We remark that the limits $k=0$ and 1 correspond to the Hamiltonian energies $H=\infty$ and $H=2$, the latter being the energy of the homoclinic loop (separatrix) of the unperturbed system. Thus the frequency increases monotonically with action I in the range $I\in(4/\pi,\infty)(k\in(1,0))$.

We now apply Theorem 3.1 to the example. For $n=1$, the resonance relation (6.9) becomes $\pi m = kK(k)$, which, given $m \in Z^+$, determines k uniquely since $kK(k)$ is monotone. Then, from (6.11), for a zero of M_1 we require

$$z = \frac{4\delta E(k)}{\pi k}, \tag{6.16}$$

which determines z, and for a zero of M_3

$$\frac{8\gamma\delta E(k)}{\pi k} = \frac{1}{k}\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right)\cos m\theta, \tag{6.17}$$

which determines precisely $2m$ values of θ in the range $[0, 2\pi)$, provided that

$$\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right) > \left|\frac{8\gamma\delta E(k)}{\pi}\right|. \tag{6.18}$$

We leave it to the reader to verify that the nondegeneracy condition of the theorem also holds.

When (6.18) is an equality we have bifurcation (the $2m$ simple zeros degenerate into m multiple roots). To verify that nondegenerate saddle-node bifurcations occur, we fix $\delta > 0$ and vary γ, thus converting the problem to a one parameter system. Applying Theorem 5.1, we require a zero of M_1, M_2 at which, in addition,

$$\frac{\partial \Omega}{\partial I}\left(\frac{\partial(M_1, M_3)}{\partial(z, \theta)}\right) = 0 \quad \text{and thus} \quad \frac{\partial(M_1, M_3)}{\partial(z, \theta)} = 0 \tag{6.19}$$

(since $\partial\Omega/\partial z \equiv 0$ and $\partial\Omega/\partial I \neq 0$ here). This implies that

$$\begin{aligned} &[2m\pi]\cdot\left[\frac{\pi m^2}{k}\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right)\sin m\theta\right] = 0 \\ &\Rightarrow m\theta = 0, \pi, \cdots \Rightarrow \theta = \frac{l\pi}{m}, \qquad l = 0, \cdots, 2m-1. \end{aligned} \tag{6.20}$$

Note that $\sin m\theta = 0$, $\cos m\theta = \pm 1$, which occurs precisely where (6.18) becomes an equality. Finally, nondegeneracy condition (b) of Theorem 5.1 holds, since $\partial\Omega/\partial z \equiv 0$ and $\partial\Omega/\partial I \neq 0$ and

$$\frac{\partial(M_1, M_3)}{\partial(\gamma, z)} = [0]\cdot[-2\pi m\gamma] - [2m\pi]\cdot[-2\pi m z] = 4\pi^2 m^2 z \neq 0,$$

when $z \neq 0$, and

$$\begin{aligned} \frac{d}{d\theta}\left(\frac{\partial(M_1, M_3)}{\partial(z, \theta)}\right) &= \frac{d}{d\theta}\left\{[2\pi m]\cdot\left[\frac{\pi m^2}{k}\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right)\sin m\theta\right] - [0]\cdot[-2\pi m\gamma]\right\} \\ &= -\frac{2\pi^2 m^4}{k}\operatorname{sech}\left[\frac{\pi m K'(k)}{K(k)}\right]\cos m\theta \neq 0, \end{aligned} \tag{6.21}$$

when $\cos m\theta (= \pm 1) \neq 0$. (Since $\gamma, \delta > 0$ we only obtain the $\cos m\theta = +1$ case). Summarizing our results, we have the following.

THEOREM 6.2. *There exists a countable set of bifurcation curves in* (γ, δ)*-space given by*

$$\gamma\delta = \frac{\pi}{8E(k)}\operatorname{sech}\left(\frac{\pi m K'(k)}{K(k)}\right), \quad kK(k) = \pi m, \quad m = 1, 2, \cdots, \tag{6.22}$$

near which, for ε sufficiently small (depending on m), saddle-node bifurcations to pairs of $2\pi m$-periodic orbits occur for the Poincaré map of (6.6).

Eigenvalue computations as outlined in § 4 reveal that each pair of subharmonics consists of a stable sink and a saddle. The requirement that $M_1 = 0$ (see (6.16)) together with the monotonicity of $kK(k)$ shows that these periodic points all lie near a bowl-like surface of revolution in (θ, I, z) (or (x, y, z)) space. Figure 6 gives our impression of the situation. Note how the periodic points accumulate near $z = 4\delta/\pi$ ($k = 1$; $H = 2$). In this connection, we note that the curves (6.22) accumulate on the curve $\gamma\delta = \pi/8 \operatorname{sech}(\pi/2)$ as $k \to 1^-(H \to 2^+)$. We plan to study questions relating to the resulting chaotic invariant sets in a subsequent paper.

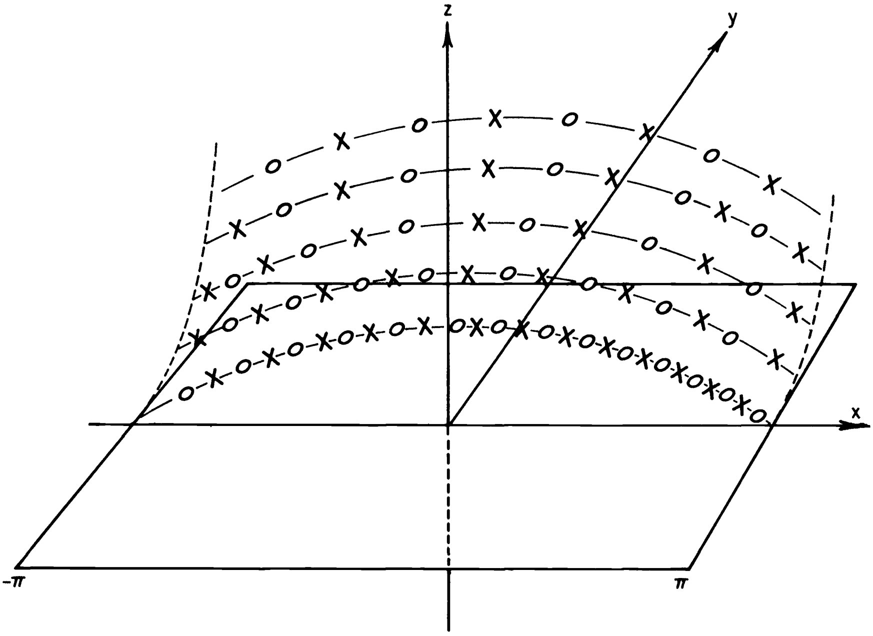


FIG. 6. *Subharmonics for the perturbed pendulum.*

We end our discussion of this example by remarking that since M_3 has no zeros (other than $z = 0$) unless $n = 1$ or $\gamma = 0$, no other periodic orbits exist near the unperturbed rotational solutions for $\delta, \gamma > 0$. We also note that the "bowl" of $m/1$ subharmonics created in (x, y, z) space (Fig. 6) has a similar structure to the unfolding of a Hopf bifurcation for diffeomorphisms of $\mathbb{R}^2$ recently studied by Chenciner [1983], [1985a], [1985b].

REFERENCES

V. I. ARNOLD (1978), *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, Berlin.

——, (1982), *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, Berlin.

N. H. BAKER, D. W. MOORE AND E. A. SPEIGEL (1971), *Aperiodic behavior of nonlinear oscillator*, Quart. J. Mech. Appl. Math., 24, pp. 391-422.

P. F. BYRD AND M. D. FRIEDMAN (1971), *Handbook of Elliptic Integrals for Scientists and Engineers*, Springer-Verlag, New York, Berlin.

J. CARR (1981), *Application of Centre Manifold Theory*, Springer-Verlag, New York, Berlin.

A. CHENCINER (1983), *Bifurcations de difféomorphismes de $\mathbb{R}^2$ au voisinage d'un point fixe elliptique*, in Les Houches Summer School Proceedings, R. Helleman, ed., G. Iooss, North-Holland, Amsterdam.

———, (1985a) *Bifurcations de points fixes elliptiques*, I.—*Courbes Invariantes*, Publ. Math. de 'l I.H.E.S., 61, pp. 67-127.

———, (1985b), *Bifurcations de points fixes elliptiques*, II. *Orbites periodiques et ensembles de Cantor invariants*, Invent. Math., 80, pp. 81-106.

H. GOLDSTEIN (1980), *Classical Mechanics*, 2nd ed., Addison-Wesley, Reading, MA.

B. D. GREENSPAN AND P. J. HOLMES (1983), *Homoclinic orbits, subharmonics and global bifurcations in forced oscillations*, in Nonlinear Dynamics and Turbulence, G. Barenblatt, G. Iooss and D. D. Joseph, eds., Pitman, London, Chap. 10, pp. 272-214.

———, (1984), *Repeated resonance and homoclinic orbits in a periodically forced family of oscillators*, this Journal, 15, pp. 69-97.

J. GRUENDLER (1985), *The existence of homoclinic orbits and the method of Melnikov for systems in* $\mathbb{R}^n$, this Journal, 16, pp. 907-931.

J. GUCKENHEIMER AND P. J. HOLMES (1983), *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, Berlin.

P. HARTMAN (1964), *Ordinary Differential Equations*, John Wiley, New York.

M. HÉNON (1976), *A two-dimensional mapping with a strange attractor*, Comm. Math. Phys., 50, pp. 69-77.

P. J. HOLMES (1983), *Bifurcation and chaos in a simple feedback control system*, Proc. 22nd IEEE Conference on Control and Decision, paper wp5, Vol. 1, pp. 365-370.

———, (1985), *Dynamics of a nonlinear oscillator with feedback control*, I: *Local analysis*, ASME J. Dyn. Sys. Meas. Control, 107, pp. 159-165.

P. J. HOLMES AND J. E. MARSDEN (1982a), *Horseshoes in perturbations of Hamiltonian systems with two degrees of freedom*, Comm. Math. Phys., 82, pp. 523-544.

———, (1982b), *Melnikov's method and Arnold diffusion for perturbations of integrable Hamiltonian systems*, J. Math. Phys., 23, pp. 669-675.

———, (1983), *Horseshoes and Arnold diffusion for Hamiltonian systems on Lie groups*, Indiana Univ. Math. J., 32, pp. 263-319.

G. IOOSS (1979), *Bifurcations of Maps and Applications*, North-Holland, Amsterdam.

E. N. LORENZ (1963), *Deterministic nonperiodic flow*, J. Atmospheric Sci., 20, pp. 130-141.

J. E. MARSDEN AND M. MCCRACKEN (1976), *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York-Heidelberg-Berlin.

C. J. MARZEC AND E. A. SPIEGEL (1980), *Ordinary differential equations with strange attractors*, SIAM J. Appl. Math., 38, pp. 403-421.

V. K. MELNIKOV (1963), *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12, pp. 1-57.

F. C. MOON AND R. H. RAND (1984), *Parametric stiffness control of flexible structures*, Proc. NASA workshop on identification and control of flexible space structures, San Diego, May 1984.

K. A. ROBBINS (1979) *Periodic solutions and bifurcation structure at high R in the Lorenz model*, SIAM J. Appl. Math., 36, pp. 457-472.

C. ROBINSON (1983), *Sustained resonance for a nonlinear system with slowly varying coefficients*, this Journal, 14, pp. 847-860.

C. T. SPARROW (1982), *Chaos in a three-dimensional single loop feedback system with a piecewise linear feedback function*, J. Math. Anal. Appl., 83, pp. 275-291.

———, (1981), *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*, Springer-Verlag, New York, Berlin, Heidelberg.

S. WIGGINS AND P. HOLMES (1987), *Homoclinic orbits in slowly varying oscillators*, this Journal, 18, pp. 612-629.

S. WIGGINS (1985), *Slowly varying oscillators*, Ph.D. thesis, Cornell Univ., Ithaca, N.Y.

———, (1986), *A generalization of the method of Melinkov for detecting chaotic invariant sets*, Caltech preprint.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

HOMOCLINIC ORBITS IN SLOWLY VARYING OSCILLATORS*

STEPHEN WIGGINS† AND PHILIP HOLMES‡

Abstract. We obtain existence and bifurcation theorems for homoclinic orbits in three-dimensional flows that are perturbations of families of planar Hamiltonian systems. The perturbations may or may not depend explicitly on time. We show how the results on periodic orbits of the preceding paper are related to the present homoclinic results, and apply them to a periodically forced Duffing equation with weak feedback.

Key words. bifurcation, Hamiltonian system, homoclinic orbit, perturbation theory, Melnikov method

AMS(MOS) subject classifications. 34CXX, 58F14, 70KXX

1. Introduction. In the preceding paper we developed perturbation methods based on ideas of Melnikov [1963] that permit us to approximate Poincaré maps for autonomous and periodically forced slowly varying oscillators, the flows of which are close to those of families of planar Hamiltonian systems. We obtained existence, stability and bifurcation results for periodic orbits in such systems. In the present paper we extend these results to deal with homoclinic orbits and show how the periodic results are related to them.

In § 2 we outline the geometry of the phase space and we describe basic perturbation results. The computational tools and existence and bifurcation theorems are developed in §§ 3 and 4, and the relationship between periodic and homoclinic orbits is discussed in § 5. The example and conclusions follow in §§ 6 and 7.

2. Structure of the phase space. As in Wiggins and Holmes [1987] we will consider systems of the form

$$\left.\begin{aligned} \dot{x} &= f_1(x, y, z) + \varepsilon g_1(x, y, z, t, \boldsymbol{\mu}) \\ \dot{y} &= f_2(x, y, z) + \varepsilon g_2(x, y, z, t; \boldsymbol{\mu}) \\ \dot{z} &= \varepsilon g_3(x, y, z, t; \boldsymbol{\mu}) \end{aligned}\right\} \quad \text{or} \quad \dot{\mathbf{q}} = \mathbf{f}(\mathbf{q}) + \varepsilon \mathbf{g}_{\mu}(\mathbf{q}, t), \tag{2.1}$$

with $0 < \varepsilon \ll 1$, **f** and **g** sufficiently smooth ($C^r, r \geqq 2$), **g** periodic in t with period T and $\boldsymbol{\mu} \in R^k$ a vector of parameters. We will write $\mathbf{g}(\mathbf{q}, t; \boldsymbol{\mu}) = \mathbf{g}_{\mu}(\mathbf{q}, t)$ and frequently drop the explicit dependence on $\boldsymbol{\mu}$. We make the following assumptions on the unperturbed system:

(A1) For $\varepsilon = 0$, (2.1) reduces to a one-parameter family of planar Hamiltonian systems with Hamiltonian $H(x, y, z)$:

$$\begin{aligned} \dot{x} &= f_1(x, y, z) = \frac{\partial H}{\partial y}, \\ \dot{y} &= f_2(x, y, z) = -\frac{\partial H}{\partial x}, \\ \dot{z} &= 0. \end{aligned} \tag{2.2}$$

* Received by the editors June 1, 1985; accepted for publication (in revised form) April 22, 1986. This research was supported in part by the Air Force Office of Scientific Research under AFSOR 84-0051.

† Department of Applied Mathematics, California Institute of Technology, Pasadena, California 91125.

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

(A2) For each value of z in some open interval $J \subseteq \mathbb{R}$ the "planar" system (2.2) possesses a homoclinic orbit to a hyperbolic saddle point. Thus, when viewed in the full three-dimensional phase space, system (2.2) possesses a normally hyperbolic invariant one-dimensional manifold, $\mathcal{N}$, given by the union of saddle points of the one-parameter family of planar systems. $\mathcal{N}$ has two-dimensional stable and unstable manifolds (denoted by $W^s(\mathcal{N})$, $W^u(\mathcal{N})$, respectively), such that their intersection $W^s(\mathcal{N}) \cap W^u(\mathcal{N}) \overset{\text{def}}{=} \Gamma$ is made up of the union of the homoclinic orbits of the one-parameter family of planar systems. Henceforth we assume that $\mathcal{N}$ is connected; if not, the theory is applied separately to each connected component of $\mathcal{N}$.

(A3) The interior of Γ contains a two-parameter family of periodic orbits, which we denote by $\mathbf{q}_0^{\alpha,z}(t-\theta)$ for $z \in J$ and $\alpha \in L(z)$, where for each $z \in J$, $L(z)$ is an open interval in R. We denote $L(z)$ by $(\alpha(z), \alpha_0(z))$ and assume that $\lim_{\alpha \to \alpha_0} T(\alpha, z) = \infty$, where $T(\alpha, z)$ denotes the period of $\mathbf{q}_0^{\alpha,z}(t-\theta)$ and that $T(\alpha, z)$ is a differentiable function of α and z with $dT(\alpha, z)/d\alpha \neq 0$ for $(\alpha, z) \in (L(z), J)$.

Note that the assumptions of Wiggins and Holmes [1987] are included in the above. As before we suspend (2.1) over the space $R^3 \times S^1$ where $S^1 = R/T$ is the circle of length T by defining the function $\phi(t) = t$, mod T and then by T-periodicity of the g_i we have

$$\begin{aligned} \dot{x} &= f_1(x, y, z) + \varepsilon g_1(x, y, z, \phi; \boldsymbol{\mu}), \\ \dot{y} &= f_2(x, y, z) + \varepsilon g_2(x, y, z, \phi; \boldsymbol{\mu}), \\ \dot{z} &= \varepsilon g_3(x, y, z, \phi; \boldsymbol{\mu}), \\ \dot{\phi} &= 1. \end{aligned} \qquad (x, y, z, \phi) \in R^3 \times S^1, \tag{2.3}$$

Again we note that this suspension makes sense even when the g_i are independent of ϕ. At $\varepsilon = 0$, for the suspended system we denote the normally hyperbolic invariant set by $\mathcal{M} \equiv (\mathcal{N}, \phi) = \mathcal{N} \times S^1$. See Fig. 1.

For computations it is convenient to have $\mathcal{M}$ in an explicit form. Recall from assumption (A2) that $\mathcal{N}$ is a one-manifold of equilibrium points for the unperturbed system such that on each $z =$ constant plane the equilibrium point of the associated planar system is hyperbolic. Since we have assumed that the unperturbed vector field is Hamiltonian, a simple computation of the eigenvalues of the linearized vector field at this point shows that $\partial(f_1, f_2)/\partial(x, y) < 0$. Thus, by the implicit function theorem,

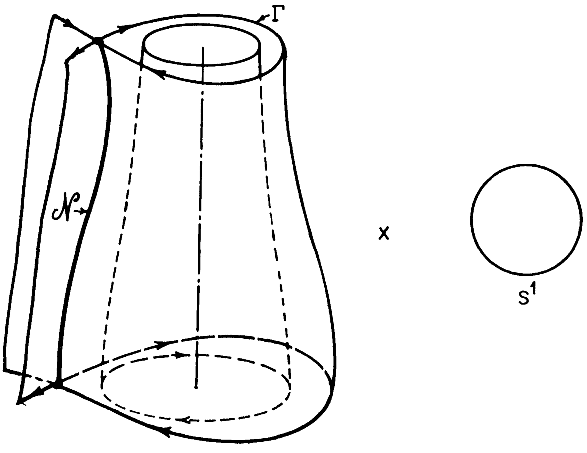


FIG. 1. *The unperturbed phase space.*

$\mathcal{M}$ can be represented as a graph over the z variables:

$$\mathcal{M}=\left\{(\gamma(z),\phi)\;\middle|\;\gamma(z)=(x(z),y(z),z), f_1(x,y,z)=f_2(x,y,z)=0,\ \left.\frac{\partial(f_1,f_2)}{\partial(x,y)}\right|_{\gamma(z)}<0,\ \phi\in S^1,\ z\in J\right\}. \tag{2.4}$$

The following results give us information about the perturbed phase space.

Proposition 2.1. *There exists $\varepsilon_0>0$ such that for $0<\varepsilon<\varepsilon_0\ll 1$ there exists a normally hyperbolic invariant one-manifold*

$$\mathcal{M}_\varepsilon=\{(\gamma(z,\phi;\varepsilon),\phi)=(\gamma(z)+\mathcal{O}(\varepsilon),\phi)\,|\,\phi\in S^1, z\in J\}, \tag{2.5}$$

where $\gamma(z,\phi;\varepsilon)$ is a C^r function of z and ε. Moreover, $\mathcal{M}_\varepsilon$ has local stable and unstable manifolds. $W^s_{\text{loc}}(\mathcal{M}_\varepsilon)$, $W^u_{\text{loc}}(\mathcal{M}_\varepsilon)$, which are C^r-close to the local stable and unstable manifolds of $\mathcal{M}$, denoted by $W^s_{\text{loc}}(\mathcal{M})$ and $W^u_{\text{loc}}(\mathcal{M})$, respectively.

Remark. $\mathcal{M}_\varepsilon$ is an invariant manifold in the weaker sense in that solutions may leave $\mathcal{M}_\varepsilon$ by virtue of their z values crossing the boundary of J. This will occur on a time scale $\mathcal{O}(1/\varepsilon)$ since motion along $\mathcal{M}_\varepsilon$ has a speed $\mathcal{O}(\varepsilon)$.

Proof. The existence of $\mathcal{M}_\varepsilon$, $W^s_{\text{loc}}(\mathcal{M}_\varepsilon)$, and $W^u_{\text{loc}}(\mathcal{M}_\varepsilon)$ follows from the persistence of normally hyperbolic invariant sets and their stable and unstable manifolds (see Hirsch, Pugh and Shub [1977] or Fenichel [1971]), with some slight technical modifications. The usual theorem requires $\mathcal{M}$ to be compact and boundaryless. However, there are two ways to get around these requirements. One, due to Robinson [1983], involves mapping $\mathcal{M}$ into a compact space (e.g. a sphere) and smoothly extending the vector field to a neighborhood of $\mathcal{M}$ via the use of bump functions, the conclusions then follow from the Hirsch, Pugh and Shub [1977] theory. The second method is due to Kopell [1985] and involves the use of the invariant manifold theory of Fenichel [1971]. Briefly, the vector field on the boundary of $\mathcal{M}$ is zero, $\mathcal{M}$ is then perturbed in a neighborhood of its boundary via a bump function in such a manner that it becomes "overflowing" (resp. "underflowing") invariant (see Fenichel [1971] for precise definitions). The existence of the perturbed manifold and its local unstable (resp. stable) manifold then follows from the Fenichel invariant manifold theorem. Furthermore, the modification of the vector field near the boundary of $\mathcal{M}$ does not affect the dynamics of our original system in the sense that, although now $\mathcal{M}_\varepsilon$ and $W^{s,u}_{\text{loc}}(\mathcal{M}_\varepsilon)$ may depend on the specific modification, asymptotic expansions of these manifolds agree to all orders for arbitrary modifications (Kopell [1985]). (Note: the situation is the same as that which arises in applications of center manifold theory, where the nonuniqueness of the center manifold does not effect recurrent motions.) □

On $\mathcal{M}$ all points are fixed, there is no motion. However, on $\mathcal{M}_\varepsilon$ this need not be the case. The following result gives us information concerning the flow on $\mathcal{M}_\varepsilon$.

Proposition 2.2. *Let $\overline{g_3(\gamma(z))}=1/T\int_0^T g_3(\gamma(z),\phi)\,d\phi$ and suppose there exists $z_0\in J$ such that $\overline{g_3(\gamma(z_0))}=0$, $(d/dz)\overline{g_3(\gamma(z_0))}\neq 0$. Then $(\gamma(z_0,\phi,\varepsilon),\phi)=(\gamma(z_0)+\mathcal{O}(\varepsilon),\phi)$ is a hyperbolic periodic orbit on $\mathcal{M}_\varepsilon$ with period T.*

Proof. This is a straightforward application of the averaging theorem (see Hale [1969] or Guckenheimer and Holmes [1983]) restricted to $\mathcal{M}_\varepsilon$. □

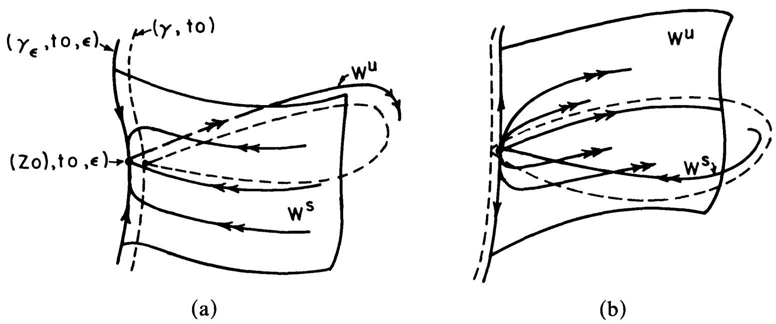
Remark. If g_3 is not explicitly time dependent, then $g_3\equiv\overline{g_3}$ and averaging is unnecessary, so the proof goes through without appeal to the averaging theorem.

In order to visualize the situation we take the following cross section to the flow induced by (2.3):

$$\Sigma^{t_0}=\{(x,y,z,\phi)\in R^3\times S^1\,|\,\phi=t_0\in[0,T)\}. \tag{2.6}$$

If there exists a hyperbolic periodic orbit on $\mathcal{M}_\varepsilon$ there are two possible situations (see Fig. 2).

In these pictures the solid lines are to be interpreted as initial conditions for solutions of the perturbed equation, while the dotted lines can be thought of as actual solutions of the unperturbed equation, since the unperturbed equation is autonomous.



$(\gamma_\varepsilon, t_0, \epsilon)$
(γ, t_0)
W^u
$(\gamma\ (Z_0), t_0, \epsilon)$
W^s
(a)
W^u
W^s
(b)

FIG. 2. *The perturbed manifolds.* (a) $\gamma(z_0)+\mathcal{O}(\varepsilon)$ *has* 1-*d unstable and* 2-*d stable manifolds.* (b) $\gamma(z_0)+\mathcal{O}(\varepsilon)$ *has* 1-*d stable and* 2-*d unstable manifolds.*

The following perturbation results allows us to approximate certain solutions in the stable and unstable manifolds of $(\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)$ for arbitrarily long time intervals. This is necessary since we wish to find homoclinic, rather than periodic, orbits.

PROPOSITION 2.3. *Suppose there exists* $z_0 \in J$ *such that* $(\gamma(z_0, \phi; \varepsilon), \phi) = (\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)$ *is a hyperbolic periodic orbit on* $\mathcal{M}_\varepsilon$. *Then, for each* ε *sufficiently small, there exists* $C>0$, $K=\{z\colon z_0 - C\varepsilon < z < z_0 + C\varepsilon\}$ *and solutions* $\mathbf{q}_\varepsilon^s(t, \theta)$, $\mathbf{q}_\varepsilon^u(t, \theta)$ *lying in the stable and unstable manifolds of* $(\gamma(z_0, \phi; \varepsilon), \phi)$ *with the following representations valid in the indicated time intervals*:

Case (a) dim $W^s[(\gamma(z_0, \phi; \varepsilon), \phi)] = 3$, dim $W^u[(\gamma(z_0, \phi; \varepsilon), \phi)] = 2$:

$$\mathbf{q}_\varepsilon^s(t, \theta) = \mathbf{q}_0(t-\theta) + \varepsilon \mathbf{q}_1^s(t, \theta) + \mathcal{O}(\varepsilon^2), \qquad z_\varepsilon^s(t_0, \theta) \in K, \quad t \in [t_0, \infty).$$

$$\mathbf{q}_\varepsilon^u(t, \theta) = \mathbf{q}_0(t-\theta) + \varepsilon \mathbf{q}_1^u(t, \theta) + \mathcal{O}(\varepsilon^2), \qquad t \in (-\infty, t_0];$$

Case (b) dim $W^s[(\gamma(z_0, \phi; \varepsilon), \phi)] = 2$, dim $W^u[(\gamma(z_0, \phi; \varepsilon), \phi)] = 3$:

$$\mathbf{q}_\varepsilon^s(t, \theta) = \mathbf{q}_0(t-\theta) + \varepsilon \mathbf{q}_1^s(t, \theta) + \mathcal{O}(\varepsilon^2), \qquad t \in [t_0, \infty),$$

$$\mathbf{q}_\varepsilon^u(t, \theta) = \mathbf{q}_0(t-\theta) + \varepsilon \mathbf{q}_1^u(t, \theta) + \mathcal{O}(\varepsilon^2), \qquad z_\varepsilon^u(t_0, \theta) \in K, \quad t \in (-\infty, t_0],$$

where $\mathbf{q}_0(t-\theta)$ *is the solution of the unperturbed equation that connects the point* $\gamma(z_0)$ *on* $\mathcal{M}$ *to itself, i.e., the homoclinic orbit on the* $z = z_0$ *plane.*

Proof. See Wiggins [1985]. □

Remarks. (1) $\mathbf{q}_1^s(t, \theta)$ and $\mathbf{q}_1^u(t, \theta)$ may be obtained by solving the first variational equations

$$\dot{\mathbf{q}}_1^s = \mathbf{Df}(\mathbf{q}_0)\mathbf{q}_1^s + \mathbf{g}(\mathbf{q}_0, t), \qquad t \in [t_0, \infty),$$

$$\dot{\mathbf{q}}_1^u = \mathbf{Df}(\mathbf{q}_0)\mathbf{q}_1^u + \mathbf{g}(\mathbf{q}_0, t), \qquad t \in (-\infty, t_0].$$

(2) We note that in the (suspended) unperturbed system, dim $W^s[\gamma(z_0), \phi] =$ dim $W^u[(\gamma(z_0), \phi)] = 2$ and that for the perturbed system the dimensions of $W^s[(\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)]$ or $W^u[(\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)]$ may increase by one (Cases (a) and (b)). So in order to uniformly approximate solutions in the unperturbed manifolds by

solutions in the unperturbed manifolds for arbitrarily long time intervals, these solutions must initially start out close together. This is the reason for the requirements $z_\varepsilon^{s,u}(t_0, \theta) \in K$ in Cases (a) and (b).

(3) For the stable manifold in Case (a) and the unstable manifold in Case (b), the theorem does not tell us explicitly which solution in the manifolds we are on, only that all solutions with initial z values in K are approximated uniformly by a corresponding solution to the unperturbed equation. Consequently we are no longer able to follow individual solutions in these manifolds during their time evolutions.

In § 5 we will be concerned with periodic orbits limiting on Γ. In this situation we need to approximate perturbed solutions arbitrarily close to Γ by unperturbed periodic orbits so we need some kind of control on the flow on $\mathcal{M}_\varepsilon$.

PROPOSITION 2.4. *Let* $(\gamma(z_0) + \mathcal{O}(\varepsilon), \phi)$ *be a hyperbolic periodic orbit on* $\mathcal{M}_\varepsilon$ *and let* $\mathbf{q}^{\alpha,z_0}(t-\theta)$ *be a periodic orbit of the unperturbed system with period* $T(\alpha, z_0)$. *Then there exists a perturbed orbit* $\mathbf{q}_\varepsilon^{\alpha,z_0}(t, \theta)$, *not necessarily periodic, which can be expressed as*

$$\mathbf{q}_\varepsilon^{\alpha,z_0}(t, \theta) = \mathbf{q}_0^{\alpha,z_0}(t-\theta) + \varepsilon \mathbf{q}_1^{\alpha,z_0}(t, \theta) + \mathcal{O}(\varepsilon^2)$$

uniformly in $t \in [t_0, t_0 + T(\alpha, z_0)]$, *for* ε *sufficiently small and all* $\alpha \in L(z_0)$.

Proof. See Wiggins [1985]. □

We will remark that Proposition 2.4 only allows us to approximate perturbed orbits by unperturbed orbits for one passage through a neighborhood of $\mathcal{M}_\varepsilon$. This is due to the fact that orbits take arbitrarily long to pass through the neighborhood and therefore the slightest error may be magnified greatly over the long time of passage. Consequently for periodic orbits near Γ we are limited to the study of resonant orbits satisfying $mT = nT(\alpha, z_0)$, $n = 1$. However, since we can pick $T(\alpha, z_0)$ at will, and by (A3) $T \to \infty$ as $\alpha \to \alpha_0$, m can be arbitrarily large.

3. Existence of homoclinic orbits. We now turn our attention to the homoclinic manifold Γ. By Proposition 2.3, in order to approximate to $\mathcal{O}(\varepsilon)$ orbits in the stable and unstable manifolds of $\mathcal{M}_\varepsilon$ by orbits in the stable and unstable manifolds of $\mathcal{M}$ it is necessary that there exist a point $z_0 \in J$ such that $(\gamma(z_0, \phi; \varepsilon), \phi)$ is a hyperbolic periodic orbit on $\mathcal{M}_\varepsilon$. Now a hyperbolic periodic orbit on $\mathcal{M}_\varepsilon$ will have either a three-dimensional stable manifold and a two-dimensional unstable manifold or vice versa. Thus in the four-dimensional phase space we expect the intersection to be generically one-dimensional. In measuring distances between manifolds of solutions in phase space it is only necessary to explore the directions transverse to the manifolds, so the number of measurements necessary in order to determine whether or not the manifolds intersect should be equal to the minimum codimension of the manifolds. In our case that number is one and we expect a single (scalar) measurement to suffice.

Now on the cross section Σ^0 the hyperbolic periodic orbit for the flow, $(\gamma(z_0, \phi; \varepsilon), \phi)$, corresponds to a hyperbolic fixed point, $\gamma(z_0) + \mathcal{O}(\varepsilon)$, for the Poincaré map P_ε, which has either a two-dimensional stable manifold and a one-dimensional unstable manifold (Case (a) of Proposition 2.3) or a one-dimensional stable manifold and a two-dimensional unstable manifold (Case (b) of Proposition 2.3). For definiteness, in the following we assume that Case (a) holds, since the argument and conclusions for Case (b) are identical.

We develop a measure for the distance between the stable manifold $W^s(\gamma(z_0) + \mathcal{O}(\varepsilon))$ and the unstable manifold $W^u(\gamma(z_0) + \mathcal{O}(\varepsilon))$ on the cross-section Σ^0. Let $\alpha_0 = \alpha(z_0)$ denote the value of α on the $z = z_0$ level that corresponds to the unperturbed homoclinic orbit on that z-level, and denote this orbit $\mathbf{q}_0(t-\theta)$, where we have dropped the explicit (α, z) dependence for ease of notation.

At the point $\mathbf{q}_0(-\theta)$ on the cross-section Σ^0 we consider the plane Π normal to the vector $\mathbf{f}(\mathbf{q}_0(-\theta))$. There exists a unique point $\mathbf{q}^u_\varepsilon(-\theta)$ in $W^u(\gamma(z_0)+\mathcal{O}(\varepsilon))\cap\Pi$ which is "closest" to $\gamma(z_0)+\mathcal{O}(\varepsilon)$ in the sense of elapsed time for a solution leaving the neighborhood of $\gamma(z_0)+\mathcal{O}(\varepsilon)$. Similarly, there exists a curve on the plane Π, namely the intersection $W^s(\gamma(z_0)+\mathcal{O}(\varepsilon))\cap\Pi$, which is closest to $\gamma(z_0)+\mathcal{O}(\varepsilon)$ in the sense of elapsed time. We choose the unique point $\mathbf{q}^s_\varepsilon(0,\theta)$ on this curve such that $\mathbf{q}^u_\varepsilon(0,\theta)-\mathbf{q}^s_\varepsilon(0,\theta)$ is parallel to $(-f_2(\mathbf{q}_0(-\theta)), f_1(\mathbf{q}_0(-\theta)), 0)$. Thus we require $z^s_1(0,\theta)=z^u_1(0,\theta)$. We are guaranteed that such a choice of points can be made for each θ by Proposition 2.3 which says that the local perturbed manifolds are C^r ε-close to the local unperturbed manifolds in a neighborhood of $(\gamma(z_0)+O(\varepsilon), \phi)$. Thus their tangent spaces are ε-close. Outside of this neighborhood, solutions remain ε-close to unperturbed solutions for finite times, hence their maximum movement in the z-direction is $\mathcal{O}(\varepsilon)$. See Fig. 3.

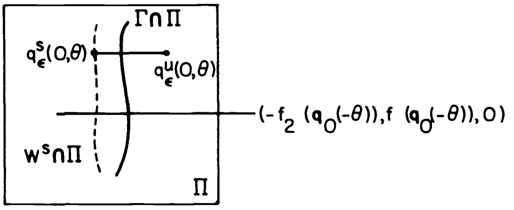


FIG. 3. *Intersections of the stable and unstable manifolds with* Π.

Clearly $|\mathbf{q}^u_\varepsilon(0,\theta)-\mathbf{q}^s_\varepsilon(0,\theta)|$ is a measure of the distance between $W^s(\gamma(z_0)+\mathcal{O}(\varepsilon))$ and $W^u(\gamma(z_0)+\mathcal{O}(\varepsilon))$. However, for easier computation and in order to account for the relative orientations between $W^s(\gamma(z_0)+\mathcal{O}(\varepsilon))$ and $W^u(\gamma(z_0)+\mathcal{O}(\varepsilon))$, we prefer to use the following distance measurement:

$$\begin{aligned} d(\alpha_0,\theta,z_0) &= \frac{(\partial H/\partial x(\mathbf{q}_0(-\theta)), \partial H/\partial y(\mathbf{q}_0(-\theta)), 0)\cdot(\mathbf{q}^u_\varepsilon(0,\theta)-\mathbf{q}^s_\varepsilon(0,\theta))}{\|\mathbf{f}(\mathbf{q}_0(-\theta))\|} \\ &= \frac{\varepsilon[(\partial H/\partial x(\mathbf{q}_0(-\theta)), \partial H/\partial y(\mathbf{q}_0(-\theta)), 0)\cdot(\mathbf{q}^u_1(0,\theta)-\mathbf{q}^s_1(0,\theta))]}{\|\mathbf{f}(\mathbf{q}_0(-\theta))\|}+\mathcal{O}(\varepsilon^2) \qquad (3.1) \\ &\overset{\text{def}}{=} \varepsilon\frac{M(\theta)}{\|\mathbf{f}(\mathbf{q}_0(-\theta))\|}+\mathcal{O}(\varepsilon^2) \end{aligned}$$

where "$\cdot$" is the usual vector dot product, $\|\cdot\|$ is the Euclidean norm, and $M(\theta)$ is defined to be the homoclinic *Melnikov function.*

We now develop a computable expression for $M(\theta)$. Recall that geometrically $M(\theta)$ is the lowest order term in an asymptotic expansion for the distance between the stable and unstable manifolds of a hyperbolic fixed point of a Poincaré map. We shall derive and solve a simple differential equation for a time dependent version of M, as in the standard planar Melnikov calculation.

Letting

$$\begin{aligned} \Delta(t,\theta) &= f_1(\mathbf{q}_0(-\theta))(y^u_1(t,\theta)-y^s_1(t,\theta))-f_2(\mathbf{q}_0(t-\theta))(x^u_1(t,\theta)-x^s_1(t,\theta)) \qquad (3.2) \\ &\overset{\text{def}}{=} \Delta^u(t,\theta)-\Delta^s(t,\theta) \end{aligned}$$

we compute

$$
\begin{aligned}
\dot{\Delta}^u(t,\theta) &= \left(\frac{\partial f_1}{\partial x}(\mathbf{q}_0(t-\theta)) + \frac{\partial f_2}{\partial y}(\mathbf{q}_0(t-\theta))\right)\Delta^u(t,\theta) \\
&\quad + f_1(\mathbf{q}_0(t-\theta))g_2(\mathbf{q}_0(t-\theta,t)) - f_2(\mathbf{q}_0(t-\theta))g_1(\mathbf{q}_0(t-\theta,t)) \\
&\quad + \left[f_1(\mathbf{q}_0(t-\theta))\frac{\partial f_2}{\partial z}(\mathbf{q}_0(t-\theta)) \right. \\
&\qquad \left. - f_2(\mathbf{q}_0(t-\theta))\frac{\partial f_1}{\partial z}(\mathbf{q}_0(t-\theta)) \right] z_1^u(t,\theta).
\end{aligned} \tag{3.3}
$$

Henceforth we suppress the arguments of the f_i, g_i and their partial derivatives. We have assumed the unperturbed vector field to be Hamiltonian, so that $(\partial f_1/\partial x) + (\partial f_2/\partial y) = 0$, and (3.3) becomes

$$
\dot{\Delta}^u(t,\theta) = f_1 g_2 - f_2 g_1 + \left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right) z_1^u(t,\theta), \tag{3.4}
$$

where $z_1^u(t,\theta)$ is obtained by solving the z component of the first variational equation:

$$
\dot{z}_1^u(t,\theta) = g_3(\mathbf{q}_0(t-\theta),t), \qquad t \in (-\infty, 0]. \tag{3.5}
$$

Equation (3.4) can be integrated immediately to give

$$
\Delta^u(0,\theta) - \Delta^u(-\infty,0) = \int_{-\infty}^{0} \left[f_1 g_2 - f_2 g_1 + \left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right) z_1^u(t,\theta)\right] dt. \tag{3.6}
$$

Similarly, we obtain an expression for $\dot{\Delta}^s(t,\theta)$, which leads to

$$
\Delta^s(\infty,\theta) - \Delta^s(0,\theta) = \int_{0}^{\infty} \left[(f_1 g_2 - f_2 g_1) + \left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right) z_1^s(t,\theta)\right] dt. \tag{3.7}
$$

Now $\Delta^s(\infty,\theta)$ and $\Delta^u(-\infty,\theta)$ are both zero since $\mathbf{q}_1^{u,s}(t,\theta)$ is bounded for all time (Proposition 2.3) and the unperturbed vector field goes to zero exponentially fast as $\gamma(z_0)$ is approached. Similarly, the improper integrals converge and we have

$$
\begin{aligned}
M(\theta) &= \int_{-\infty}^{-\infty} (f_1 g_2 - f_2 g_1)\, dt + \int_{-\infty}^{0} \left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right) z_1^u(t,\theta)\, dt \\
&\quad + \int_{0}^{\infty} \left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right) z_1^s(t,\theta)\, dt.
\end{aligned} \tag{3.8}
$$

Since we assume that the unperturbed vector field is Hamiltonian, the reader can easily verify that

$$
f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z} = \frac{\partial H}{\partial y}\frac{\partial^2 H}{\partial x\,\partial z} + \frac{\partial H}{\partial x}\frac{\partial^2 H}{\partial y\,\partial x} = -\frac{d}{dt}\left(\frac{\partial H}{\partial z}\right) + \frac{\partial^2 H}{\partial z^2}\dot{z}, \tag{3.9}
$$

and that

$$
\left(f_1 \frac{\partial f_2}{\partial z} - f_2 \frac{\partial f_1}{\partial z}\right)(\mathbf{q}_0(t-\theta)) = -\frac{d}{dt}\left(\frac{\partial H}{\partial z}(\mathbf{q}_0(t-\theta))\right), \tag{3.10}
$$

since $z = \text{constant}$ on an unperturbed orbit. Using this fact and integrating by parts once, we find

$$\int_{-\infty}^{0}\left(f_1\frac{\partial f_2}{\partial z}-f_2\frac{\partial f_1}{\partial z}\right)(\mathbf{q}_0(t-\theta),t)z_1^u(t,\theta)\,dt$$

$$(3.11)\qquad =\frac{\partial H}{\partial z}(\mathbf{q}_0(-\infty))z_1^u(-\infty,\theta)-\frac{\partial H}{\partial z}(\mathbf{q}_0(-\theta))z_1^u(0,\theta)$$

$$+\int_{-\infty}^{0}\frac{\partial H}{\partial z}(\mathbf{q}_0(t-\theta))g_3(\mathbf{q}_0(t-\theta),t)\,dt,$$

and similarly

$$\int_{0}^{\infty}\left(f_1\frac{\partial f_2}{\partial z}-f_2\frac{\partial f_1}{\partial z}\right)(\mathbf{q}_0(t-\theta))z_1^s(t,\theta)\,dt$$

$$(3.12)\qquad =\frac{\partial H}{\partial z}(\mathbf{q}_0(-\theta))z_1^s(0,\theta)-\frac{\partial H}{\partial z}(\mathbf{q}_0(\infty))z_1^s(\infty,\theta)$$

$$+\int_{0}^{\infty}\frac{\partial H}{\partial z}(\mathbf{q}_0(t-\theta))g_3(\mathbf{q}_0(t-\theta),t),\,dt.$$

Thus we have

$$(3.13)\qquad \begin{aligned} M(\theta)=&\int_{-\infty}^{+\infty}\left(f_1g_2-f_2g_1+\frac{\partial H}{\partial z}g_3\right)(\mathbf{q}_0(t-\theta),t)\,dt+\frac{\partial H}{\partial z}(\mathbf{q}_0(-\infty))z_1^u(-\infty,\theta)\\ &-\frac{\partial H}{\partial z}(\mathbf{q}_0(\infty))z_1^s(\infty,\theta)+\frac{\partial H}{\partial z}(\mathbf{q}_0(-\theta))(z_1^s(0,\theta)-z_1^u(0,\theta)).\end{aligned}$$

We note that $(\partial H/\partial z)(\mathbf{q}_0(-\infty))=(\partial H/\partial z)(\mathbf{q}_0(\infty))$, since the unperturbed orbit approaches $\gamma(z_0)$ for $t\to\pm\infty$, and that $z_1^u(-\infty,\theta)$ and $z_1^s(\infty,\theta)$ converge to the saddle point on the section Σ^0. See Robinson [1985] or Wiggins [1985] for a discussion of this limit process. It follows that

$$(3.14)\qquad \frac{\partial H}{\partial z}(q(-\infty))z_1^u(-\infty,\theta)-\frac{\partial H}{\partial z}(q(\infty))z_1^s(\infty,\theta)=0,$$

and by our original choice of $q_\varepsilon^s(0,\theta)$ and $q_\varepsilon^u(0,\theta)$ we have $z_1^s(0,\theta)-z_1^u(0,\theta)=0$; thus we arrive at an expression for the Melnikov function:

$$(3.15)\qquad M(\theta)=\int_{-\infty}^{\infty}\left(f_1g_2-f_2g_1+\frac{\partial H}{\partial z}g_3\right)(\mathbf{q}_0(t-\theta),t)\,dt.$$

Finally, using $f_1=\partial H/\partial y$, $f_2=-\partial H/\partial x$, and transforming $t\to t+\theta$, (3.15) can be rewritten in the compact form

$$(3.16)\qquad M(\theta)=\int_{-\infty}^{\infty}(\nabla H\cdot g)(\mathbf{q}_0(t),t+\theta)\,dt.$$

Now on the cross-section Σ^0, the distance between the stable and unstable manifolds of $\gamma(z_0)+\mathcal{O}(\varepsilon)$ is measured by $d(\theta)=\varepsilon(M(\theta)/\|f(\mathbf{q}_0(-\theta))\|)+\mathcal{O}(\varepsilon^2)$, so if $M(\theta)$ has a simple zero at $\hat\theta$ $(M(\hat\theta)=0,\ (dM/d\theta)(\hat\theta)\neq 0)$, then by the implicit function theorem, $d(\theta)$ also has a zero near $\hat\theta$. We have now proved the following.

THEOREM 3.1. *Suppose $M(\theta)$ has at least one simple zero; then for ε sufficiently small, near this point $W^s(\gamma(z_0)+\mathcal{O}(\varepsilon))$ and $W^u(\gamma(z_0)+\mathcal{O}(\varepsilon))$ intersect transversely. On the other hand, if $M(\theta)$ is bounded away from zero for all θ then $W^s(\gamma(z_0)+\mathcal{O}(\varepsilon))\cap W^u(\gamma(z_0)+\mathcal{O}(\varepsilon))=\varnothing$.*

This result is important since it allows us to test for transverse homoclinic points in the Poincaré map of specific differential equations. Thus, by the Smale–Birkhoff homoclinic theorem, we know that some iterate of the Poincaré map, $(P_\varepsilon)^N$, has an invariant hyperbolic set near such a point; i.e., a Smale horseshoe with its attendant chaotic dynamics (Guckenheimer and Holmes [1983, Chapter 5]).

Remark. In the formulation (3.14) $f_1g_2-f_2g_1 \overset{\text{def}}{=} (f\wedge g)_{1,2}$ is the usual planar Melnikov function, while $(\partial H/\partial z)g_3$ is the additional contribution due to the slow variation of z. We note that analogous expressions involving such extra terms occur in multidegree of freedom Hamiltonian examples, of Koiller [1984], Holmes and Marsden [1983] and Holmes [1986]. Also see Robinson [1983] and Gruendler [1985]. We remark that, if g is a time-independent perturbation, then, as formulation (3.16) makes clear, $M=\int_{-\infty}^{\infty}(\nabla H\cdot g)(\mathbf{q}_0(t))\,dt$ is θ-independent and thus simple zeros cannot occur and transversal intersections cannot be found. This is not surprising, since in that case we have an *autonomous* three-dimensional vector field and $\gamma(z_0)+\mathcal{O}(\varepsilon)$ is a fixed point with a one-dimensional unstable manifold and a two-dimensional stable manifold (or vice versa). If such manifolds intersect, they necessarily do so along a solution curve and thus the intersection cannot be transversal (cf. Guckenheimer and Holmes [1983, § 1.8]). However, if g depends upon parameters, then such "autonomous" homoclinic orbits can occur naturally as a parameter varies (see Theorem 4.2 below).

4. Bifurcations. In this section we give two theorems relevant to the case where the slowly varying oscillator depends upon a parameter $\mu\in R$.

THEOREM 4.1. (*Nonautonomous*). *Consider system* (2.1) *depending on a scalar parameter $\mu\in K$, where K is some open interval in R. Suppose there exists a point* (θ_0,μ_0) *such that*

(a) $M(\theta_0,\mu_0)=0,$

(b) $\left.\dfrac{\partial M}{\partial\theta}\right|_{(\theta_0,\mu_0)}=0.$

(c) $\left.\dfrac{\partial^2 M}{\partial\theta^2}\right|_{(\theta_0,\mu_0)}\neq 0,$

(d) $\left.\dfrac{\partial M}{\partial\mu}\right|_{(\theta_0,\mu_0)}\neq 0.$

Then, for $\varepsilon\neq 0$ sufficiently small, near μ_0 there is a bifurcation value $\hat{\mu}$ at which quadratic homoclinic tangencies occur.

As we have noted, if g is time-independent then M is necessarily θ-independent. Hence hypothesis (c) of Theorem 4.1 cannot be satisfied, and Theorem 3.1 cannot be applied. However, in this case we have the following.

THEOREM 4.2. (*Autonomous*). *Consider system* (2.1), *where* $\mathbf{g}(\mathbf{q};\mu)$ *is time independent but depends on a scalar parameter $\mu\in K\subseteq\mathbb{R}$. Suppose there exists a point $\mu_0\in K$ such that*

(a) $M(\mu_0)=0,$

(b) $\left.\dfrac{\partial M}{\partial\mu}\right|_{\mu=\mu_0}\neq 0.$

Then, for $\varepsilon\neq 0$ sufficiently small, near μ_0 there is a bifurcation value μ at which (*nontransverse*) *homoclinic orbits occur.*

Proofs. These two results are proved by straightforward Taylor series expansion of M about the point (θ_0, μ_0) (resp. μ_0) and application of the implicit function theorem. See Guckenheimer and Holmes [1983, § 4.5]. □

We remark that in the autonomous case it does not immediately follow that homoclinic orbits imply horseshoes, although that conclusion does follow for certain types of saddle-point with complex eigenvalues and in some cases with real eigenvalues (Silnikov [1965], [1967], [1970], Devaney [1976], Holmes [1980], Sparrow [1982]).

5. Interaction of periodic and homoclinic orbits. In the preceding paper (Wiggins and Holmes [1987]) we studied the two parameter family of periodic orbits inside Γ which remain bounded away from Γ, i.e., their periods were uniformly bounded above by some constant.

Now we will relate the periodic Melnikov theory to the homoclinic results of the present paper. From Wiggins and Holmes [1987, (2.7) and (3.11)-(3.19)], we have, omitting the arguments $\mathbf{q}_0^{\alpha,z}(t)$

$$\begin{aligned}\mathbf{M}^{m/n}(I, \theta, z) &= \left[\frac{1}{\Omega(1, Z)}\left\{\int_0^{mT} (\nabla H \cdot \mathbf{g})\, dt - \frac{\partial H}{\partial z}\bigg|_I \int_0^{mT} g_3\, dt\right\}, \int_0^{mT} g_3\, dt\right] \\ &\overset{\text{def}}{=} (M_1^{m/n}, M_3^{m/n}).\end{aligned} \tag{5.1}$$

In that paper (5.1) was obtained by a computation involving action angle variables; however, a cartesian (x, y, z) computation analogous to that of § 3 above yields precisely the same expression. To see this, refer to (3.13) of the present paper, use periodicity of $\mathbf{q}_0^{\alpha,z}(t)$, change the limits of integration to $-mT/2 \to mT/2$ and observe that $\Omega(I, z) = \|\mathbf{f}(\mathbf{q}_0^{\alpha,z}(-\theta))\|$. Finally, in dealing with periodic orbits the term $(\partial H/\partial z)(\mathbf{q}_0(-\infty))z_1^u(-\infty, \theta) - (\partial H/\partial z)(\mathbf{q}_0(\infty))z_1^s(\infty, \theta)$ becomes

$$\frac{\partial H}{\partial z}\bigg|_{I=\text{const}}\left(z_1\left(\frac{-mT}{2}, \theta\right) - z_1\left(\frac{mT}{2}, \theta\right)\right)\bigg| = \frac{-\partial H}{\partial z}\bigg|_I \int_{-mT/2}^{mT/2} g_3\, dt,$$

as required, and the other boundary term of (3.13) vanishes.

We now show that the limit of $\mathbf{M}^{m/1}$ exists on the homoclinic manifold and give an interpretation of that limit in terms of the dynamics of the slowly varying oscillators. It is clear that the homoclinic manifold cannot be approached in an arbitrary manner, since not all solutions in the perturbed manifolds can be uniformly approximated by solutions in the unperturbed manifolds for arbitrarily long time intervals: by Propositions 2.3 and 2.4 this can be done only for perturbed solutions with initial z values in $O(\varepsilon)$ neighborhood of a point $z_0 \in J$ such that $(\gamma(z_0) + O(\varepsilon), \phi)$ is a hyperbolic periodic orbit (or fixed point in the autonomous case) on M_ε. However, we have the following result.

PROPOSITION 5.1. *Suppose that there exists a point* $z_0 \in J$ *such that* $(\gamma(z_0) + \mathcal{O}(\varepsilon), \phi)$ *is a hyperbolic periodic orbit on* $\mathcal{M}_\varepsilon$. *Then*

(1) $$\lim_{\substack{m\to\infty \\ z\to z_0}} M_3^{m/1} = \int_{-\infty}^{\infty} g_3(q_0(t), t+\theta)\, dt = 0,$$

(2) $$\lim_{\substack{m\to\infty \\ z\to z_0}} M_1^{m/1} = \frac{1}{\|\mathbf{f}(\mathbf{q}_0(-\theta))\|}\int_{-\infty}^{\infty} (\nabla H \cdot \mathbf{g})(\mathbf{q}_0(t), t+\theta)\, dt = M(\theta),$$

where the integrands are evaluated on the unperturbed homoclinic orbit with $z = z_0$.

Proof. The proof of (1) follows by integrating the z_1 component of the first variational equation and examining the limiting behavior as $t \to \pm\infty$.

The proof of (2) involves a straightforward modification of arguments given in Theorem 4.6.4 of Guckenheimer and Holmes [1983]. □

This proposition shows that the periodic Melnikov functions have a meaning on the homoclinic manifold, although the dynamical interpretations are very different. Inside the homoclinic manifold, a simple zero of $(M_1^{m/1}, M_3^{m/1})$ implies the existence of an isolated periodic orbit of period mT. On the homoclinic manifold, a simple zero of M_3 implies a hyperbolic periodic orbit, $(\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)$, on $\mathcal{M}_\varepsilon$ and a simple zero of M_1 implies a transversal intersection between the stable and unstable manifolds of $(\gamma(z_0)+\mathcal{O}(\varepsilon), \phi)$. However, note that Proposition 5.1 allows us to think of M_1 and M_3 as functions of (α, z, θ) with $\theta \in R$, $z \in J$ and $\alpha \in L(z)=[\alpha(z), \alpha_0(z)]$, where $\alpha_0(z)$ is the value of α which gives an orbit on the homoclinic manifold for that particular z-value.

We end this section by remarking on the case where the system (2.1) is autonomous. In this case the limits of integration for the subharmonic Melnikov vector are $-T(\alpha, z)/2 \rightarrow T(\alpha, z)/2$ where $T(\alpha, z)$ is the period of an unperturbed orbit. In showing that the subharmonic Melnikov vector has a meaning on the homoclinic manifold in this case we take limits as $\alpha \rightarrow \alpha_0$, $z \rightarrow z_0$, where α_0 is the value of α on the homoclinic manifold and z_0 is a z value such that $\gamma(z_0)+\mathcal{O}(\varepsilon)$ is a hyperbolic fixed point on $\mathcal{M}_\varepsilon$.

Now we will show that the hypotheses of the homoclinic bifurcation Theorems 4.1 and 4.2 also imply the existence of nearby families of periodic orbits, which converge to the homoclinic orbits as $\mu \rightarrow \hat{\mu}$.

$M^{m/1}(I, z, \mu)$ and $M^{m/1}(I, \theta, z, \mu)$ denote the autonomous and nonautonomous subharmonic Melnikov vectors respectively (we will drop the superscript $m/1$ for notational convenience), unless $(I(\theta), z)$ belongs to the homoclinic orbit, in which case they denote the homoclinic Melnikov functions of Proposition 5.1. In that case, the requirement of a hyperbolic set in $\mathcal{M}_\varepsilon$ fixes $z=z_0$ (via $\bar{g}_3(\gamma(z_0)=0)$ and $I=I_0$ is fixed by the unperturbed homoclinic orbit on the plane $z=z_0$. Thus, in § 4 we merely wrote $M(\mu)$ for $M_1(I_0, z_0, \mu)$ and $M(\theta, \mu)$ for $M_1(I_0, \theta, z_0, \mu)$.

We remark that we have replaced the parameter α with I, the action. This is convenient since we are interested in periodic orbits limiting on homoclinic orbits and is justifiable by Proposition 5.1 and the fact that I represents the *area* bounded by an orbit on a fixed z plane, and this area is defined even for the homoclinic orbit.

Our two main results are the following.

THEOREM 5.2 (*Autonomous*). *Consider the parametrized Melnikov functions* $\mathbf{M}(I, z, \mu)$ *for a parameter* $\mu \in R$. *Suppose there exists* $z_0=z_0(\mu) \in J$ *such that* $\gamma(z_0(\mu), \mu)+\mathcal{O}(\varepsilon)$ *is a hyperbolic fixed point on* $\mathcal{M}_\varepsilon$ *for each* μ *in an open interval* K *containing a value* μ_0 *and let* $I=I_0$ *be the value of the action corresponding to the homoclinic orbit on the* $z=z_0(\mu)$ *level at which*

(a) $M_1(I_0, z_0, \mu_0)=0,$

(b) $\dfrac{\partial M_1}{\partial \mu}(I_0, z_0, \mu_0) \neq 0,$

(c) $\dfrac{\partial g_1}{\partial x}(\gamma(z_0, \mu_0))+\dfrac{\partial g_2}{\partial y}(\gamma(z_0, \mu_0)) \neq 0,$

(d) $\dfrac{\partial g_3}{\partial z}(\gamma(z_0, \mu_0)) \neq 0.$

Then for $\varepsilon \neq 0$ *sufficiently small the solutions of* (2.1) *contain a family* $\Lambda(\mu)$, *of periodic orbits* $(\mu \in K)$, *which converge on the homoclinic orbit with periods approaching infinity as* $\mu \rightarrow \hat{\mu}=\mu_0+\mathcal{O}(\varepsilon)$, *where* $\hat{\mu}$ *is the homoclinic bifurcation value.*

THEOREM 5.3 (*Nonautonomous*). *Consider the parametrized Melnikov functions* $\mathbf{M}(I, \theta, z, \mu)$ *for a parameter* $\mu \in R$. *Suppose there exists* $z_0 = z_0(\mu) \in J$ *such that* $(\gamma(z_0(\mu), \mu) + \mathcal{O}(\varepsilon), \phi)$ *is a hyperbolic periodic orbit on* $\mathcal{M}_\varepsilon$ *for each* μ *in an open interval* K *containing a value* μ_0, *and let* $I = I_0$ *be the value of the action corresponding to the homoclinic orbit on the* $z = z_0(\mu)$ *level such that at the point* $(I_0, \theta_0, z_0(\mu_0), \mu_0)$ *we have*

$$\text{(a)}\qquad M_1(I_1, \theta_0, z_0(\mu_0), \mu_0) = 0,$$

$$\text{(b)}\qquad \frac{\partial M_1}{\partial \theta}(I_1, \theta_0, z_0(\mu_0), \mu_0) = 0,$$

$$\text{(c)}\qquad \frac{\partial^2 M_1}{\partial \theta^2}(I_1, \theta_0, z_0(\mu_0), \mu_0) \neq 0,$$

$$\text{(d)}\qquad \frac{\partial M_1}{\partial \mu}(I_1, \theta_0, z_0(\mu_0), \mu_0) \neq 0,$$

$$\text{(e)}\qquad \frac{\partial \bar{g}_3}{\partial z}(\gamma(z_0(\mu_0), \mu_0)) \neq 0.$$

Then, for $\varepsilon \neq 0$ *sufficiently small, the homoclinic bifurcation is a countable limit of subharmonic saddle-node bifurcations to higher and higher periods.*

Proofs. The proofs of Theorems 5.2 and 5.3 involve straightforward, though tedious, calculations with the Melnikov functions. The interested reader is referred to Wiggins [1985] for the details. These results generalize the autonomous planar homoclinic bifurcation theorems of Andronov et al. [1971] and the nonautonomous planar Melnikov [1963] methods of Greenspan and Holmes [1983]. We remark that the theorems can also be proved using the more "geometric" arguments of Silnikov [1965], [1967], [1970], in which a local analysis near the hyperbolic set γ is combined with a near identity global return map (cf. Guckenheimer and Holmes [1983, § 6.5]). □

6. An example. In this section we apply the theory developed above to the equation

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - x^3 - z - \varepsilon\delta y, \qquad (6.1) \\ \dot{z} &= \varepsilon(\gamma x - \alpha z + \beta \cos t), \end{aligned}$$

which models a single degree of freedom nonlinear oscillator subject to weak linear damping and weak feedback control (Holmes and Moon [1983], Holmes [1983], [1985]). If $\beta = 0$, the system is autonomous and, for sufficiently strong damping ($\varepsilon\delta$) and feedback ($\varepsilon\gamma$) the feedback stabilizes the equilibrium position $(0, 0, 0)$, which is a saddle-point for small $\varepsilon\gamma$. In Holmes [1985] local bifurcation results were obtained for a slight variant of this system and in Holmes [1983] an ad hoc perturbation method was used to argue that transverse homoclinic orbits would occur in the nonautonomous ($\beta \neq 0$) case. (The term $\beta \cos t$ represents a desired response characteristic, which is relayed to the system via the feedback loop). This example therefore illustrates both the autonomous and nonautonomous theories developed above.

The unperturbed Hamiltonian corresponding to (6.1) is

$$H(x, y; z) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4} + xz, \qquad (6.2)$$

and straightforward analysis reveals the unperturbed phase space structure sketched in Fig. 4. A hyperbolic manifold $\mathcal{N}$, given by $x=\bar{x}(z)$, $y=0$, where $\bar{x}$ is the intermediate size root of

$$x^3-x+z=0, \tag{6.3}$$

exists for $-2/3\sqrt{3}<z<2/3\sqrt{3}$. For each $z=z_0$ in this range, the Hamiltonian system $H(x, y; z_0)$ restricted to $z=z_0$ has a hyperbolic saddle-point $(\bar{x}(z_0), 0)$ with a "figure 8" double homoclinic loop enclosing two elliptic fixed points (the other two roots of (6.3)). For $|z|>2/3\sqrt{3}$ (6.3) has a single root and H has a single elliptic fixed point.

First we apply Propositions 2.1 and 2.2 to this system, and seek periodic orbits for the perturbed flow in the manifold $\mathcal{M}_\varepsilon=(\mathcal{N}, \phi)+\mathcal{O}(\varepsilon)$. This necessitates computation of

$$\begin{aligned}\overline{g_3(\gamma(z))}&=\frac{1}{2\pi}\int_0^{2\pi}(\gamma\bar{x}(z)-\alpha z+\beta\cos t)\,dt\\&=\gamma\bar{x}(z)-\alpha z.\end{aligned} \tag{6.4}$$

We note that the same result obtains for $\beta=0$ or $\beta\neq 0$. We also check that

$$\frac{\partial(f_1, f_2)}{\partial(x, y)}=3x^2-1 \tag{6.5}$$

is strictly negative on $\mathcal{N}$, since the middle root of (6.3) lies in the range $(-1/\sqrt{3}, 1/\sqrt{3})$ for $-2/3\sqrt{3}<z<2/3\sqrt{3}$.

From (6.4), for a zero of $\bar{g}_3$ we require $z=\gamma\bar{x}(z)/\alpha$, and thus, using (6.3), we must solve

$$x^3+\left(\frac{\gamma}{\alpha}-1\right)x=0, \qquad \Rightarrow x=0 \quad\text{or}\quad x=\pm\sqrt{1-\frac{\gamma}{\alpha}}, \qquad \frac{\gamma}{\alpha}<1. \tag{6.6}$$

Also, note that the derivative

$$\frac{d}{dz}\bar{g}_3=\gamma\bar{x}'-\alpha=\frac{\gamma}{1-3\bar{x}^2}-\alpha, \tag{6.7}$$

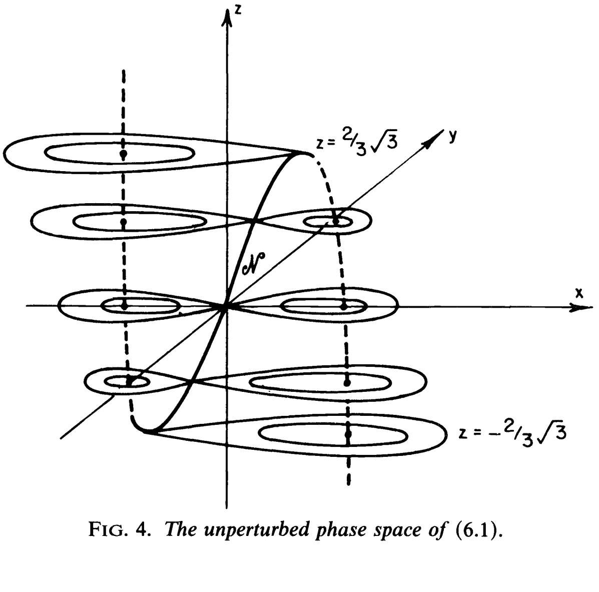


FIG. 4. *The unperturbed phase space of* (6.1).

takes the values $\gamma-\alpha$ for $x=0$ and $2\alpha(\alpha-\gamma)/(3\gamma-2\alpha)$ for $x=\pm\sqrt{1-\gamma/\alpha}$. Thus, for simple zeros we require $\gamma\neq\alpha$ at $x=0$ and $2\alpha/3<\gamma<\alpha$ at $x=\pm\sqrt{1-\gamma/\alpha}$. If these conditions are met, then Proposition 2.2 guarantees that the perturbed Poincaré map has hyperbolic fixed points near $(0,0,0)$ and $(\pm\sqrt{1-\gamma/\alpha},\ 0,\ \pm(\gamma/\alpha)\sqrt{1-\gamma/\alpha})$. Examination of the sign of $d\bar{g}_3/dz$ shows that the dynamics on $\mathcal{M}_\varepsilon$ is as sketched in Fig. 5.

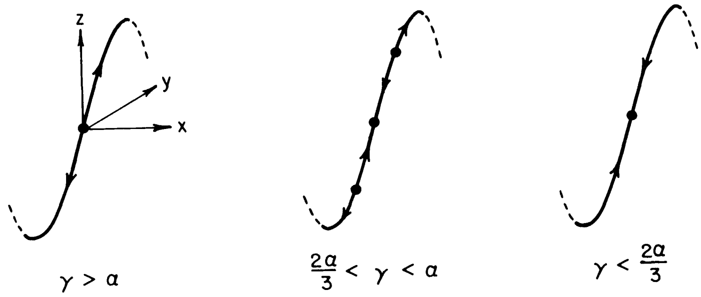

z
y
x
$\gamma > \alpha$
$\frac{2\alpha}{3} < \gamma < \alpha$
$\gamma < \frac{2\alpha}{3}$

FIG. 5. *Dynamics on* $\mathcal{M}_\varepsilon$, $\varepsilon\neq 0$.

We next apply the perturbation theory of § 3, computing the Melnikov functions for $z=0$ and $z=\pm(\gamma/\alpha)\sqrt{1-\gamma/\alpha}$. The unperturbed solutions on these z-planes are given by

$$\mathbf{q}_0(t)=(x,y,z)(t)=(\pm\sqrt{2}\operatorname{sech} t,\ \mp\sqrt{2}\operatorname{sech} t\tanh t,\ 0),\tag{6.8}$$

$$\mathbf{q}_0^+(t)=\left(\frac{2cS+ab}{2bS-a},\frac{-2ad^3ST}{(2bS-a)^2},\frac{\gamma}{\alpha}\sqrt{1-\frac{\gamma}{\alpha}}\right)\tag{6.9a}$$

and

$$\mathbf{q}_0^-(t)=\left(\frac{2cS-ab}{2bS+a},\frac{2ad^3ST}{(2bS+a)^2},\frac{\gamma}{\alpha}\sqrt{1-\frac{\gamma}{\alpha}}\right),\tag{6.9b}$$

on the plane $z=\gamma\sqrt{1-\gamma/\alpha}/\alpha$. Here we define

$$\begin{gathered}a=\sqrt{\frac{2\gamma}{\alpha}},\quad b=\sqrt{1-\frac{\gamma}{\alpha}},\quad c=1-\frac{2\gamma}{\alpha},\quad d=\sqrt{\frac{3\gamma}{\alpha}-2},\\ S=\operatorname{sech}(dt)\quad\text{and}\quad T=\tanh(dt).\end{gathered}\tag{6.10}$$

On $z=-\gamma\sqrt{1-\gamma/\alpha}/\alpha$ we have $\bar{\mathbf{q}}_0^+=-\mathbf{q}_0^+$ and $\bar{\mathbf{q}}_0^-=-\mathbf{q}_0^-$. Note that, as $\gamma/\alpha\to 1$, (6.9a, b) → (6.8).

The Melnikov integral is

$$\int_{-\infty}^{\infty}(\nabla H\cdot g)(\mathbf{q}_0(t),t+\theta)\,dt=\int_{-\infty}^{\infty}(-\delta y^2+\gamma x^2-\alpha xz+\beta x\cos(t+\theta))\,dt.\tag{6.11}$$

In the first case this yields

$$\begin{aligned}M&=\int_{-\infty}^{\infty}(-2\delta\operatorname{sech}^2 t\tanh^2 t+2\gamma\operatorname{sech}^2 t\pm\sqrt{2}\beta\operatorname{sech} t\cos(t+\theta))\,dt\\&=-\frac{4\delta}{3}+4\gamma\pm\sqrt{2}\beta\pi\operatorname{sech}\left(\frac{\pi}{2}\right)\cos\theta.\end{aligned}\tag{6.12}$$

In the second case, on the plane $z=(\gamma/\alpha)\sqrt{(1-(\gamma/\alpha))}$, we obtain the expressions

$$\begin{aligned}M=-4\delta\left[\frac{d^4}{3}+b^2d^2+\frac{\gamma bd}{\sqrt{2}\alpha}\left(\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}b\pm\frac{\pi}{2}\right)\right]\\ +2\gamma\left[2d+\sqrt{2}b\left(\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}b\pm\frac{\pi}{2}\right)\right]\mp 2\sqrt{2}\pi\beta\frac{\sinh(1/d\sin^{-1}\sqrt{d\alpha/\gamma})}{\sinh(\pi/d)}\cos\theta,\end{aligned}\tag{6.13}$$

where the upper choice of sign refers to the larger homoclinic loop q_0^+ and the lower choice to the smaller loop, q_0^- (cf. Fig. 5). On $z=-(\gamma/\alpha)\sqrt{1-(\gamma/\alpha)}$, we find

$$\begin{aligned}M=-4\delta\left[\frac{d^4}{3}+b^2d^2+\frac{\gamma bd}{\sqrt{2}\alpha}\left(\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}b\pm\frac{\pi}{2}\right)\right]\\ +2\gamma\Bigl[2d+\sqrt{2}b\Bigl[\Bigl(\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}b\pm\frac{\pi}{2}\Bigr)\Bigr]\pm 2\sqrt{2}\pi\beta\frac{\sinh(1/d\sin^{-1}\sqrt{d\alpha/\gamma})}{\sinh(\pi/d)}\cos\theta.\end{aligned}\tag{6.14}$$

In all four cases the principal value $0\leqq\sin^{-1}(\cdot)\leqq\pi/2$ is to be taken. We note that, when $\gamma/\alpha=1$, so that $b=0$ and $d=1$, both (6.13) and (6.14) reduce to (6.12).

We present the bifurcation results that follow from these computations and Theorem 4.2, for the autonomous case, ($\beta=0$) in Fig. 6(a), where we show the bifurcation sets $M(\delta,\gamma)=0$ for fixed $\alpha=1$ computed from (6.12)-(6.14) using the definitions of a, b, c, d in (6.10). The linear set (6.12) $\delta=3\gamma$ and the two curves from $(6.13)^\pm$ are indicated on the figure. For $\beta=0$ $(6.14)^\pm$ gives curves coincident with those of $(6.13)^\pm$. Note that as $\gamma/\alpha\to 1^-$ (where the three fixed points on M_ε coalesce) all three curves meet, and also that the curves for the homoclinic orbits near $z=\pm(\gamma/\alpha)(1-(\gamma/\alpha))^{1/2}$ go to infinity as $\gamma\to 2/3^+$ (where the two nontrivial fixed points reach the boundary of M_ε). We remark that a branch of the curve labeled $(6.13)^-$ and $(6.14)^-$ has not been shown in Fig. 6(a) since it assumes δ values outside the range of our graph for γ values of physical interest. Figure 7 gives schematic phase portraits corresponding to parameter values labeled in Fig. 6(a).

In the nonautonomous case ($\beta\neq 0$) we see that the effect of the nonautonomous perturbation β cost is to open each of these curves into a band of width $\mathcal{O}(\beta)$ (see Fig. 6(b)). By Theorem 4.1 we conclude that quadratic homoclinic tangencies occur

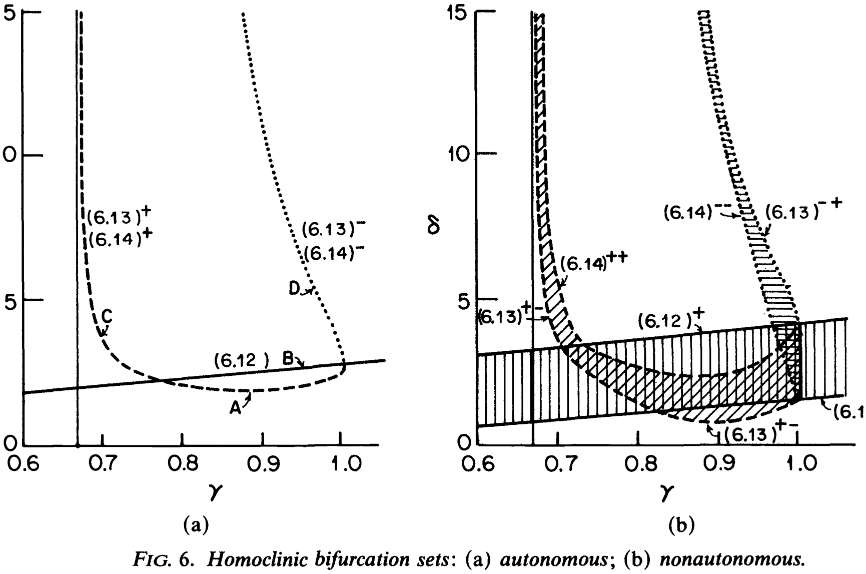


FIG. 6. *Homoclinic bifurcation sets*: (a) *autonomous*; (b) *nonautonomous*.

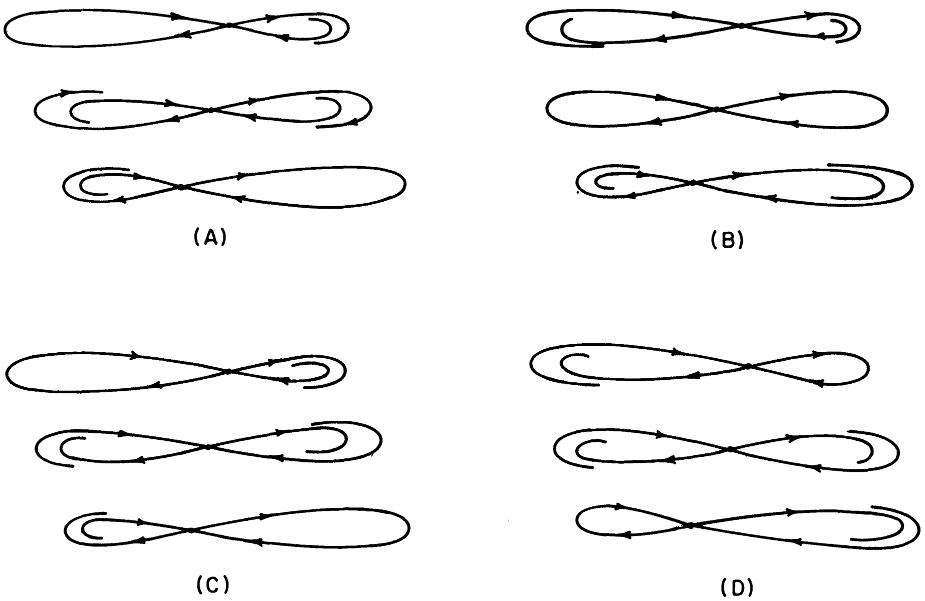


FIG. 7. *Phase portraits at points* A, B, C, D of *Fig.* 6(a).

on the boundaries of these bands with transversal intersections inside. Therefore, from Theorem 5.3 we know that the points on the boundaries of these bands are countable limits of saddle-node bifurcation points of periodic orbits to higher and higher periods. In this case, then, we have deterministic chaos for parameter values in the bands indicated in Fig. 6(b).

We conclude this section by remarking on the situation that occurs when the gain, γ, goes to zero. From (6.1) we see that the z component of the vector field decouples from the x and y components. Thus z can be solved for as an explicit function of time which is asymptotically periodic ($z \sim \varepsilon\beta \sin t + O(\varepsilon^2)$ as $t \to \infty$), and this solution can be substituted into the x and y components, resulting in an equation for a planar forced oscillator. Then one would expect to recover the usual Melnikov function for the equation

$$\dot{x} = y, \qquad \dot{y} = x - x^3 - \varepsilon[\beta \sin t + \delta y] + O(\varepsilon^2)$$

as studied by Greenspan and Holmes [1983], and inspection of (6.12) shows that this is indeed the case. In this respect, we note that the gain γ acts as a destabilizing influence resulting in the effective damping $((4\delta/3) - 4\gamma)$ in (6.12) in comparison with the term $4\delta/3$ in the uncoupled "planar" Duffing equation. Consequently the critical force level for the appearance of transverse homoclinic orbits and chaos is

$$\beta_{\text{crit}} = \frac{((4\delta/3) - 4\gamma)}{\sqrt{2}\pi} \cosh\left(\frac{\pi}{2}\right)$$

rather than

$$\beta_{\text{crit}} = \frac{4\delta}{3\sqrt{2}\pi} \cosh\left(\frac{\pi}{2}\right).$$

These results go some way in explaining the destabilizing effect of gain observed in numerical integrations of this and similar systems by Moon (see Holmes and Moon [1983]).

7. Conclusions. In this and the preceding paper we have developed a global perturbation theory for slowly varying oscillators that collapse to one parameter families of Hamiltonian systems in the limit $\varepsilon = 0$. As such, they typically possess two parameter (α, z) families of periodic orbits and one parameter (z) families of homoclinic orbits to hyperbolic manifolds of equilibria. The perturbation theory we have developed uses these highly degenerate structures to seek isolated periodic and homoclinic orbits for $\varepsilon \neq 0$, small. We have given existence, stability and codimension one bifurcation theorems for periodic orbits in resonance with an external forcing and an existence theorem for transverse homoclinic orbits in the nonautonomous case and homoclinic bifurcation theorems for both cases. The hypotheses of the theorems can be checked explicitly in examples by computations involving integration around the unperturbed closed orbits. We have illustrated such computations with examples of a nonlinear oscillator subject to weak feedback control and external forcing.

In the interests of providing detailed results and specific applications, we have chosen to limit our analyses to three-dimensional systems, but we remark that the methods generalize in a natural way to systems in which x and y are each n-dimensional and z is m-dimensional: i.e., slowly varying perturbations of m-parameter families of n-degree of freedom Hamiltonians (see Wiggins [1986]).

REFERENCES

A. A. Andronov, E. A. Leontovich, I. I. Gordon and A. G. Maier (1971), *Theory of Bifurcations of Dynamic Systems on a Plane*, Jerusalem, Israel Program For Scientific Translations.

R. L. Devaney (1976), *Homoclinic orbits in Hamiltonian systems*, J. Differential Equations, 21, pp. 431-438.

N. Fenichel (1971), *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21, pp. 193-226.

B. D. Greenspan and P. J. Holmes (1983), *Homoclinic orbits, subharmonics and global bifurcations in forced oscillations*, in Nonlinear Dynamics and Turbulence, G. Barenblatt, G. Iooss and D. D. Joseph, eds., Pitman, London, pp. 172-214.

J. Gruendler (1982), *A generalization of the method of Melnikov to arbitrary dimension*, Ph.D. thesis, Univ. of North Carolina, Chapel Hill, NC.

——— (1985), *The existence of homoclinic orbits and the method of Melnikov for systems in* $\mathbb{R}^n$, this Journal, 16, pp. 907-931.

J. Guckenheimer and P. J. Holmes (1983), *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, Berlin, Heidelberg.

J. K. Hale (1969), *Ordinary Differential Equations*, John Wiley, New York.

M. W. Hirsch, C. C. Pugh and M. Shub (1977), *Invariant Manifolds*, Lecture Notes in Mathematics No. 583, Springer-Verlag, New York, Berlin, Heidelberg.

P. J. Holmes (1980), *A strange family of three-dimensional vector fields near a degenerate singularity*, J. Differential Equations, 37, pp. 382-404.

——— (1983), *Bifurcation and chaos in a simple feedback control system*, Proc. 22nd IEEE Conference on Control and Decision, paper WP5, Vol. I, pp. 365-370.

——— (1985), *Dynamics of a nonliner oscillator with feedback control*, I: *Local analysis*, Trans. ASME J. Dyn. Systems, Meas. Control.

——— (1986), *Chaotic motions in a weakly nonlinear model for surface waves*, J. Fluid Mech., 162, pp. 365-388.

P. J. Holmes and J. E. Marsden (1983), *Horseshoes and Arnold diffusion for Hamiltonian systems on Lie groups*, Indiana Univ. Math. J., 32, pp. 273-310.

P. J. Holmes and F. C. Moon (1983), *Strange attractors and chaos in nonlinear mechanics*, Trans. ASME Ser. E, J. Appl. Mech., 50, pp. 1021-1032.

J. Koiller (1984), *Notes on a general Melnikov criterion*, unpublished.

N. Kopell (1985), *Invariant manifolds and the initialization problem for some atmospheric equations*, Phys. D., 14, pp. 203-215.

V. K. MELNIKOV (1963), *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12, pp. 1-57.

C. ROBINSON (1983), *Sustained resonance for a nonlinear system with slowly varying coefficients*, this Journal, 14, pp. 847-860.

——— (1985), *Horseshoes for autonomous Hamiltonian systems using the Melnikov integral*, preprint, Northwestern Univ.

L. P. SILNIKOV (1965), *A case of the existence of a denumerable set of periodic motions*, Soviet Math. Dokl., 6, pp. 163-166.

——— (1967), *The existence of a denumerable set of periodic motions in four-dimensional space in an extended neighborhood of a saddle-focus*, Soviet Math. Dokl., 8, pp. 54-58.

——— (1970), *A contribution to the problem of the structure of an extended neighborhood of a rough equilibrium state of saddle-focus type*, Math. USSR-Sb., 10, pp. 91-102.

C. T. SPARROW (1982), *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*, Springer-Verlag, New York, Berlin, Heidelberg.

S. WIGGINS AND P. HOLMES (1987), *Periodic orbits in slowly varying oscillators*, this Journal, 18, pp. 592-611.

S. WIGGINS (1985), *Slowly varying oscillators*, Ph.D. thesis, Cornell Univ., Ithaca, NY.

——— (1986), *A generalization of the method of Melnikov for the detection of chaotic invariant sets*, preprint, California Inst. of Technology.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

004

TOPOLOGICAL TRANSVERSALITY: APPLICATIONS TO THIRD ORDER BOUNDARY VALUE PROBLEMS*

DANIEL J. O'REGAN†

Abstract. In this paper we use topological transversality to obtain existence theorems for certain classes of third order boundary value problems. Our analysis is based on the notions of an essential map and on a priori bounds on solutions.

Key words. topological transverality, third order boundary value problems, higher order equations

AMS(MOS) subject classifications. 34B10, 34B15

1. Introduction. In this paper we study the existence of solutions to third order boundary value problems of the form

$$(1.1) \qquad y'''=f(t,y,y',y''), \qquad t\in[0,1], y\in B$$

where $f:[0,1]\times R^3\to R$ is continuous. Here B denotes suitable boundary conditions.

In §2 results of Granas, Guenther and Lee [4], [5] on second order boundary value problems are extended so that existence theorems can be obtained for a certain class of third order boundary value problems. The existence theorems obtained in § 2, however, are rather specialized. In § 3 by placing different types of monotonicity and growth conditions on the nonlinearity f, we obtain new and interesting existence theorems for a wide class of problems.

2. The Bernstein theory of the equation $y'''=f(t,y,y',y'')$. In this section we extend the Bernstein theory and results of Granas, Guenther and Lee [4] to discuss problems of the form (1.1). Fix a point c in $[0,1]$. Let B denote either the boundary conditions

$$\text{(i)} \qquad y(c)=0, \quad y'(0)=0, \quad y'(1)=0,$$

$$\text{(ii)} \qquad y(c)=0, \quad y''(0)=0, \quad y''(1)=0$$

or

$$\text{(iii)} \qquad \begin{aligned} &sy(c)+dy'(c)=0, && s\neq 0,\\ &-\alpha y'(0)+\beta y''(0)=0, && \alpha,\beta>0,\\ &ay'(1)+by''(1)=0, && a,b>0.\end{aligned}$$

Theorem 2.1 of [5] was proved for two point boundary value problems; however, no change in the proof is necessary if we consider multipoint boundary value problems. Hence, specializing Theorem 2.1 of [5] for the case $n=3$ we obtain the following.

THEOREM 2.1. *Let $f:[0,1]\times R^3\to R$ be continuous and $0\leqq\lambda\leqq 1$. Suppose there is a constant K independent of λ such that $\|y\|_3\leqq K$ for each solution $y(t)$ to*

$$(2.1)_\lambda \qquad y'''-y'=\lambda[f(t,y,y',y'')-y'], \qquad t\in[0,1], \quad y\in B.$$

Then the boundary value problem (1.1) *has at least one solution in* $C^3[0,1]$.

Suppose $y(t)$ is a solution to (1.1) and $[y'(t)]^2$ has a maximum at $t_0\in(0,1)$. Then

$$(2.2) \qquad y''(t_0)=0 \quad \text{and} \quad y'(t_0)f(t_0,y(t_0),y'(t_0),0)\leqq 0.$$

* Received by the editors October 20, 1985; accepted for publication (in revised form) April 22, 1986.

† Department of Mathematics and Applied Statistics, University of Idaho, Moscow, Idaho 83843.

THEOREM 2.2. *Suppose there is a constant $M \geqq 0$ such that*

$$pf(t, u, p, 0) > 0 \quad \text{for } |p| > M,$$

and (t, u) in $[0, 1] \times R$.

Then any solution y to (1.1) *satisfies*

$$|y'(t)| \leqq M \quad \text{for } t \in [0, 1].$$

Furthermore, there exists a constant $M_1 \equiv M[1 + |d/s|]$ such that

$$|y(t)| \leqq M_1 \quad \text{for } t \in [0, 1].$$

Proof. Suppose first $|y'|$ achieves a maximum at $t_0 \in (0, 1)$. Assume $|y'(t_0)| > M$. Then $y'(t_0)f(t_0, y(t_0), y'(t_0), 0) > 0$, which contradicts (2.2). Thus $|y'(t_0)| \leqq M_0$.

Now if y satisfies (i) or (iii), an easy argument shows $|y'|$ cannot have a nontrivial maximum at 0 or 1.

Finally, if y is a solution to (ii) and if $|y'|$ assumes its maximum at $t_0 = 0$ or $t_0 = 1$, then $|y'(t_0)| \leqq M$. To see this suppose $|y'(0)|$ is the maximum value of $|y'|$. If we assume $|y'(0)| > M$, then $y'(0)y'''(0) > 0$. Now if $y'(0) > 0$, then $y'''(0) > 0$, so $y''(t) = \int_0^t y'''(z)\, dz$ is strictly increasing near $t = 0$. We then have $y''(t) > y''(0) = 0$ for $t > 0$ and near zero and so $|y'(0)| = y'(0)$ is not the maximum of $|y'|$ on $[0, 1]$, a contradiction. We obtain a similar contradiction if we assume $y'(0) < 0$. Hence $|y'(t)| \leqq M$ for $t \in [0, 1]$. Finally, integration yields the stated bounds on $|y|$ immediately.

We couple the monotonicity condition in Theorem 2.2 with the analogue of the Bernstein growth condition to obtain our basic existence theorem.

THEOREM 2.3. *Let $f: [0, 1] \times R^3 \to R$ be continuous.*

(a) *Suppose there is a constant $M \geqq 0$ such that*

$$pf(t, u, p, 0) > 0 \quad \text{for } |p| > M$$

and $(t, u) \in [0, 1] \times R$.

(b) *Suppose that*

$$|f(t, u, p, q)| \leqq A(t, u, p)q^2 + B(t, u, p)$$

where $A(t, u, p)$, $B(t, u, p) \geqq 0$ are functions bounded on bounded (t, u, p) sets.

Then the boundary value problem (1.1) *has at least one solution in $C^3[0, 1]$.*

Proof. Existence follows immediately from Theorem 2.1 once a priori bounds are established for solutions y to $(2.1)_\lambda$. If $\lambda = 0$, $y \equiv 0$. Otherwise for $0 < \lambda \leqq 1$, $pf(t, u, p, 0) > 0$ for $|p| > M$ implies $\lambda pf(t, u, p, 0) + (1 - \lambda)p^2 > 0$ for $|p| > M$. Thus Theorem 2.2 yields a priori bounds M, M_1 for $|y'|$ and $|y|$ respectively. Finally we obtain a priori bounds on y'' and y'''. Now each of the boundary conditions (i), (ii), or (iii) implies that y'' vanishes at least once on $[0, 1]$. We also have

$$|f(t, u, p, q)| \leqq Aq^2 + B$$

where A and B denote upper bounds of $A(t, u, p)$, $B(t, u, p)$ respectively for $(t, u, p) \in [0, 1] \times [-M_1, M_1] \times [-M, M]$. Now each point $t \in [0, 1]$ for which $y''(t) \neq 0$ belongs to an interval $[\mu, v]$ such that y'' maintains a fixed sign on $[\mu, v]$ and $y''(\mu)$ and/or $y''(v)$ is zero. Assume that $y''(\mu) = 0$ and $y'' \geqq 0$ on $[\mu, v]$. Now with $A_0 = A$, $B_0 = B + M$ the differential equation yields

$$\frac{2A_0 y'' y'''}{A_0(y'')^2 + B_0} \leqq 2A_0 y''.$$

Integrating from μ to t we obtain

$$|y''(t)| \leqq \left[\frac{B_0}{A_0}(e^{4A_0M} - 1)\right]^{1/2} \equiv M_2.$$

The other cases are treated similarly and the same bound is obtained. With these bounds the differential equation yields a priori bounds independent of λ for $|y'''|$.

COROLLARY 2.4. *Let* $f:[0,1]\times R^3\to R$ *be continuous.*

(a) *Suppose there is a constant* $M\geqq 0$ *such that*

$$pf(t,u,p,0)\geqq 0 \quad \text{for } |p|>M$$

and (t,u) *in* $[0,1]\times R$.

(b) *Suppose* (b) *of Theorem* 2.3 *holds.*

Then the boundary value problem (1.1) *has at least one solution in* $C^3[0,1]$.

Proof. Consider

$$y'''=f_n(t,y,y',y''),\qquad t\in[0,1],\quad y\in B \tag{2.3}$$

where $f_n=f+y'/n$, n an integer. Apply Theorem 2.3 to (2.3) to obtain solutions y_n to (2.3) for $n=1,2,\cdots$. Now a simple compactness argument implies that a subsequence of $\{y_n\}$ converges to a solution of $y'''=f(t,y,y',y'')$, $y\in B$ and the proof is complete. □

To conclude this section we examine the inhomogeneous boundary value problems

$$y'''=f(t,y,y',y''),\qquad t\in[0,1],\quad y\in\tilde B \tag{2.4}$$

where $\tilde B$ denotes either the boundary conditions

$$y(c)=r,\quad y'(0)=l,\quad y'(1)=T \tag{iv}$$

or

$$\begin{aligned} &sy(c)+dy'(c)=r, && s\neq 0,\\ &-\alpha y'(0)+\beta y''(0)=l, && \alpha,\beta>0,\\ &ay'(1)+by''(1)=T, && a,b>0. \end{aligned} \tag{v}$$

Here c is a fixed point of $[0,1]$.

THEOREM 2.5. *Let* $f:[0,1]\times R^3\to R$ *be continuous. Suppose* (a) *and* (b) *of Corollary* 2.4 *are also satisfied. Then the boundary value problem* (2.4) *has at least one solution in* $C^3[0,1]$.

Proof. Consider the family of problems

$$y'''=\lambda f(t,y,y',y''),\qquad 0\leqq\lambda\leqq 1,\quad y\in\tilde B. \tag{2.4$_\lambda$}$$

The existence of a solution in $C^3[0,1]$ follows immediately from Theorem 5.1 of [7] once a priori bounds independent of λ are established for solutions y to $(2.4)_\lambda$. We assume at first that $pf(t,u,p,0)>0$ for $|p|>M$ and (t,u) in $[0,1]\times R$. Now if $\lambda=0$ we have a unique solution, and thus $|y'(t)|\leqq L$ for some constant $L<\infty$. Otherwise for $0<\lambda\leqq 1$, it follows immediately from Theorem 2.2 that $|y'|\leqq M_0=\max\{M,|l|,|T|\}$, if y satisfies (iv) and $|y'|\leqq M_1=\max\{M,|l/\alpha|,|T/a|\}$ if y satisfies (v). As an example suppose y satisfies (v) and $|y'(t)|$ assumes its maximum at $t=0$. Then $y'(0)y''(0)\leqq 0$ so

$$0\geqq y'(0)\beta y''(0)=\alpha[y'(0)]^2\left[\frac{l}{\alpha y'(0)}+1\right]$$

and consequently $|y'(0)|\leqq|l/\alpha|$. Hence a priori bounds for $|y'|$ and $|y|$ are immediate. A priori bounds for y'' and y''' follow exactly as in the proof of Theorem 2.3 once we observe that

$$|y''(\mu)|\leqq K,$$

$K\geqq 0$ a fixed constant independent of λ, for some point $\mu\in[0,1]$.

Now assume $pf(t,u,p,0)\geqq 0$ for $|p|>M$ and (t,u) in $[0,1]\times R$. The existence of a solution in this case follows by an argument similar to Corollary 2.4.

Example 1. (Sandwich beam). Beams formed by a few lamina of different materials are known as sandwich beams. In the analysis of such beams Krajcinovic [9] found that the distribution of shear deformation ψ is governed by the differential equation

$$\psi''' - k^2(x, \psi)\psi' + a(x, \psi) = 0.$$

Here $k^2 \neq 0$. For further information on k^2 and a see Krajcinovic [9].

For the case of free ends, the condition of zero shear bimoment at both ends leads to the boundary condition $\psi'(0) = \psi'(1) = 0$. Also symmetry considerations yields $\psi(1/2) = 0$. Thus we are interested in solving the boundary value problem

$$\text{(2.5)} \qquad \begin{aligned} &\psi''' = k^2(x, \psi)\psi' - a(x, \psi), \qquad x \in [0, 1], \\ &\psi'(0) = \psi'(1) = \psi(1/2) = 0. \end{aligned}$$

Now we make the following assumptions on k and a.

Suppose $k^2(x, \omega)$ and $a(x, \psi)$ are continuous functions on $[0, 1] \times R$. In addition, suppose there exists a constant $L < \infty$ such that

$$\left| \frac{a(x, \psi)}{k^2(x, \psi)} \right| \leqq L \quad \text{for } (x, \psi) \in [0, 1] \times R.$$

Then $\psi' f(x, \psi, \psi', 0) = \psi'(k^2\psi' - a) > 0$ for $|\psi'| > L$ and $(x, \psi) \in [0, 1] \times R$, and so Theorem 2.3 implies that (2.5) has at least one solution in $C^3[0, 1]$.

3. Another approach to third order boundary value problems. In this section we place essentially different types of monotonicity and growth conditions on the nonlinearity f to obtain existence theorems for a wide class of third order problems. We consider problems of the form

$$\text{(3.1)} \qquad y''' = f(t, y, y', y''), \qquad t \in [0, 1], y \in B_0$$

where $f: [0, 1] \times R^3 \to R$ is continuous. Here B_0 denotes either the boundary conditions

$$\text{(vi)} \qquad y(0) = 0, \quad y(1) = 0, \quad y'(0) = 0,$$

$$\text{(vii)} \qquad y(0) = 0, \quad y'(0) = 0, \quad y'(1) = 0,$$

or

$$\text{(viii)} \qquad \begin{aligned} &-\alpha y(0) + \beta y'(0) = 0, \quad \alpha, \beta > 0, \\ &ay(1) + by'(1) = 0, \quad a, b > 0, \\ &y''(0) = 0. \end{aligned}$$

Remark. It should be noted here that many of the boundary conditions in § 2 will also be considered in this section. The behaviour of the nonlinearity f will determine which existence theorem to use.

The following theorem, although not the main result in this section, is a powerful existence theorem in its own right.

THEOREM 3.1. *Let* $f: [0, 1] \times R^3 \to R$ *be continuous and* $0 \leqq \lambda \leqq 1$. *Suppose*

$$|f(t, u, p, q)| \leqq [A(t, u)|p| + B(t, u)][C(t, u)|q| + D(t, u)]$$

where $A(t, u)$, $B(t, u)$, $C(t, u)$, $D(t, u) \geqq 0$ are functions bounded on bounded (t, u) sets. Finally, we assume that there is a constant M such that

$$|y(t)| \leqq M, \qquad t \in [0, 1]$$

for each solution $y(t)$ to

$$(3.1)_\lambda \qquad y''' = \lambda f(t, y, y', y''), \qquad t \in [0, 1], \quad y \in B_0.$$

Then the boundary value problem (3.1) *has at least one solution in* $C^3[0, 1]$.

Proof. To prove existence of a solution in $C^3[0, 1]$ we apply Theorem 2.1 of [5]. To establish a priori bounds for $(3.1)_\lambda$, let $y(t)$ be a solution to $(3.1)_\lambda$. All that remains is to obtain a priori bounds for y', y'' and y'''. We first observe that boundary conditions (vi), (vii) or (viii) imply that y'' vanishes at least once on $[0, 1]$. We also have

$$|f(t, u, p, q)| \leqq [A|p| + B][C|q| + D]$$

where A, B, C and D denote upper bounds of $A(t, u)$, $B(t, u)$, $C(t, u)$, $D(t, u)$ respectively for $(t, u) \in [0, 1] \times [-M, M]$, and so

$$|\lambda f(t, y, y', y'')| \leqq (A|y'| + B)(C|y''| + D).$$

Now each point $t \in [0, 1]$ for which $y''(t) \neq 0$ belongs to an interval $[\mu, v]$ such that y'' maintains a fixed sign on $[\mu, v]$ and $y''(\mu)$ and/or $y''(v)$ is zero. Assume $y''(\mu) = 0$ and $y'' \geqq 0$ on $[\mu, v]$. Then the differential equation yields

$$(3.2) \qquad \frac{y'''}{Cy'' + D} \leqq A|y'| + B.$$

Also since $y'' \geqq 0$ on $[\mu, v]$ we have y' increasing on $[\mu, v]$, so in particular $y'(s) \geqq y'(\mu)$ for $s \in [\mu, v]$. At this stage of the proof the argument breaks up into two cases, $y'(\mu) \geqq 0$ and $y'(\mu) < 0$. Assume at first $y'(\mu) \geqq 0$, and so $y'(s) \geqq 0$ for $s \in [\mu, v]$. It follows from (3.2) that

$$\frac{Cy'''}{Cy'' + D} \leqq ACy' + BC,$$

and so integrating from μ to t we obtain

$$|y''(t)| \leqq \frac{D}{C}[\exp(2ACM + BC) - 1] \equiv M_0.$$

On the other hand assume $y'(\mu) < 0$. Again the argument breaks up into two subcases, $y'(s) \leqq 0$ for $s \in [\mu, t]$ or there exists $\zeta \in (\mu, t)$ such that $y'(\zeta) = 0$ and $y'(s) > 0$ for $s \in (\zeta, v]$. Suppose at first $y'(s) \leqq 0$ for $s \in [\mu, t]$; then (3.2) implies

$$\frac{Cy'''}{Cy'' + D} \leqq -ACy' + BC,$$

and so integrating from μ to t yields

$$|y''(t)| \leqq M_0,$$

as before. Finally suppose there exists $\zeta \in (\mu, t)$ such that $y'(\zeta) = 0$ and $y'(s) > 0$ for $s \in (\zeta, v]$; then for $\eta \in [\mu, \zeta]$, (3.2) implies

$$\frac{Cy'''}{Cy'' + D} \leqq -ACy' + BC,$$

which yields $|y''(\zeta)| \leqq M_0$. So for $s \in [\zeta, v]$, (3.2) again implies

$$\frac{Cy'''}{Cy''+D} \leqq ACy' + BC,$$

and thus

$$|y''(t)| \leqq \frac{[CM_0 + D] \exp(2ACM + BC) - D}{C} \equiv M_1.$$

Thus $|y''(t)| \leqq M_1$. The other cases are treated similarly and the same bound $M_1 = \max\{M_0, M_1\}$ is obtained. Thus $|y''| \leqq M_1$ for each solution y to $(3.1)_\lambda$. Now if y satisfies (vi) then

$$|y'(t)| = \left| \int_0^t y''(z)\, dz \right| \leqq M_1,$$

while if y satisfies (vii)

$$|y'(t)| = \left| \int_t^1 y''(z)\, dz \right| \leqq M_1.$$

Finally if y satisfies (viii)

$$|y'(t)| \leqq \left| \int_0^t y''(z)\, dz \right| + |y'(0)| \leqq M_1 + \frac{\alpha M}{\beta} \equiv M_2.$$

Thus $|y'| \leqq M_2$ for each solution y to $(3.1)_\lambda$. With these bounds the differential equation yields a priori bounds independent of λ for $|y'''|$ i.e. $|y'''| \leqq \max\{|f(t, u, p, q)|\} \equiv M_3$ where the maximum is computed over $[0, 1] \times [-M, M] \times [-M_2, M_2] \times [-M_1, M_1]$. Thus $|y|_3 \leqq K = \max\{M, M_1, M_2, M_3\}$ and the existence of a solution to (3.1) is established.

For notational purposes, let $(3.1)_{(\mathrm{vi})}$ denote the boundary value problem $y''' = f(t, y, y', y'')$, $t \in [0, 1]$, with y satisfying (vi). Similarly, define $(3.1)_{(\mathrm{vii})}$ and $(3.1)_{(\mathrm{viii})}$. Next sufficient conditions on f are given which imply a priori bounds on any solution $y(t)$ to $(3.1)_{(\mathrm{vi})}$, $(3.1)_{(\mathrm{vii})}$ or $(3.1)_{(\mathrm{viii})}$. Suppose $y(t)$ is a solution to (3.1) and $[y(t)]^2$ has a maximum at $t_0 \in (0, 1)$. Then $y'(t_0) = 0$ and $y(t_0)y''(t_0) \leqq 0$.

THEOREM 3.2. *Suppose there is a constant $M \geqq 0$ such that*

$$\int_0^t u'(z)[f(z, u(z), u'(z), u''(z)) + L(u'(z))^n u''(z)]\, dz > 0$$

for $|u(t)| > M$, where L and $n > -2$ are constants, with $u \in C^2[0, 1]$ and $u'(0) = 0$.

Then any solution y to $(3.1)_{(\mathrm{vi})}$ or $(3.1)_{(\mathrm{vii})}$ satisfies

$$|y(t)| \leqq M \quad \textit{for } t \in [0, 1].$$

Proof. Suppose $|y|$ achieves a positive maximum at $t_0 \in (0, 1)$, then $y'(t_0) = 0$. Assume $|y(t_0)| > M$, and so

$$\int_0^{t_0} [y'(z)y'''(z) + L(y'(z))^{n+1} y''(z)]\, dz > 0.$$

Integration by parts together with $y'(t_0) = 0$ yields

$$-\int_0^{t_0} (y''(z))^2\, dz > 0,$$

a contradiction. Thus, $|y(t_0)| \leqq M$.

At this stage we divide the proof into two cases. Suppose first y is a solution to $(3.1)_{(\mathrm{vi})}$. If $|y|$ assumes its maximum value at either $t=0$ or $t=1$ then trivially $|y(t)| \leqq M$ for $t \in [0,1]$. So the conclusion of the theorem follows for $(3.1)_{(\mathrm{vi})}$. Now suppose y is a solution to $(3.1)_{(\mathrm{vii})}$. If y assumes its maximum value at $t=0$ then trivially $|y(t)| \leqq M$ for $t \in [0,1]$. On the other hand, suppose $|y|$ achieves its maximum value at $t=1$. Suppose $|y(1)|>M$; if so,

$$\int_0^1 [y'(z)y'''(z)+L(y'(z))^{n+1}y''(z)]\,dz>0,$$

which yields

$$-\int_0^1 [y''(z)]^2\,dz>0,$$

a contradiction. Thus $|y(1)| \leqq M$ and the conclusion of the theorem follows for $(3.1)_{(\mathrm{vii})}$.

Remark. M is independent of L in Theorem 3.2.

An analogous theorem holds for $(3.1)_{(\mathrm{viii})}$.

THEOREM 3.3. *Suppose there is a constant $M \geqq 0$ such that*

$$\int_0^t u'(z)[f(z,u(z),u'(z),u''(z))+L(u'(z))^n u''(z)]\,dz>0$$

for $|u(t)|>M$, where n is an even integer greater than or equal to zero and $L \geqq 0$ is a constant, with $u \in C^2[0,1]$ and $u''(0)=0$.

Then any solution y to $(3.1)_{(\mathrm{viii})}$ satisfies

$$|y(t)| \leqq M \quad \text{for } t \in [0,1].$$

Proof. Suppose $|y|$ achieves a positive maximum at $t_0 \in (0,1)$ and assume $|y(t_0)|>M$. Then $y'(t_0)=0$ and

$$\int_0^{t_0} [y'(z)y'''(z)+L(y'(z))^{n+1}y''(z)]\,dz>0$$

yield

$$-\int_0^{t_0} [y''(z)]^2\,dz - L\frac{(y'(0))^{n+2}}{n+2}>0,$$

a contradiction. Thus $|y(t_0)| \leqq M$.

On the other hand $|y|$ cannot have a nontrivial maximum at 0 or 1. For suppose the maximum of $|y|$ occurs at 0. Then $y(0)y'(0) \leqq 0$. However, from (viii), $y(0)y'(0)=(\beta/\alpha)[y(0)]^2>0$, a contradiction. A similar argument works for the case $t=1$. Thus

$$|y(t)| \leqq M \quad \text{for } t \in [0,1].$$

We are now in a position to prove our main existence theorems for this section.

THEOREM 3.4. *Let $f:[0,1]\times R^3 \to R$ be continuous.*

(a) *Suppose there is a constant $M \geqq 0$ such that*

$$\int_0^t u'(z)[f(z,u(z),u'(z),u''(z))+L(u'(z))^n u''(z)]\,dz>0$$

for $|u(t)|>M$, where L and $n>-2$ are constants, with $u \in C^2[0,1]$ and $u'(0)=0$.

(b) *Suppose*

$$|f(t,u,p,q)| \leqq [A(t,u)|p|+B(t,u)][C(t,u)|q|+D(t,u)]$$

where $A(t,u)$, $B(t,u)$, $C(t,u)$, $D(t,u) \geqq 0$ are functions bounded on bounded (t,u) sets.

Then the boundary value problem $(3.1)_{(\mathrm{vi})}$ *and* $(3.1)_{(\mathrm{vii})}$ *have at least one solution in* $C^3[0,1]$.

Proof. To prove existence of a solution in $C^3[0,1]$ we apply Theorem 3.1. We need to establish a priori bounds for any solution $y(t)$ to $(3.1)_\lambda$. Now if $\lambda=0$ we have the unique solution $y\equiv 0$. Otherwise for $0<\lambda\leqq 1$

$$\int_0^t u'(z)[f(z,u(z),u'(z),u''(z))+L(u'(z))^n u''(z)]\,dz>0,$$

for $|u(t)|>M$ implies

$$\int_0^t u'(z)[\lambda f(z,u(z),u'(z),u''(z))+\lambda L(u'(z))^n u''(z)]\,dz>0$$

for $|u(t)|>M$. So Theorem 3.2 together with its remark implies $|y|\leqq M$ for any solution y to $(3.1)_\lambda$. Hence the existence of a solution to $(3.1)_{(\mathrm{vi})}$ and $(3.1)_{(\mathrm{vii})}$ is established.

COROLLARY 3.5. *Let* $f:[0,1]\times R^3\to R$ *be continuous.*

(a) *Suppose there is a constant* $M\geqq 0$ *such that*

$$\int_0^t u'(z)[f(z,u(z),u'(z),u''(z))+L[u'(z)]^n u''(z)]\,dz\geqq 0$$

for $|u(t)|>M$, *where L and* $n>-2$ *are constants, with* $u\in C^2[0,1]$ *and* $u'(0)=0$.

(b) *Suppose*

$$|f(t,u,p,q)|\leqq[A(t,u)|p|+B(t,u)][C(t,u)|q|+D(t,u)]$$

where $A(t,u)$, $B(t,u)$, $C(t,u)$, $D(t,u)\geqq 0$ *are functions bounded on bounded* (t,u) *sets.*

Then the boundary value problem $(3.1)_{(\mathrm{vi})}$ *and* $(3.1)_{(\mathrm{vii})}$ *have at least one solution in* $C^3[0,1]$.

Proof. Let us consider

$$(3.3)\qquad \begin{gathered} y'''=f_n(t,y,y',y''),\\ y\in(\mathrm{vi})\text{ or }(\mathrm{vii})\end{gathered}$$

where $f_n(t,y,y',y'')=f(t,y,y',y'')+(y/n)$ for $n=1,2,\cdots$. Clearly

$$\begin{aligned}&\int_0^t y'(z)[f_n(z,y(z),y'(z),y''(z))+L(y'(z))^n y''(z)]\,dz\\ &=\int_0^t y'(z)[f(z,y(z),y'(z),y''(z))+L(y'(z))^n y''(z)]\,dz+\frac{1}{n}\frac{y^2(t)}{2}\\ &>0\end{aligned}$$

for $|y(t)|>M$ since $y(0)=0$.

Thus Theorem 3.2 implies $|y_n|\leqq M$ for any solution y_n to (3.3) and $n=1,2,\cdots$. Also

$$|f_n(t,y,y',y'')|\leqq[A(t,y)|y'|+B(t,y)][C(t,y)|y''|+D(t,y)]+M.$$

Now we can apply Theorem 3.4 to (3.3): If y_n is a solution to (3.3) for $n=1,2,\cdots$ we have $|y_n|_3\leqq K$ for some constant K independent of n. By the Arzela-Ascoli Theorem there is a subset N of the natural numbers and a function $y\in C^2[0,1]$ so that $|y_n-y|_2\to 0$ as $n\to\infty$ in N. If $G(t,z)$ is the Green's function for (L,B_0) where $Ly=y'''$ and B_0 denotes the boundary conditions (vi) or (vii) then

$$y_n(t)=\int_0^1 G(t,z)f_n(z,y_n(z),y_n'(z),y_n''(z))\,dz.$$

Let $n \to \infty$ through N to obtain

$$y(t) = \int_0^1 G(t, z) f(z, y(z), y'(z), y''(z))\, dz.$$

Thus $y \in C^3_{B_0}$ and y satisfies $y''' = f(t, y, y', y'')$.

We can obtain a similar result to Theorem 3.4 for $(3.1)_{(\mathrm{viii})}$.

THEOREM 3.6. *Let $f: [0, 1] \times R^3 \to R$ be continuous.*

(a) *Suppose there is a constant $M \geqq 0$ such that*

$$\int_0^t u'(z)[f(z, u(z), u'(z), u''(z)) + L(u'(z))^n u''(z)]\, dz > 0$$

for $|u(t)| > M$, where n is an even integer greater than or equal to zero and $L \geqq 0$ is a constant, with $u \in C^2[0, 1]$ and $u''(0) = 0$.

(b) *Suppose*

$$|f(t, u, p, q)| \leqq [A(t, u)|p| + B(t, u)][C(t, u)|q| + D(t, u)]$$

where $A(t, u)$, $B(t, u)$, $C(t, u)$, $D(t, u) \geqq 0$ are functions bounded on bounded (t, u) sets.

Then the boundary value problem $(3.1)_{(\mathrm{viii})}$ has at least one solution in $C^3[0, 1]$.

Remark. We can obtain similar results to those in Theorems 3.4 and 3.6 if the boundary conditions (vi), (vii) or (viii) are replaced by any of the following:

$$\begin{array}{ll}
(\mathrm{ix}) & y(0) = 0, \quad y(1) = 0, \quad y''(0) = 0, \\
(\mathrm{x}) & y(0) = 0, \quad y'(1) = 0, \quad y''(0) = 0, \\
(\mathrm{xi}) & y(1) = 0, \quad y'(0) = 0, \quad y'(0) = 0, \\
(\mathrm{xii}) & y(0) = 0, \quad y(1) = 0, \quad y'(1) = 0, \\
(\mathrm{xiii}) & y(0) = 0, \quad y(1) = 0, \quad y''(1) = 0, \\
(\mathrm{xiv}) & y(1) = 0, \quad y'(0) = 0, \quad y''(1) = 0,
\end{array}$$

or

$$\begin{array}{lll}
(\mathrm{xv}) & -\alpha y(0) + \beta y'(0) = 0, & \alpha, \beta > 0, \\
 & ay(1) + by'(1) = 0, & a, b > 0, \\
 & y''(1) = 0. &
\end{array}$$

An example of this is the following theorem.

THEOREM 3.7. *Let $f: [0, 1] \times R^3 \to R$ be continuous.*

(a) *Suppose there is a constant $M \geqq 0$ such that*

$$\int_t^1 u'(z)[f(z, u(z), u'(z), u''(z)) + L(u'(z))^n u''(z)]\, dz > 0$$

for $|u(t)| > M$, where n is an even integer greater than or equal to zero and $L \leqq 0$ is a constant, with $u \in C^2[0, 1]$ and $u''(1) = 0$.

(b) *Suppose*

$$|f(t, u, p, q)| \leqq [A(t, u)|p| + B(t, u)][C(t, u)|q| + D(t, u)]$$

where $A(t, u)$, $B(t, u)$, $C(t, u)$, $D(t, u) \geqq 0$ are functions bounded on bounded (t, u) sets.

Then the boundary value problem

$$y''' = f(t, y, y', y''), \qquad t \in [0, 1], \quad y \in (xv)$$

has at least one solution in $C^3[0, 1]$.

The following example illustrates the ideas and results of this section.

Example 2. Consider the boundary value problem

$$\text{(BVP)}\quad \begin{aligned} &y'''(t)=A_0+B_0[y(t)]^p+C_0[y(t)]^m y'(t)+D_0y''(t)+E_0y'(t)y''(t), \qquad t\in[0,1],\\ &y(0)=y(1)=y'(0)=0 \end{aligned}$$

where A_0, $B_0>0$, $C_0\geqq 0$, D_0, E_0 are constants with $m\geqq 0$ even and $p>0$ odd.

We will now show that (BVP) has a solution in $C^3[0,1]$ via Theorem 3.4. Now if $f(t,y,y',y'')=A_0+B_0y^p+C_0y^my'+D_0y''+E_0y'y''$ and $L=-E_0$, $n=1$, $\tilde{L}=-D_0$, $\tilde{n}=0$ we have

$$\begin{aligned} &\int_0^t y'(z)[f(z,y(z),y',y'')+L(y'(z))^n y''(z)+\tilde{L}(y'(z))^{\tilde{n}}y''(z)]\,dz\\ &\qquad=\int_0^t (A_0y'(z)+B_0[y(z)]^p y'(z)+C_0[y(z)]^m[y'(z)]^2)\,dz\\ &\qquad\geqq\int_0^t [A_0y'(z)+B_0(y(z))^p y'(z)]\,dz\\ &\qquad=y(t)\left(A_0+\frac{B_0}{p+1}[y(t)]^p\right)\quad \text{since } y(0)=0\\ &\qquad>0\quad \text{for } |y(t)|>\left[\left|\frac{-(p+1)A_0}{B_0}\right|\right]^{1/p}. \end{aligned}$$

Finally it is clear that we can find constants A, B, C, D such that

$$|f(t,y,y',y'')|\leqq(A|y'|+B)(C|y''|+D)$$

for (t,y) in a bounded set.

Hence all conditions in Theorem 3.4 are satisfied, so (BVP) has at least one solution in $C^3[0,1]$. □

To conclude this section we examine the inhomogeneous boundary value problem

$$\text{(3.4)}\quad \begin{aligned} &y'''=f(t,y,y',y''), \qquad t\in[0,1],\\ &y(0)=r,\quad y(1)=s,\quad y'(0)=l, \end{aligned}$$

or

$$\text{(3.5)}\quad \begin{aligned} &y'''=f(t,y,y',y''), \qquad t\in[0,1],\\ &y(0)=r,\quad y(1)=s,\quad y''(0)=l, \end{aligned}$$

and establish the existence of a solution to (3.4), (3.5) in $C^3[0,1]$ under essentially the same hypothesis on f used in Theorem 3.4.

THEOREM 3.8. *Let $f:[0,1]\times R^3\to R$ be continuous.*

(a) *Suppose there is a constant $M\geqq 0$ such that*

$$\int_0^t u'(z)[f(z,u(z)+lz,u'(z)+l,u''(z))+L(u'(z))^n u''(z)]\,dz>0$$

for $|u(t)|>M$, where L and $n>-2$ are constants with $u\in C^2[0,1]$ and $u'(0)=0$.

(b) *Suppose*

$$|f(t,u,p,q)|\leqq[A(t,u)|p|+B(t,u)][C(t,u)|q|+D(t,u)]$$

where $A(t,u)$, $B(t,u)$, $C(t,u)$, $D(t,u)\geqq 0$ are functions bounded on bounded (t,u) sets.

Then the boundary value problem (3.4) *has at least one solution in* $C^3[0, 1]$.

Proof. Consider the family of problems

$$(3.4)_\lambda \qquad \begin{aligned} &y''' = \lambda f(t, y, y', y''), \qquad 0 \leqq \lambda \leqq 1, \\ &y(0) = r, \quad y(1) = s, \quad y'(0) = l. \end{aligned}$$

The existence of a solution in $C^3[0, 1]$ follows immediately from Theorem 5.1 of [7] once a priori bounds independent of λ are established for solutions y to $(3.4)_\lambda$. To establish a priori bounds for $(3.4)_\lambda$, let $y(t)$ be a solution to $(3.4)_\lambda$. Now if $\lambda = 0$ we have a unique solution and thus $|y(t)| \leqq M_0$, for some constant M_0. Otherwise for $0 < \lambda \leqq 1$, let $w = y - lt$, so $w(0) = r$, $w(1) = s - l$ and $w'(0) = 0$. Now

$$\int_0^t w'(z)[f(z, w(z) + lz, w'(z) + l, w''(z)) + L(w'(z))^n w''(z)]\, dz > 0$$

for $|w(t)| > M$ implies

$$\int_0^t w'(z)[\lambda f(z, w(z) + lz, w'(z) + l, w''(z)) + \lambda L(w'(z))^n w''(z)]\, dz > 0$$

for $|w(t)| > M$. It follows from Theorem 3.2 and its remark that

$$|w(t)| \leqq \max\{M, |r|, |s - l|\} = \tilde{K} \quad \text{for } t \in [0, 1].$$

Thus $|y(t)| \leqq M_1 \equiv \max\{\tilde{K} + |l|, M_0\}$ for any solution y to $(3.4)_\lambda$, $0 \leqq \lambda \leqq 1$. A priori bounds independent of λ for y', y'', y''' will follow from a slight modification of the proof in Theorem 3.1 once we observe that $|y''(\mu)| \leqq K$, $K \geqq 0$ a fixed constant independent of λ, for some $\mu \in [0, 1]$. We accomplish this by letting $v(t) = y(t) - (1 - t^2)r - t^2 s + lt^2$ and noticing that $v(0) = 0$, $v(1) = l$, $v'(0) = l$. Hence by the Mean Value Theorem there exists $\mu \in (0, 1)$ such that $v''(\mu) = 0$, i.e. $y''(\mu) = 2s - 2l - 2r$. The existence of a solution to (3.4) follows from Theorem 5.1 of [7].

We can obtain a corresponding existence theorem for (3.5).

THEOREM 3.9. *Let* $f: [0, 1] \times R^3 \to R$ *be continuous.*

(a) *Suppose there is a constant* $M \geqq 0$ *such that*

$$\int_0^t u'(z)[f(z, u(z) + \frac{lz^2}{z}, u'(z) + lz, u''(z) + l)) + L(u'(z))^n u''(z)]\, dz > 0$$

for $|u(t)| > M$, *where n is an even integer greater than or equal to zero and* $L \geqq 0$ *is a constant, with* $u \in C^2[0, 1]$ *and* $u''(0) = 0$.

(b) *Suppose*

$$|f(t, u, p, q)| \leqq [A(t, u)|p| + B(t, u)][C(t, u)|q| + D(t, u)]$$

where $A(t, u)$, $B(t, u)$, $C(t, u)$, $D(t, u) \geqq 0$ *are functions bounded on bounded* (t, u) *sets.*

Then the boundary value problem (3.5) *has at least one solution in* $C^3[0, 1]$.

REFERENCES

[1] P. B. BAILEY, L. F. SHAMPINE AND P. E. WALTMAN, *Nonlinear Two Point Boundary Value Problems*, Academic Press, New York, London, 1968.

[2] J. DUGUNDJI AND A. GRANAS, *Fixed Point Theory, Vol.* 1, Monographie Matematyczne, PNW, Warsaw, 1982.

[3] A. GRANAS, *Sur la méthode de continuité de Poincaré*, C.R. Acad. Sci. Paris Sér. I. Math., 282 (1976), pp. 983-985.

[4] A. Granas, R. B. Guenther and J. W. Lee, *On a theorem of S. Bernstein*, Pacific J. Math., 74 (1978), pp. 67-82.

[5] ———, *Nonlinear boundary value problems for some classes of ordinary differential equations*, Rocky Mountain J. Math., 10 (1980), pp. 35-38.

[6] ———, *Topological transversality* I (*Some nonlinear diffusion problems*), Pacific J. Math., 89 (1980), pp. 53-67.

[7] R. B. Guenther, *Problèmes aux limites non linéaires pour certaines classes d'équations différentielles ordinaires*, Les Presses de L'Université de Montréal, 1985.

[8] J. E. Innes and L. K. Jackson, *Nagumo conditions for ordinary differential equations*, International Conference on Differential Equations, H. A. Artosiewicz, ed., Academic Press, New York, 1975, pp. 385-398.

[9] D. Krajcinovic, *Sandwich beam analysis*, Trans. ASME Ser. E, J. Appl. Mech., 94 (1972), pp. 773-778.

[10] T. Y. Na, *Computational Methods in Engineering Boundary Value Problems*, Academic Press, New York, 1979.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

STRICTLY COOPERATIVE SYSTEMS WITH A FIRST INTEGRAL*

JANUSZ MIERCZYŃSKI†

Abstract. We consider systems of differential equations $dx_i/dt = F_i(x_1, \cdots, x_n)$ in the nonnegative orthant in the n-space satisfying the following hypotheses: i) $F(0) = 0$; ii) if $x_i < y_i$ and $x_j = y_j$ for $j \neq i$ then $F_k(x) < F_k(y)$ for $k \neq i$; iii) F possesses a first integral with positive gradient. We prove that every solution to such a system either converges to an equilibrium or eventually leaves any compact set.

Key words. equilibrium, first integral, Lyapunov function, ω-limit set, nonnegative orthant, order preserving, strictly cooperative system

AMS(MOS) subject classification. Primary 34C05

1. Introduction. The purpose of the present paper is to study the limiting behavior of solutions of systems of ordinary differential equations possessing a first integral, where the right sides of equations as well as the first integrals are subject to some monotonicity conditions. We prove that any solution to such a system either converges to an equilibrium or eventually leaves any compact set.

Particular classes of systems of differential equations $\dot{x} = F(x)$, $x \in U \subset \mathbb{R}^n$, satisfying conditions $\partial F_i/\partial x_j \geqq 0$ for $i \neq j$, were studied by many authors (see references in [1]; see also [3, Chap. III]). Recently, in [1] and [2], M. W. Hirsch initiated investigation of systems of that type (which he calls cooperative systems), using ideas taken from the dynamical systems theory. Such systems may describe, for instance, competition between biological species or chemical reactions. In cooperative systems it is natural to expect convergence of bounded solutions to an equilibrium or to a closed orbit. A decisive step toward answering this conjecture was made by M. W. Hirsch, who in [2] proved that, for systems that are cooperative and irreducible (that is, the matrices $[(\partial F_i/\partial x_j)(p)]$ are irreducible), almost all (with respect to the Lebesgue measure) points whose forward orbits are bounded approach the equilibrium set. In [11] J. Smillie has found a class of cooperative irreducible systems for which the following holds: every solution either converges to an equilibrium or eventually leaves any compact set.

To our knowledge, a general class of first integrals for cooperative systems was considered only in [2, Thm. 4.7]. However, that result is negative: if the set of equilibria is countable then every continuous invariant function is constant. On the other side, many authors (e.g. [5], [6], [7], [9], [10]) considered cooperative (or related) systems on the nonnegative orthant $\mathbb{R}^n_+$ having $\sum_i x_i$ as a first integral. For such systems it was proved that every solution converges to an equilibrium. In [9] and [10] these results were extended to the case of nonautonomous cooperative systems periodic (resp. almost periodic) in t.

The results contained in the present paper are a generalization of the theorems mentioned above to the case of not necessarily linear first integral. Methods used here are geometric, and the only nonelementary tool made use of is the Brouwer fixed-point theorem. The exposition is independent of any other work on this subject; however, the idea of a Lyapunov function L is taken from the author's previous work [8].

2. Definitions and preliminary lemmas. We define $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geqq 0\}$, $\partial\mathbb{R}^n_+ = \{x \in \mathbb{R}^n_+ : x_i = 0 \text{ for some } i\}$, $\operatorname{Int}\mathbb{R}^n_+ = \mathbb{R}^n_+ \backslash \partial\mathbb{R}^n_+$. Moreover, we denote $e_i = (0, \cdots, 0, {}_i1, 0, \cdots, 0)$—the ith vector of the standard base in $\mathbb{R}^n$, $B = \{x \in \mathbb{R}^n : 0 \leqq x_i \leqq 1$

* Received by the editors February 25, 1985; accepted for publication (in revised form) January 13, 1986.

† Institute of Mathematics, Technical University of Wrocław, PL-50-370 Wrocław, Wybrzeże Wyspiańskiego 27, Poland.

for each $i\}$, $B_+ = \{x \in B \colon x_i = 1$ for some $i\}$, $x < y$ if $x_i < y_i$ for each i, and $x <_i y$ if $x_i < y_i$ and $x_j = y_j$ for $j \neq i$. $\langle \cdot, \cdot \rangle$ will denote the usual inner product in $\mathbb{R}^n$, $\|\cdot\|$—the corresponding norm.

Let $H \colon \mathbb{R}^n_+ \to \mathbb{R}$ be a C^1 function. By a gradient of H at p we mean the vector $((\partial H/\partial x_1)(p), \cdots, (\partial H/\partial x_n)(p))$ (written grad $H(p)$).

Let $F \colon \mathbb{R}^n_+ \to \mathbb{R}^n$ be a vector field. By a *first integral* for F we understand a function $H \colon \mathbb{R}^n_+ \to \mathbb{R}_+$, of class C^1, such that $\operatorname{grad} H(p) \neq 0$ at each $p \in \mathbb{R}^n_+$ and $\langle \operatorname{grad} H(p), F(p) \rangle = 0$.

A point $p \in \mathbb{R}^n_+$ for which $F(p) = 0$ is called an *equilibrium.*

Let $x \colon [0, T) \to \mathbb{R}^n_+$ be a nonextendible solution to the initial value problem $dx/dt = F(x)$, $x(0) = x$. We say the set $\{x(t) \colon 0 \leqq t < T\}$ is a *forward orbit* of x.

The set $\omega(x)$ consists of those points $y \in \mathbb{R}^n_+$ for which there exists a sequence $t_k \to T$ such that $x(t_k) \to y$. This set is called an *ω-limit set* of x. The following facts are well known.

THEOREM 2.1. *If $\omega(x) = \{y\}$ then the solution x is defined on $[0, +\infty)$ and converges to an equilibrium y.*

THEOREM 2.2. *If $\omega(x) = \varnothing$ then $x(\cdot)$ eventually leaves any compact set.*

The set $A \subset \mathbb{R}^n_+$ is called *positively invariant* if for each $a \in A$ its forward orbit is contained in A.

We consider a system of ordinary differential equations in $\mathbb{R}^n_+$ defined by a C^1 vector field $F \colon \mathbb{R}^n_+ \to \mathbb{R}^n$,

$$(2.1) \qquad \frac{dx_i}{dt} = F_i(x_1, \cdots, x_n) = F_i(x), \quad x \in \mathbb{R}^n_+, \qquad F = (F_1, \cdots, F_n),$$

satisfying:

(A1) $F(0) = 0$;

(A2) If $x <_i y$ then $F_j(x) < F_j(y)$ for $j \neq i$;

(A3) There exists a first integral H for F such that $\operatorname{grad} H(x) > 0$ for $x \in \mathbb{R}^n_+$ and (for convenience) $H(0) = 0$. Systems satisfying (A2) will be called *strictly cooperative.*

Let M denote the least upper bound of the values of H. From (A3) it follows that H is onto $[0, M)$, where $0 < M \leqq +\infty$.

By $\operatorname{Int} H^{-1}(h)$ we denote the set $\{x \in \operatorname{Int} \mathbb{R}^n_+ \colon H(x) = h\}$.

LEMMA 2.1. *Let c be an equilibrium. Then $c + \mathbb{R}^n_+$ is positively invariant. Moreover, c is a unique equilibrium on $c + \partial\mathbb{R}^n_+$.*

Proof. By performing, if necessary, the change of coordinates $\tilde{x}_i = x_i - c_i$, we reduce a general case to the case $c = 0$. Let $x \in \mathbb{R}^n_+$. Then let I denote the subset of $\{1, \cdots, n\}$ such that $x_i = 0$ exactly for $i \in I$. If $x \neq 0$ then $x_j > 0$ for indices j belonging to some nonvoid subset J of $\{1, \cdots, n\}$. For convenience assume $I = \{1, \cdots, k\}$, $J = \{k+1, \cdots, n\}$. By (A2) we obtain

$$F_i(0, \cdots, 0, x_{k+1}, \cdots, x_n) > F_i(0, \cdots, 0, x_{k+2}, \cdots, x_n) > \cdots > F_i(0) = 0 \quad \text{for } i \in I.$$

Therefore on the boundary of $\mathbb{R}^n_+$ (except 0) the vector field F is directed inward, which in a standard way implies that $\mathbb{R}^n_+$ is positively invariant. Q.E.D.

LEMMA 2.2. *Let c be an equilibrium. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that for each $h \in [H(c) - \delta, H(c) + \delta] \cap [0, M)$ there exists an equilibrium c_h such that $H(c_h) = h$, $c_h > c$ for $h > H(c)$ (resp. $c_h < c$ for $h < H(c)$) and $\|c_h - c\| \leqq \varepsilon$.*

Proof. We consider the case $h > H(c)$. From (A3) it follows that $H(c + n^{-1/2}e_i) > H(c)$. Set $\delta = \min_i H(c + n^{-1/2}e_i) - H(c)$. For $x \in c + n^{-1/2}B_+$ we have $H(x) \geqq H(c) + \delta$. Let for a moment $h \in (H(c), H(c) + \delta)$ be fixed. By $\ell(x)$ we denote the straight line passing through c and x. Let the mapping $K \colon (c + n^{-1/2}B_+) \to H^{-1}(h) \cap (c + \mathbb{R}^n_+)$ be defined in the following way: $K(x)$ is the unique point on $\ell(x)$ at which $H(K(x))=h$. The existence and uniqueness of this point follow from (A3). Having in mind our choice of h, we infer that K is in fact onto $H^{-1}(h)\cap(c+n^{-1/2}B)=H^{-1}(h)\cap(c+\mathbb{R}_+^n)$. The definition of K implies that K is one-to-one. Moreover, K^{-1} is continuous as a "central projection" onto $(c+n^{-1/2}B_+)$. Therefore, from the compactness of the domain of K, K is a homeomorphism. Thus $H^{-1}(h)\cap(c+\mathbb{R}_+^n)$ is homeomorphic to the $(n-1)$-dimensional disk. This set, as an intersection of positively invariant sets, is positively invariant. The Brouwer fixed-point theorem tells us that there is an equilibrium $c_h\in H^{-1}(h)\cap(c+\mathbb{R}_+^n)$. From Lemma 2.1 it follows that $(c_h)_i>c_i$. Moreover, for every $y\in c+n^{-1/2}B$ we have $\|y-c\|\leqq\varepsilon$, so $\|c_h-c\|\leqq\varepsilon$. The case $h<H(c)$ is treated in an analogous way. Q.E.D.

PROPOSITION 2.1. *The set S of equilibria is linearly ordered by* $<$.

Proof. Suppose there exist $c, d\in S$ such that $c_i=d_i$ for $i\in I$, $c_j<d_j$ for $j\in J$, $c_l>d_l$ for $l\in L$ and at least two of these sets are nonempty. Consider the point z, $z_k=\max\{c_k, d_k\}$ for $1\leqq k\leqq n$. Proceeding as in the proof of Lemma 2.1 we obtain $F_i(z)>F_i(d)=0$ for $i\in I$, $F_j(z)>F_j(d)=0$ for $j\in J$ and $F_l(z)>F_l(c)=0$ for $l\in L$. But from this it follows that $\langle F(z), \operatorname{grad} H(z)\rangle>0$, a contradiction. Q.E.D.

COROLLARY. *For each $h\in[0, M)$ there is at most one equilibrium on $H^{-1}(h)$.*

PROPOSITION 2.2. *There exists U, $0<U\leqq M$, such that there is an order-preserving homeomorphism between $[0, U)$ and the set S of equilibria of F.*

Proof. The function $H|S$ is continuous. From the corollary and Proposition 2.1 it follows that $H|S$ is one-to-one, so Z, the function inverse to $H|S$, exists. Lemma 2.2 implies that Z is continuous and preserves order. Therefore Z is a homeomorphism. The statement on the domain of Z also follows from Lemma 2.2. Q.E.D.

Let Z denote, as in the above proof, the function inverse to $H|S$. We define the function $L\colon \mathbb{R}_+^n\to\mathbb{R}_+$, $L(x)=\min\{Z_i^{-1}(x_i)\colon Z_i^{-1}(x_i) \text{ is defined}\}$, where Z_i^{-1} denotes the function inverse to the ith coordinate of Z.

LEMMA 2.3. *L is well defined and continuous.*

Proof. Suppose that for each i, $Z_i(h)\to a_i$ as $h\to U$. Then $a=(a_1,\cdots,a_n)$ would be an equilibrium, and $H(a)=U$. But by Lemma 2.2 there would exist an equilibrium $b>a$ and $H(b)>U$. Thus for some i, $Z_i(h)\to\infty$ as $h\to U$, so L is well defined.

In order to prove the continuity of L notice that, by Lemma 2.2, the image of Z_i is a right-open interval. Therefore if in the definition of $L(x)$ the minimum is realized exactly by indices $i\in I$ then for y belonging to some sufficiently small neighborhood of x the minimum is realized by indices from some subset of I. The minimum of a finite family of continuous functions is continuous, so we have obtained the desired result. Q.E.D.

Remark. As we have said in the Introduction, L has a simple geometric interpretation. Namely, the level surface corresponding to h is equal to the set $Z(h)+\partial\mathbb{R}_+^n$.

LEMMA 2.4. *For $h\in[0, U)$, $\sup\{L(x)\colon x\in H^{-1}(h)\}$ is attained only at $Z(h)$. For $h\in[U, M)$, $\sup\{L(x)\colon x\in H^{-1}(h)\}$ is attained nowhere.*

Proof. First assume $h\in[0, U)$. It is easy to check out that $L(Z(h))=h$. Suppose that for some $x\neq Z(h)$, $x\in H^{-1}(h)$, we have $L(x)\geqq h$. From the definition of L it follows that $x_i\geqq Z_i(h)$ for every i. But for some j, $x_j\neq Z_j(h)$. (A3) implies that $H(x)>h$, a contradiction.

Now we assume $h\in[U, M)$. Let x be any point of $H^{-1}(h)$. Denote $L(x)$ by g. This means that $x\in Z(g)+\partial\mathbb{R}_+^n$ (of course $g<h$). Let I stand for the set of indices i for which $x_i=Z_i(g)$. Define the point $\tilde{x}$ in the following way: $\tilde{x}_i=x_i+1$ for $i\in I$, $\tilde{x}_j=x_j$ for $j\notin I$. From (A3) it follows that $H(\tilde{x})>h$. Considering, as in the proof of Lemma 2.2, the straight line passing through $Z(g)$ and x, we can find a point $y\in Z(g)+\operatorname{Int}\mathbb{R}_+^n$

such that $H(y)=h$. But the fact that $y \in Z(g)+\operatorname{Int}\mathbb{R}_+^n$ means that $H(y)>h$, a contradiction. Q.E.D.

3. The main result.

THEOREM 3.1. *For $x \in \mathbb{R}_+^n$ we have either $\omega(x)=\{Z(H(x))\}$ or $\omega(x)=\varnothing$.*

Proof. Let I denote the set of indices realizing the minimum in the definition of $L(x(t))$. Analogously, as in the proof of Lemma 2.1 we can show that $\dot{x}_i(t)>0$ for $i \in I$. Hence L is strictly increasing along the forward orbits of (2.1), except for the constant solutions. Let $x \in \mathbb{R}_+^n$ be fixed, and let E denote the set $H^{-1}(H(x))$. H is a first integral, so the forward orbit of x is contained in E. Furthermore, from the closedness of E it follows that $\omega(x) \subset E$. Now it suffices to show that $\omega(x) \subset \{y \in E: L(y)=\sup_{z\in E} L(z)\}$. Suppose not. Then there exists $y \in \omega(x)$ such that $L(y)<\sup_{z\in E} L(z)$. From Lemma 2.4, y is not an equilibrium, so L strictly increases along the orbit of y. Let v be any point·on the forward orbit of y, distinct from y. Obviously $L(v)>L(y)$. Choose neighborhoods V of v, Y of y, such that $\inf_{z\in V} L(z)>\sup_{z\in Y} L(z)$. From the very definition of the ω-limit set we can find moments $t_1<t_2$ such that $x(t_1)\in V$, $x(t_2)\in Y$. From this we deduce that $L(x(t_1))<L(x(t_2))$, which contradicts our choice of V and Y. Q.E.D.

The following example shows that the case $\omega(x)=\varnothing$ is possible.

Example. Consider the nonnegative quadrant $\mathbb{R}_+^2$. As a first integral choose a function $H(x_1,x_2)=G_1(x_1)+G_2(x_2)$, where $G_2(x_2)=x_2$, and $G_1: \mathbb{R}_+ \to \mathbb{R}_+$ is a function of class C^2 such that $G_1(0)=0$, G_1' is positive and $\lim_{x_1\to\infty} G_1(x_1)=1$. Let $W: \mathbb{R}_+ \to \mathbb{R}_+$ be a function of class C^1 such that $W(0)=0$, W' is positive and $\lim_{x\to\infty} W(x)=1$. As the equilibrium set S we choose the graph of W, i.e., $S=\{(x_1,x_2): x_2=W(x_1), x_1 \in \mathbb{R}_+\}$. We define

$$F_2(x_1,x_2)=W(x_1)-x_2,\ F_1(x_1,x_2)=-(F_2(x_1,x_2))/(G_1'(x_1))=(x_2-W(x_1))/(G_1'(x_1)).$$

It is easy to check out that F is C^1 and that H is a first integral for our system. Moreover

$$(\partial F_2/\partial x_1)(x_1,x_2)=W'(x_1)>0, \qquad (\partial F_1/\partial x_2)(x_1,x_2)=[G_1'(x_1)]^{-1}>0,$$

so that the system is strictly cooperative. From the choice of W it follows that $\sup_{x\in S} H(x)=2$. Hence the level surface $H^{-1}(3)$ does not intersect S, so for each $x\in H^{-1}(3)$, $\omega(x)=\varnothing$.

Remark. Professor Morris W. Hirsch has informed the author that he obtained the same result under slightly different hypotheses [4].

Acknowledgments. The author would like to thank Professor M. W. Hirsch for sending him a preprint of the paper [2]. He is also indebted to Professor G. R. Sell for calling his attention to [9] and [10].

REFERENCES

[1] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative.* I: *Limit sets*, this Journal, 13 (1982), pp. 167-179.

[2] ———, *Systems of differential equations that are competitive or cooperative.* II *Convergence almost everywhere*, this Journal, 16 (1985), pp. 432-439.

[3] ———, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc. (N.S.), 11 (1984), pp. 1-64.

[4] ———, private communication.

[5] R. D. JENKS, *Homogeneous multidimensional differential systems for mathematical models*, J. Differential Equations, 4 (1968), pp. 549-565.

[6] ———, *Irreducible tensors and associated homogeneous nonnegative transformations*, J. Differential Equations, 4 (1968), pp. 566-572.

[7] R. D. JENKS, *Quadratic differential systems for interactive population models*, J. Differential Equations, 5 (1969), pp. 497-514.

[8] J. MIERCZYŃSKI, *Functional differential equations with certain monotonicity properties*, Math. Methods Appl. Sci., 6 (1984), pp. 97-103.

[9] F. NAKAJIMA, *Periodic time-dependent gross-substitute systems*, SIAM J. Appl. Math., 36 (1979), pp. 421-427.

[10] G. R. SELL AND F. NAKAJIMA, *Almost periodic gross-substitute dynamical systems*, Tôhoku Math. J., (2), 32 (1980), pp. 255-263.

[11] J. SMILLIE, *Competitive and cooperative tridiagonal systems of differential equations*, this Journal, 15 (1984), pp. 530-534.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

006

ON THE ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF A REACTING-DIFFUSING SYSTEM: A TWO PREDATORS-ONE PREY MODEL*

L. HSIAO†, Y. SU† AND Z. P. XIN‡

Abstract. We discuss a predator-prey system consisting of two predator species and a single prey species which obeys Michaelis-Menten dynamics. We establish global existence for different initial-boundary conditions by either the semigroup approach or the comparison method. Furthermore, we investigate the asymptotic properties of the solutions.

Key words. predator, prey, reaction, diffusion, sectorial operator, fractional power, linearization, spectrum, monotonicity

AMS(MOS) subject classifications. 35, 92

1. Introduction. Evolution equations of reaction-diffusion type enter a wide number of phenomena in ecology, biology, biochemistry, chemistry and physics. An important and interesting problem about these systems is the time evolution of the various density distributions and their relations to the corresponding stationary distributions. This kind of problem has been discussed quite a lot in various fields for a single reaction-diffusion equation. In recent years, attention has been given to coupled reaction-diffusion equations from various fields of applied sciences, but mostly to two equations. For instance, the question as to whether one prey can support two predators is an intriguing one in ecology. But very little theoretical work has been done on such systems, although competition between species exploiting a common prey population is of frequent occurrence in nature.

The present paper discusses a predator-prey system consisting of two predator species, V_1 and V_2, and a single prey species, S. We assume that the predator species compete purely exploitatively, with no interference between rivals. Both species have access to the prey and compete only by lowering the population of shared prey. For death rates it is assumed that the number dying is proportional to the number currently alive. We also assume that there are no significant time lags in the system, that growth rates are logistic in the prey species in the absence of predation, and that the predators' functional response obeys the Holling "nonlearning" curve (for an ecological problem) [5], [6], or Michaelis-Menten Dynamics (for enzyme reactions). The model is given by

$$\begin{aligned}
&\frac{\partial S}{\partial t}=k_1\Delta S+\gamma S\left[1-\frac{S}{K}\right]-\left(\frac{M_1}{y_1}\right)\left(\frac{V_1S}{a_1+S}\right)-\left(\frac{M_2}{y_2}\right)\left(\frac{V_2S}{a_2+S}\right),\\
(1.1)\qquad &\frac{\partial V_1}{\partial t}=k_2\Delta V_1+\frac{M_1V_1S}{a_1+S}-D_1V_1,\\
&\frac{\partial V_2}{\partial t}=k_3\Delta V_2+\frac{M_2V_2S}{a_2+S}-D_2V_2,
\end{aligned}$$

with initial conditions

$$(1.2)\qquad S(X,0)=S_0(X),\quad V_i(X,0)=V_{i0}(X),\quad i=1,2,\quad X\in\Omega$$

* Received by the editors October 1, 1984; accepted for publication (in revised form) January 13, 1986.

† Academia Sinica, Institute of Mathematics, Beijing, People's Republic of China.

‡ Department of Mathematics, Northwest University, Xian, People's Republic of China.

and boundary conditions

$(1.3)_{\mathrm{I}}$ Dirichlet conditions $(S, V_1, V_2)=(0,0,0)$ on $\partial\Omega\times\mathbb{R}^+$, or

$(1.3)_{\mathrm{II}}$ Neumann conditions $\left(\frac{\partial S}{\partial n}, \frac{\partial V_1}{\partial n}, \frac{\partial V_2}{\partial n}\right)=(0,0,0)$ on $\partial\Omega\times\mathbb{R}^+$,

where γ is the growth rate and K the carrying capacity for the prey s, while y_i, M_i, D_i, a_i are the yield constants, the intrinsic growth rate, the death rate and the Michaelis-Menten constant, respectively, for the ith predator V_i, $i=1,2$. k_1, k_2, k_3 are nonnegative diffusion coefficients, respectively, for S, V_1 and V_2. Ω is an open bounded set of R^3 with boundary $\partial\Omega$.

When the effect of diffusion is not considered, namely, $(k_1, k_2, k_3)=(0,0,0)$, this system (1.1) becomes an ordinary differential equation. Hsu, Hubbell and Waltman [7], [8] and Butler and Waltmann [1] analyzed the behavior of solutions of this system of ordinary differential equations in order to answer the biological question: Under what conditions would neither, one, or both species of predator survive? If only one predator survives, how do we determine the limiting behavior of the surviving predator and its prey?

In this paper we analyze the behavior of solutions of system (1.1) with (1.2), $(1.3)_{\mathrm{I}}$ (Case I) or (1.2), $(1.3)_{\mathrm{II}}$ (Case II). In § 2 we deal with Case I. As for the existence and positivity properties of the solutions of (1.1)-$(1.3)_{\mathrm{I}}$ we shall prove Theorems 2.1 and 2.2 by a semigroup approach. Our results on the asymptotic stability of the equilibrium solutions are summarized in Theorems 2.3-2.6. In § 3 we consider Case II. By employing a monotonicity argument we establish existence-comparison Theorems 3.1 and 3.2. We investigate the asymptotic properties of the solution in Theorems 3.3-3.8.

2. The case of Dirichlet conditions. We are interested in the nonnegative solution of problem (1.1), (1.2), $(1.3)_{\mathrm{I}}$. In view of the special form of the reaction terms in (1.1), it is easy to know that [12] (1.1) has the following invariant region Σ:

$$\Sigma=\{(S, V_1, V_2)\colon S\geqq 0,\ V_1\geqq 0,\ V_2\geqq 0\}. \tag{2.1}$$

We change the form of (1.1) into

$$\begin{aligned}\frac{\partial S}{\partial t}&=(k_1\Delta S+\gamma S)+\left(-\frac{\gamma}{K}S^2-\frac{M_1V_1S}{y_1(a_1+S)}-\frac{M_2V_2S}{y_2(a_2+S)}\right),\\ \frac{\partial V_1}{\partial t}&=(k_2\Delta V_1+\beta_1V_1)+\left(-\frac{a_1M_1V_1}{a_1+S}\right),\\ \frac{\partial V_2}{\partial t}&=(k_3\Delta V_2+\beta_2V_2)+\left(-\frac{a_2M_2V_2}{a_2+S}\right),\end{aligned} \tag{2.2}$$

where $\beta_i=M_i-D_i$, $i=1,2$.

It will be convenient to think of system (2.2) as an evolution equation for the unknown $U(t)=(S(t), V_1(t), V_2(t))$ in a Banach space X.

Define $X=(L_2(\Omega))^3$ with norm $\|U\|=(\|S\|_2^2+\|V_1\|_2^2+\|V_2\|_2^2)^{1/2}$ for any $U=(S, V_1, V_2)\in X$, where $\|\cdot\|_2$ denotes the usual norm in $L_2(\Omega)$. Clearly, X is a Banach space.

If one chooses the above Banach space X to work with, system (2.2) can be rewritten as follows.

Let $S(t)=S(\cdot,t)$, $V_1(t)=V_1(\cdot,t)$, $V_2(t)=V_2(\cdot,t)$, $U(t)=(S(t), V_1(t), V_2(t))$. Then (2.2), (1.2), $(1.3)_{\mathrm{I}}$ is changed into the form

$$\frac{dU}{dt}=AU+F(U), \qquad U(0)=U_0 \tag{2.3}$$

where

$$A=\begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{pmatrix}=\begin{pmatrix} k_1\Delta+\gamma & 0 & 0 \\ 0 & k_2\Delta+\beta_1 & 0 \\ 0 & 0 & k_3\Delta+\beta_2 \end{pmatrix}, \tag{2.4}$$

$$D(A_i)=\{u\in L^2(\Omega), u|_{\partial\Omega}=0, A_iu\in L^2(\Omega)\},$$

$$F(U)=\begin{pmatrix} -\dfrac{r}{K}S^2-\dfrac{M_1V_1S}{y_1(a_1+S)}-\dfrac{M_2V_2S}{y_2(a_2+S)} \\ -\dfrac{M_1a_1V_1}{a_1+S} \\ -\dfrac{M_2a_2V_2}{a_2+S} \end{pmatrix}. \tag{2.5}$$

It is well known that $D(A_i)\subseteq H_0^1(\Omega)$ and is dense in $L^2(\Omega)$.

Suppose Y, Z are Banach spaces; we shall denote the norm of a linear bounded operator $T\colon Y\to Z$ by $\|T\|_{Y,Z}$. For a closed operator T, the spectrum, resolvent set and resolvent operator will be denoted by the usual symbols $\sigma(T)$, $\rho(T)$ and $R(\mu,t)=(T-\mu)^{-1}$.

LEMMA 2.1. *A generates an analytic semigroup e^{At} on λ.*

Before we prove the lemma, let us recall the following definition and Theorem [4], which is written down as a proposition.

DEFINITION 2.1. We call a linear operator A in a Banach space X a sectorial operator if it is a closed densely defined operator such that, for some θ in $(0,(\pi/2))$ and some $M\geqq 1$ and real a, the sector

$$S_{a,\theta}=\{\mu|\theta\leqq|\arg(\mu-a)|\leqq\pi, \mu\neq a\}$$

is in the resolvent set of A and

$$\|R(\mu,A)\|\leqq\frac{M}{|\mu-a|} \quad \text{for all } \mu\in S_{a,\theta}.$$

PROPOSITION 2.1. *If A is a sectorial operator, then $-A$ is the infinitesimal generator of an analytic semigroup $\{e^{-tA}\}_{t\geqq 0}$, where*

$$e^{-tA}=\frac{1}{2\pi i}\int_\Gamma(\mu+A)^{-1}e^{\mu t}\,d\mu,$$

and Γ is a contour in $\rho(-A)$ with $\arg\mu\to\pm\theta$ *as $|\mu|\to\infty$ for some θ in $((\pi/2),\pi)$.*

Furthermore, e^{-At} can be continued analytically into a sector $\{t\neq 0\colon |\arg t|<\varepsilon\}$ containing the positive real axis, and if $\operatorname{Re}\sigma(A)>a$; *i.e., if* $\operatorname{Re}\lambda>a$ *whenever $\lambda\in\sigma(A)$, then for $t>0$*

$$\|e^{-At}\|\leqq c\,e^{-at}, \qquad \|A\,e^{-At}\|\leqq\frac{c}{t}e^{-at}$$

for some constant c.

Finally

$$\frac{d}{dt} e^{-At} = -A\, e^{-At} \quad \textit{for } t > 0.$$

The proof of Lemma 2.1. Due to the above proposition, it suffices to prove that $-A$ is a sectorial operator. In view of $D(A_i) = H^2(\Omega) \cap H_0^1(\Omega)$, where $H^i(\Omega) = W^{i,2}(\Omega)$, $i \geqq 0$; $H_0^i(\Omega) = W_0^{i,2}(\Omega)$, it follows that $D(A) = (H^2(\Omega) \cap H_0^1(\Omega))^3$, which implies $-A$ is a closed densely defined operator on X. Since $-A_i$ is a self-adjoint densely defined operator in $L_2(\Omega)$ and is bounded below, it is easy to show that $-A_i$ is a sectorial operator on X_i. Namely, there exist $M_i \geqq 1$, a_i and θ_i satisfying $0 < \theta_i < \pi/2$ such that

$$\rho(-A_i) \supset S_{a_i,\theta_i} = \{\mu \,|\, \theta_i \leqq |\arg(\mu - a_i)| \leqq \pi\}$$

and

$$\|R(\mu, -A_i)\|_{L^2;L^2} \leqq \frac{M_i}{|\mu - a_i|} \quad \forall \mu \in S_{a_i,\theta_i}, \quad i = 1, 2, 3.$$

Clearly, a_i can be taken as the smallest eigenvalue λ_0^i of $(-A_i)$ [14]. ($-\lambda_0^i$ is the largest eigenvalue of A_i.)

Define $\gamma_0 = \min\{\lambda_0^1, \lambda_0^2, \lambda_0^3\}$, $\theta = \max\{\theta_1, \theta_2, \theta_3\}$. It is easy to show, by the form (2.4) of A, that

$$R(\mu, -A) = \begin{pmatrix} R(\mu, -A_1) & 0 & 0 \\ 0 & R(\mu, -A_2) & 0 \\ 0 & 0 & R(\mu, -A_3) \end{pmatrix} \quad \forall \mu \in \rho(A). \tag{2.6}$$

Therefore, it can be claimed that

$$\rho(-A) \supset S_{\gamma_0,\theta}$$

and there exists $M_0 \geqq 1$ such that

$$\|R(\mu, -A)\| \leqq \frac{M_0}{|\mu - \gamma_0|} \quad \forall \mu \in S_{\gamma_0,\theta}.$$

This means $-A$ is a sectorial operator on X, so that by Proposition 2.1, it follows that A generates an analytic semigroup e^{At} on X and satisfies

$$\|e^{At}\| \leqq M\, e^{-\gamma_0 t} \tag{2.7}$$

where M denotes a suitable constant.

In order to investigate the existence and asymptotic properties of the solution to (1.1), (1.2), $(1.3)_{\text{I}}$, we recall the concepts of fractional powers of operators and some results concerned with them [4].

Definition 2.2. Suppose T is a sectorial operator and $\operatorname{Re} \sigma(T) > 0$, then for any $\alpha > 0$

$$T^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-Tt}\, dt.$$

Definition 2.3. With T as above, define T^α = inverse of $T^{-\alpha}$ ($\alpha > 0$), $D(T^\alpha) = R(T^{-\alpha})$; T^0 = identity on X.

DEFINITION 2.4. If T is a sectorial operator in a Banach space X, define for each $\alpha \geqq 0$,

$$X^{\alpha} = D(T_1^{\alpha}) \quad \text{with the graph norm}$$

$$\|X\|_{\alpha} = \|T_1^{\alpha}x\|, \quad x \in X^{\alpha},$$

where $T_1 = T + aI$ with a chosen so Re $\sigma(T_1) > 0$. It can be shown that different choices of a give equivalent norms on X^{α}, so we suppress the dependence on the choice of a.

PROPOSITION 2.2. *If T is sectorial in a Banach space X, then X^{α} is a Banach space in the norm $\|\cdot\|_{\alpha}$ for $\alpha \geqq 0$, $X^0 = X$, and for $\alpha \geqq \beta \geqq 0$, X^{α} is a dense subspace of X^{β} with continuous inclusion.*

DEFINITION 2.5. Suppose there is an extension map $E: C_c^m(\bar{\Omega}) \to C_c^m(\mathbb{R}^n)$, so $E(\phi)$ restricted to $\bar{\Omega}$ is ϕ, such that for the norms of any of the spaces C^{ν} or $W^{k,q}$ $(0 \leqq \nu$, $k \leqq m$ and $1 \leqq q < \infty)$ there is a constant $B > 0$ with

$$B^{-1}\|\phi\|_{\Omega} \leqq \|E(\phi)\|_{\mathbb{R}^n} \leqq B\|\phi\|_{\Omega}.$$

When such an extension map exists, we say Ω has the C^m-extension property.

PROPOSITION 2.3. *Suppose $\Omega \subset \mathbb{R}^n$ is an open set having the C^m extension property, $1 \leqq p < \infty$, and T is a sectorial operator in $X = L_p(\Omega)$ with $D(T) = X^1 \subset W^{m,p}(\Omega)$ for some $m \geqq 1$. Then for $0 \leqq \alpha \leqq 1$.*

$$X^{\alpha} \subset W^{k,q}(\Omega) \quad \textit{when } k - \frac{n}{q} < m\alpha < \frac{n}{p}, \quad q \geqq p$$

$$X^{\alpha} \subset C^{\nu}(\Omega) \quad \textit{when } 0 \leqq \nu < m\alpha - \frac{n}{p}.$$

Furthermore, these are continuous inclusions, provided $\alpha > \theta$, where θ is the number in Nirenberg-Gagliardo inequalities.

Remark. Such an extension map is easily constructed if Ω is bounded and $\partial\Omega$ is a C^m hypersurface separating Ω from $\mathbb{R}^n \backslash \Omega$ [3], and a more complicated construction shows $\partial\Omega$ needs only to be Lipschitzian [13].

LEMMA 2.2. *Let $0 \leqq \alpha \leqq 1$. Suppose X^{α} is a Banach space defined as Definition 2.4 with $T = -A$. Assume $\alpha = \alpha_0 > \frac{3}{4}$. Then*

$$X^{\alpha_0} \subset (L^{\infty}(\Omega) \cap W^{1,2}(\Omega))^3,$$

continuously.

Proof. In view of the form of A_i, it is easy to see that $D(A_i) = W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)$, $i = 1, 2, 3$. By using Proposition 2.3 for $T = -A_i$ with $m = 2$, $p = 2$, $n = 3$, it follows, for $\alpha_0 > \frac{3}{4}$ that

$$X_i^{\alpha_0} \subset C^{\nu}(\Omega) \quad \text{when } 0 \leqq \nu \leqq \frac{4\alpha_0 - 3}{2}.$$

Furthermore, taking $k = 1$, $q = 2$ in Proposition 2.3, it turns out that

$$X_i^{\alpha} \subset W^{1,2}(\Omega).$$

Thus,

$$X_i^{\alpha_0} \subset L^{\infty}(\Omega) \cap W^{1,2}(\Omega), \quad i = 1, 2, 3.$$

The continuity is easily checked. Then, by the special form of A, we obtain Lemma 2.2.

Define

$$\mathscr{D}=\mathbb{R}^+\times\left\{(S,V_1,V_2)\in\bar{X}^{\alpha_0},\ S>-\frac{\min(a_1,a_2)}{2}\right\}.$$

LEMMA 2.3. *The nonlinear operator F is locally Lipschitz continuous on $\mathscr{D}$.*

Proof. Since F does not depend on t explicitly, it suffices to prove that for any $U_i=\{S^{(i)},V_1^{(i)},V_2^{(i)}\}$ with

$$S^i\geqq-\frac{\min(a_1,a_2)}{2}\quad\text{and}\quad\|U_i\|_{\alpha_0}\leqq R^*,\quad R^*>0,\quad i=1,2,$$

$$\|F(U_1)-F(U_2)\|\leqq C\|U_1-U_2\|_{\alpha_0} \tag{2.8}$$

where C depends only on R^* and the parameters in F.

Set

$$F(U)=\begin{pmatrix}f_1(U)\\ f_2(U)\\ f_3(U)\end{pmatrix},$$

$$F(U_1)-F(U_2)=\begin{pmatrix}f_1(U_1)-f_1(U_2)\\ f_2(U_1)-f_2(U_2)\\ f_3(U_1)-f_3(U_2)\end{pmatrix}.$$

By Lemma 2.2, there exists R_1 depending on R^* such that

$$|U_i|_\infty\leqq R_1.$$

From (2.5),

$$\begin{aligned}|f_1(U_1)-f_1(U_2)|\leqq{}&\frac{M_1}{y_1}|V_1^{(1)}-V_1^{(2)}|+\frac{M_2}{y_2}|V_2^{(1)}-V_2^{(2)}|+\frac{\gamma}{K}|S^{(1)}+S^{(2)}||S^{(1)}-S^{(2)}|\\&+\frac{M_1a_1}{y_1}\left|\frac{V_1^{(1)}}{a_1+S^{(1)}}-\frac{V_1^{(2)}}{a_1+S^{(2)}}\right|+\frac{M_2a_2}{y_2}\left|\frac{V_2^{(1)}}{a_2+S^{(1)}}-\frac{V_2^{(2)}}{a_2+S^{(2)}}\right|.\end{aligned}$$

Since

$$\left|\frac{V_1^{(1)}}{a_1+S^{(1)}}-\frac{V_1^{(2)}}{a_1+S^{(2)}}\right|\leqq\frac{2}{a_1}|V_1^{(1)}-V_1^{(2)}|+\frac{4R_1}{a_1^2}|S^{(1)}-S^{(2)}|,$$

$$\left|\frac{V_2^{(1)}}{a_2+S^{(1)}}-\frac{V_2^{(2)}}{a_2+S^{(2)}}\right|\leqq\frac{2}{a_2}|V_2^{(1)}-V_2^{(2)}|+\frac{4R_1}{a_2^2}|S^{(1)}-S^{(2)}|,$$

it follows that

$$\begin{aligned}|f_1(U_1)-f_1(U_2)|\leqq{}&\frac{3M_1}{y_1}\{|V_1^{(1)}-V_1^{(2)}|+|V_2^{(1)}-V_2^{(2)}|\}\\&+R_1\left(\frac{2\gamma}{K}+\frac{4M_1}{a_1}+\frac{4M_2}{a_2}\right)|S^{(1)}-S^{(2)}|.\end{aligned}$$

Similarly,

$$|f_2(U_1)-f_2(U_2)|\leqq 2M_1|V_1^{(1)}-V_1^{(2)}|+\frac{4M_1K_1}{a_1}|S^{(1)}-S^{(2)}|,$$

$$|f_3(U_1)-f_3(U_2)|\leqq 2M_2|V_2^{(1)}-V_2^{(2)}|+\frac{4M_2R_1}{a_2}|S^{(1)}-S^{(2)}|.$$

Then, $\|F(U_1)-F(U_2)\|^2 \leqq C\|U_1-U_2\|^2$ which, together with Proposition 2.3, implies (2.8). Therefore, $F:\mathcal{D}\to X$ is locally Lipschitz continuous.

By Lemmas 2.1 and 2.3, the local existence result follows.

LEMMA 2.4. *For any $U_0=(S_0(x), V_{10}(x), V_{20}(x))\in X^{\alpha_0}$, there exists $T>0$ such that a unique classical solution $t\to U(t)=(S(t), V_1(t), V_2(t))$ of* (2.3) *exists in* $[0, T)$.

Proof. In view of Lemmas 2.1 and 2.3, it can be proved, by local contraction arguments (cf. [4, Thm. 3.3.3]) that for any $(t_0, U_0)\in\mathcal{D}$ there exists $T(t_0, U_0)>0$ such that (2.3) has a unique solution $U(t)$ on (t_0, t_0+T) with initial value $U(t_0)=U_0$, where $t_0=0$. Namely, $U(t)$ is a continuous function: $(0, T)\to X$ such that $U(0)=U_0$, and on $(0, T)$ we have $(t, U(t))\in\mathcal{D}$, $U(t)\in D(A)$. $(dU(t)/dt)$ exists, $t\to F(U(t))$ is locally Hölder continuous, and $\int_0^\rho \|F(U(t))\|\,dt<+\infty$ for some $\rho>0$, the differential equation (2.3) is satisfied on $(0, T)$. In addition $U(t)$ can be expressed as

$$U(t)=\exp(At)U_0+\int_0^t \exp(A(t-s))F(U(s))\,ds. \tag{2.9}$$

But, in fact; when $t>0$, $t\to(dU(t)/dt)\in X^{\alpha_0}$ is locally Hölder continuous due to Lemmas 2.1 and 2.3 (cf. [4, Thm. 3.5.2]), so $(t,x)\to U(t,x)$, $(\partial U/\partial t)(t,x)$ are also continuous on $0<t<T$ and $x\in\bar{\Omega}$. Since $U(t)\in D(A)$ implies $U(t)\in(W^{2,2}(\Omega)\cap W_0^{1,2}(\Omega))^3$, it can be shown, by the embedding theorem, that there exists $\nu>0$ such that $U(t)\in(C^\nu(\Omega))^3$. Then, $F(U(t))\in(C^\nu(\Omega))^3$ follows easily. Therefore we obtain that $U(x,t)\in C^{2+\nu}(\Omega)$. Thus, for $t>0$, $(t,x)\to U(t,x)$ is continuously differentiable in t, twice continuously differentiable in x and we have a classical solution.

Now, we are ready to prove the following.

THEOREM 2.1. *For any $U_0\in X^{\alpha_0}$, $U_0\geqq 0$, the corresponding solution exists globally and is nonnegative for any $t\geqq 0$.*

Proof. Due to $U_0\geqq 0$ and (2.1), it follows that

$$U(t)\geqq 0, \qquad 0\leqq t<T.$$

In addition, since $-R(\mu, A_i)$ maps nonnegative elements of X_i into themselves for real $\mu\in\rho(A_i)$, $-R(\mu, A)$ does the same for real $\mu\in\rho(A)$. On the other hand, for any $t>0$

$$e^{At}x=\lim_{n\to\infty}\left(I-\frac{t}{n}A\right)^{-n}x. \tag{2.10}$$

In view of $-A$ being a sectorial operator, it can be shown that for any fixed $t>0$, $n/t\in\rho(A)$ if n is sufficiently large. Then $e^{At}\cdot U\geqq 0$, $t>0$, if $U\geqq 0$ by (2.10). This, combined with the fact of $F(U)\leqq 0$ for $U\geqq 0$, implies the estimate

$$U(t)\leqq\exp(At)U_0 \quad \text{for all } t\in(0, T).$$

Therefore, by (2.7), we have

$$\|U(t)\|\leqq Me^{-\gamma_0 t}\|U_0\| \quad \text{for all } t\in(0, T), \tag{2.11}$$

which provides an a priori estimate for $U(t)$, whence the global existence of $U(t)$ and its nonnegativity follow in a standard way, together with the validity of (2.11) on $\mathbb{R}^+$.

Next, we prove an existence theorem for nonnegative and bounded solutions of (1.1), (1.2), $(1.3)_{\mathrm{I}}$.

Let λ_0 denote the principal eigenvalue of Laplace operator Δ on $\Omega\in\mathbb{R}^2$ with homogeneous Dirichlet boundary condition. Then $\lambda_0<0$. Set $\Lambda=\max\{k_1\lambda_0+\gamma, k_2\lambda_0+\beta_1, k_3\lambda_0+\beta_2\}$.

THEOREM 2.2. *Suppose* $\Lambda<0$. *Then, for any* $U_0\in X^{\alpha_0}$, $U_0\geqq 0$, *there exists globally a unique classical solution to* (1.1), (1.2), $(1.3)_{\mathrm{I}}$ *that is nonnegative and bounded in the sense of* L_∞.

Proof. Recalling the special form of A and the property of the spectrum for the operator A_i which is strongly elliptic, we may take Λ as $-\gamma_0$ in (2.7), i.e., $-\gamma_0=\Lambda<0$.

From the proof of Theorem 2.1, we read off the estimate

$$0\leqq U(t)\leqq \exp(At)U_0,$$

which implies that

$$|U(t)|_{(L_\infty)^3}\leqq|\exp(At)U_0|_{(L_\infty)^3}.$$

For $U_0\in X^{\alpha_0}$, it has been shown that $\exp(At)U_0\in X^{\alpha_0}$ and $U(t)\in X^{\alpha_0}$. By Lemma 2.2, $X^{\alpha_0}\hookrightarrow(L_\infty(\Omega))^3$ continuously. Therefore, by a known estimate with fractional powers of operators (cf. [4, Thm. 1.4.3]),

$$\begin{aligned}|U(t)|_{(L_\infty)^3}&\leqq C\|\exp(At)U_0\|_{\alpha_0}\leqq C\|(-A)^{\alpha_0}e^{At}\|\,\|U_0\|\\&\leqq C_\alpha t^{-\alpha_0}e^{-\gamma_0 t}\quad\text{for } t>0.\end{aligned}\tag{2.12}$$

Thus, the theorem is proved.

It is easily seen that the state $(0,0,0)$ is an equilibrium solution for (1.1)-$(1.3)_{\mathrm{I}}$. Now, we will discuss the stability of the solution.

THEOREM 2.3. *Suppose* $\Lambda<0$. *Then the unique equilibrium solution* $(0,0,0)$ *of* (1.1)-$(1.3)_{\mathrm{I}}$ *is globally stable. Namely, for any* $U_0\in X^{\alpha_0}$, $U_0\geqq 0$, *assume* $U(t)$ *is the corresponding solution of* (1.1), (1.2), $(1.3)_{\mathrm{I}}$ *as investigated in Theorem* 2.2; *then* $\lim_{t\to\infty}U(t)=0$ *in the sense of* L_∞, *uniformly for* x.

Proof. Since $-\gamma_0=\Lambda<0$, (2.12) gives the result.

Next, we analyze the possibility of relaxing the assumption $\Lambda<0$.

LEMMA 2.5. *Assume* $k_1\lambda_0+\gamma<0$; *then there exists a unique nonnegative equilibrium solution for* (1.1)-$(1.3)_{\mathrm{I}}$ *that is* $(0,0,0)$.

Proof. Let $U^*=(s^*(x),V_1^*(x),V_2^*(x))$ be another nonnegative equilibrium solution. Namely,

$$\left.\begin{aligned}&k_1\Delta S^*+S^*\left(\gamma-\frac{\gamma}{K}S^*-\frac{M_1V_1^*}{y_1(a_1+S^*)}-\frac{M_2V_2^*}{y_2(a_2+S^*)}\right)=0,\\&k_2\Delta V_1^*+k_1\Delta V_1^*+V_1^*\left(\frac{M_1S^*}{a_1+S^*}-D_1\right)=0,\\&k_3\Delta V_2^*+V_2^*\left(\frac{M_2S^*}{a_2+S^*}-D_2\right)=0,\end{aligned}\right\}\quad x\in\Omega,\tag{2.13}$$

$$S^*(x)=V_1^*(x)=V_2^*(x)=0,\qquad x\in\partial\Omega,\tag{2.14}$$

which means $S^*(x)$ is a nonnegative solution of the following linear problem

$$\begin{aligned}&k_1\Delta S+a(x)S=0,\qquad x\in\Omega,\\&S|_{\partial\Omega}=0\end{aligned}$$

where

$$a(x)=\gamma-\frac{\gamma}{K}S^*-\frac{M_1V_1^*}{y_1(a_1+S^*)}-\frac{M_2V_2^*}{y_2(a_2+S^*)}.$$

Due to $S^*(x)\geqq 0$, $V_1^*(x)\geqq 0$, $V_2^*(x)\geqq 0$, it follows that

$$a(x)\leqq \gamma,$$

which, combining with $k_1\lambda_0+\gamma<0$, implies

$$-\lambda_0> a(x)/k_1.$$

Therefore, it can be shown that $S^*(x)=0$, $x\in\bar{\Omega}$. Then it turns out that $V_i^*(x)=0$, $x\in\Omega$, $i=1,2$.

THEOREM 2.4. *Assume* $k_1\lambda_0+\gamma<0$. *Then the unique nonnegative equilibrium state* $(0,0,0)$ *is globally stable in* X; *namely, for any* $U_0\in X^{\alpha_0}$, $U_0\geqq 0$, *suppose* $U(t)$ *is the corresponding solution for* (1.1)-$(1.3)_{\mathrm{I}}$, *then*

$$\lim_{t\to\infty}\|U(t)\|=0.$$

Proof. In view of $k_1\lambda_0+\gamma<0$ and the fact $U(t)\geqq 0$, it is easily proved that

$$\lim_{t\to\infty} S(t)=0 \quad \text{in } L_\infty(\Omega). \tag{2.15}$$

Define a Lyapunov functional as

$$\mathscr{L}(t)=L(V_1,V_2)=\frac{1}{2}\int_\Omega [V_1^2(x,t)+V_2^2(x,t)]\,dx.$$

Then

$$\frac{d\mathscr{L}(t)}{dt}=\int_\Omega\left[V_1\frac{\partial V_1}{\partial t}+V_2\frac{\partial V_2}{\partial t}\right]dx.$$

Noticing the following

$$\frac{\partial V_1}{\partial t}=k_2\Delta V_1+V_1\left(\frac{M_1S}{a_1+S}-D_1\right),$$
$$\frac{\partial V_2}{\partial t}=k_3\Delta V_2+V_2\left(\frac{M_2S}{a_2+S}-D_2\right),$$
$$V_1(0)=V_{10}(x),\qquad V_2(0)=V_{20}(x),$$
$$V_1(t)=V_2(t)\quad\text{on }\partial\Omega,$$

we find that

$$\frac{d\mathscr{L}(t)}{dt}=\int_\Omega\left[k_2V_1\Delta V_1+V_1^2\left(\frac{M_1S}{a_1+S}\right)-D_1V_1^2\right]dx$$
$$+\int_\Omega\left[k_3V_2\Delta V_2+V_2^2\left(\frac{M_2S}{a_2+S}\right)-D_2V_2^2\right]dx.$$

For the first term, we have, by using the Poincaré inequality

$$\int_\Omega\left[k_2V_1\Delta V_1+V_1^2\left(\frac{M_1S}{a_1+S}\right)-D_1V_1^2\right]dx$$
$$=k_2V_1\frac{\partial V_1}{\partial n}\bigg|_{\partial\Omega}-k_2\int_\Omega(\nabla V_1)^2\,dx-D_1\int_\Omega V_1^2\,dx+\int_\Omega V_1^2\left(\frac{M_1S}{a_1+S}\right)dx$$
$$\leqq -k_2\int_\Omega(\nabla V_1)^2-D_1\int_\Omega V_1^2\,dx+\frac{M_1}{a_1}|S|_{L_\infty}\int_\Omega V_1^2\,dx$$
$$\leqq\left(k_2\lambda_0-D_1+\frac{M_1}{a_1}|S(t)|_{L_\infty}\right)\int_\Omega|V_1|^2\,dx.$$

Similarly,

$$\int_\Omega \left[k_3 V_2 \Delta V_2 + V_2^2\left(\frac{M_2 S}{a_2+S}\right) - D_2 V_2^2\right] dx \leqq \left(k_3\lambda_0 - D_2 + \frac{M_2}{a_2}\cdot |S|_{L_\infty}\right)\int_\Omega V_1^2\,dx.$$

In view of $\lim_{t\to\infty} S(t) = 0$ (in the sense of $L_\infty(\Omega)$), it is known that there exists $t_0 > 0$, $\sigma > 0$ such that

$$\max\left\{\left(k_2\lambda_0 - D_1 + \frac{M_1}{a_1}|S|_{L_\infty}\right), \left(k_3\lambda_0 - D_2 + \frac{M_2}{a_2}|S|_{L_\infty}\right)\right\} \leqq -\sigma \quad \text{for } t \geqq t_0.$$

Thus,

$$\frac{d\mathscr{L}(t)}{dt} \leqq -2\sigma\mathscr{L}(t) \quad \text{for } t \geqq t_0.$$

This implies that

$$\mathscr{L}(t) \leqq \mathscr{L}(t_0)\, e^{-2\sigma(t-t_0)}, \qquad t \geqq t_0,$$

so

$$\lim_{t\to\infty} \mathscr{L}(t) = \lim_{t\to\infty} L(V_1, V_2) = 0,$$

which with (2.15) gives the result.

We will show that the theorem is not true, if the assumption $k_1\lambda_0 + \gamma < 0$ does not hold.

THEOREM 2.5. *Assume* $k_1\lambda_0 + \gamma \geqq 0$. *Then there exist at least two nonnegative equilibrium solutions for* (1.1)–(1.3)$_{\text{I}}$. *Therefore it is impossible for the solution* (0, 0, 0) *to be globally stable.*

Proof. It suffices to prove that there exists a nontrivial nonnegative solution for the following problem:

$$\begin{aligned} &k_1\Delta S + S\left(\gamma - \frac{\gamma}{K}S - \frac{M_1 V_1}{y_1(a_1+S)} - \frac{M_2 V_2}{y_2(a_2+S)}\right) = 0, \\ &k_2\Delta V_1 + V_1\left(\frac{M_1 S}{a_1+S} - D_1\right) = 0, \qquad x \in \Omega, \\ &k_3\Delta V_2 + V_2\left(\frac{M_2 S}{a_2+S} - D_2\right) = 0, \\ &S(x) = V_1(x) = V_2(x) = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

In fact, we may take $V_i(x)$ as $V_i(x) \equiv 0$, $i = 1, 2$, and $S(x)$ as the nonnegative solution of the following problem:

$$\begin{aligned} &k_1\Delta S + S\left(\gamma - \frac{\gamma}{K}S\right) = 0, \qquad x \in \Omega, \\ &S|_{\partial\Omega} = 0, \end{aligned}$$

i.e.

$$\begin{aligned} &-\Delta S = \frac{\gamma}{k_1}S\left(1 - \frac{S}{K}\right), \\ &S|_{\partial\Omega} = 0. \end{aligned} \tag{2.16}$$

Due to $k_1\lambda_0+\gamma \geqq 0$ it can be shown, by comparison theorems and monotonicity methods, (cf. [12]) that there exists a unique solution $S^*(x)$ for (2.16) such that $0 < S^*(x) \leqq K$, $x \in \Omega$. Thus, $(S^*(x), 0, 0)$ is a nontrivial solution for (1.1)-(1.3)$_{\mathrm{I}}$ which is nonnegative. Therefore, $(0, 0, 0)$ cannot be globally stable.

We discuss the stability of $U^* = (S^*(x), 0, 0)$ now. We adopt a standard "linearization procedure" relative to the nontrivial equilibrium solution U^*. In doing so, we take the linear expansion of F at U^*, i.e., $F(U^*+U) = F(U^*) + BU + g(U)$ and make a careful analysis of the spectrum of $\mathcal{U} = A + B$.

Consider

$$F(U^*+U) = \begin{pmatrix} f_1(U^*+U) \\ f_2(U^*+U) \\ f_3(U^*+U) \end{pmatrix}$$

where

$$\begin{aligned} f_1(U^*+U) &= -\frac{M_1V_1}{y_1} - \frac{M_2V_2}{y_2} - \frac{\gamma}{K}(S+S^*)^2 + \frac{M_1V_1a_1}{y_1(a_1+S+S^*)} + \frac{M_2V_2a_2}{y_2(a_2+S+S^*)} \\ &= -\frac{M_1V_1}{y_1} - \frac{M_2V_2}{y_2} - \frac{\gamma}{K}S^2 - \frac{2\gamma}{K}S^*S - \frac{\gamma}{K}S^* + \frac{M_1a_1}{y_1(a_1+S^*)}V_1 \\ &\quad - \frac{M_1a_1}{y_1(a_1+S^*)(a_1+S+S^*)}V_1S + \frac{M_2a_2}{y_2(a_2+S^*)}V_2 \\ &\quad - \frac{M_2a_2}{y_2(a_2+S^*)(a_2+S+S^*)}V_2S \\ &= f_1(U^*) - \frac{S^*M_1V_1}{y_1(a_1+S^*)} - \frac{S^*M_2V_2}{y_2(a_2+S^*)} - \frac{2\gamma}{K}S^*S \\ &\quad - \frac{\gamma}{K}S^2 - \frac{M_1a_1V_1S}{y_1(a_1+S^*)(a_1+S^*+S)} - \frac{M_2a_2V_2S}{y_2(a_2+S^*)(a_2+S^*+S)}, \\ f_2(U^*+U) &= -\frac{M_1a_1V_1}{a_1+S+S^*} = -\frac{M_1a_1}{a_1+S^*}V_1 + \frac{M_1a_1}{(a_1+S^*)(a_1+S^*+S)}V_1S, \\ f_3(U^*+U) &= -\frac{M_2a_2V_2}{a_2+S+S^*} = -\frac{M_2a_2}{a_2+S^*}V_2 + \frac{M_2a_2}{(a_2+S^*)(a_2+S^*+S)}V_2S. \end{aligned}$$

Thus

$$B = \begin{bmatrix} -\dfrac{2\gamma}{K}S^* & -\dfrac{M_1S^*}{y_1(a_1+S^*)} & -\dfrac{M_2S^*}{y_2(a_2+S^*)} \\ 0 & -\dfrac{M_1a_1}{a_1+S^*} & 0 \\ 0 & 0 & -\dfrac{M_2a_2}{a_2+S^*} \end{bmatrix}, \tag{2.17}$$

$$g(U) = \begin{pmatrix} -\dfrac{M_1a_1V_1S}{y_1(a_1+S^*)(a_1+S^*+S)} - \dfrac{M_2a_2V_2S}{y_2(a_2+S^*)(a_2+S^*+S)} - \dfrac{\gamma}{K}S^2 \\ \dfrac{M_1a_1V_1S}{(a_1+S^*)(a_1+S^*+S)} \\ \dfrac{M_2a_2V_2S}{(a_2+S^*)(a_2+S^*+S)} \end{pmatrix}. \tag{2.18}$$

Let us recall a theorem about stability by the linear approximation [4].

Let A be a sectorial linear operator in a Banach space X, and let $f: U \to X$ where U is a cylindrical neighborhood in $\mathbb{R} \times X^\alpha$ (for some $\alpha < 1$) of $(\tau, \infty) \times \{x_0\}$. We say x_0 is an equilibrium point if $x(t) \equiv x_0$ is a solution of $(dx/dt) + Ax = f(t, x)$, $t > t_0$, i.e. if $x_0 \in D(A)$ and $Ax_0 = f(t, x_0)$ for all $t > t_0$.

A solution $\bar{x}(\cdot)$ on $[t_0, \infty)$ is stable (in X^α) if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that any solution x with $\|x(t_0) - \bar{x}(t_0)\|_\alpha < \delta$ exists on $[t_0, \infty)$ and satisfies $\|x(t) - \bar{x}(t)\|_\alpha < \varepsilon$ for all $t \geqq t_0$; that is, if $x_0 \mapsto x(t; t_0, x_0)$ is continuous (in X^α) at $x_0 = \bar{x}(t_0)$, uniformly in $t \geqq t_0$. The solution $\bar{x}$ is uniformly stable if $x_1 \mapsto x(t; t_1, x_1)$ is continuous as $x_1 \mapsto \bar{x}(t_1)$, uniformly in $t \geqq t_1$ and $t_1 \geqq t_0$.

The solution $\bar{x}(\cdot)$ is uniformly asymptotically stable if it is uniformly stable and $\|x(t; t_1, x_1) - \bar{x}(t)\| \to 0$ as $|t - t_1| \to +\infty$, uniformly in $t_1 \geqq t_0$ and $\|x_1 - \bar{x}(t_1)\|_\alpha < \delta$ for some constant $\delta > 0$.

PROPOSITION 2.4 (cf. [4, Thm. 5.1.1]). *Let A, f be as above and let x_0 be an equilibrium point. Suppose*

$$f(t, x_0 + z) = f(t, x_0) + Bz + g(t, z)$$

where B is a bounded linear map from X^α to X and $\|g(t, z)\| = O(\|z\|_\alpha)$ as $\|z\|_\alpha \to 0$, uniformly in $t > \tau$, and $f(t, x)$ is locally Hölder continuous in t, locally Lipschitzian in x, on U.

If the spectrum of $A-B$ lies in $\{\operatorname{Re} \lambda > \beta\}$ for some $\beta > 0$, or equivalently if the linearization

$$\frac{dz}{dt} + Az = Bz$$

is uniformly asymptotically stable, then the original equation has the solution x_0 uniformly asymptotically stable in X^α. More precisely, there exist $\rho > 0$, $M \geqq 1$ such that if $t_0 > \tau$ and $\|x_1 - x_0\|_\alpha \leqq \rho/2M$, then there exists a unique solution of

$$\frac{dx}{dt} + Ax = f(t, x), \quad t > t_0, \quad x(t_0) = x_1$$

existing on $t_0 \leqq t < \infty$ and satisfying for $t \geqq t_0$

$$\|x(t; t_0, x_1) - x_0\|_\alpha \leqq 2M e^{-\beta(t - t_0)} \|x_1 - x_0\|_\alpha.$$

From (2.17), (2.18), it is easy to check the validity of Proposition 2.4 for our case. Then, in order to investigate the stability property of U^*, it suffices to analyze the spectrum of $\mathcal{U} = A + B$.

Let us write down the resolvent equation for $\mathcal{U}$:

$$(A + B - \mu I) U = V, \quad \mu \in \mathbb{C}, \quad V \in X,$$

where

$$U = (u_1, u_2, u_3), \qquad V = (V_1, V_2, V_3).$$

Namely,

$$\begin{aligned} &\left(A_1 - \frac{2\gamma}{K} S^* - \mu\right) u_1 - \frac{M_1 S^*}{y_1(a_1 + S^*)} u_2 - \frac{M_2 S^*}{y_2(a_2 + S^*)} u_3 = V_1, \\ &\left(A_2 - \frac{M_1 a_1}{a_1 + S^*} - \mu\right) u_2 = V_2, \\ &\left(A_3 - \frac{M_2 a_2}{a_2 + S^*} - \mu\right) u_3 = V_3. \end{aligned} \tag{2.19}$$

As

$$\mu \in \rho\left(A_2 - \frac{M_1 a_1}{a_1 + S^*}\right) \cap \rho\left(A_3 - \frac{M_2 a_2}{a_2 + S^*}\right),$$

the system (2.19) can be written as

$$\left(A_1 - \frac{2\gamma}{K} S^* - \mu\right) u_1 = V_1 + \frac{M_1 S^*}{y_1(a_1 + S^*)} R\left(\mu, A_2 - \frac{a_1 M_1}{a_1 + S^*}\right) V_2 + \frac{M_2 S^*}{y_2(a_2 + S^*)} R\left(\mu, A_3 - \frac{a_2 M_2}{a_2 + S^*}\right) V_3,$$

$$u_2 = R\left(\mu, A_2 - \frac{M_1 a_1}{a_1 + S^*}\right) V_2,$$

$$u_3 = R\left(\mu, A_3 - \frac{M_2 a_2}{a_2 + S^*}\right) V_3.$$

Denote

$$\bar{A}_1 = A_1 - \frac{2\gamma}{K} S^* - \mu, \qquad \bar{A}_2 = A_2 - \frac{M_1 a_1}{a_1 + S^*}, \qquad \bar{A}_3 = A_3 - \frac{M_2 a_2}{a_2 + S^*}.$$

Set $\mathcal{B} = \{\mu \colon \mu \in \rho(\bar{A}_2) \cap \rho(\bar{A}_3),\ \bar{A}_1(\mu) \text{ is invertible in } L^2(\Omega)\}$. The following lemma is easily proved.

LEMMA 2.6. *$\rho(\mathcal{U}) \supset \mathcal{B}$ and $R(\mu, \mathcal{U})$ is given by*

$$R(\mu, \mathcal{U}) = \begin{bmatrix} \bar{A}_1^{-1}(\mu) & \bar{A}_1^{-1} \dfrac{M_1 S^*}{y_1(a_1 + S^*)} R(\mu, \bar{A}_2) & \bar{A}_1^{-1} \dfrac{M_2 S^*}{y_2(a_2 + S^*)} R(\mu, \bar{A}_3) \\ 0 & R(\mu, \bar{A}_2) & 0 \\ 0 & 0 & R(\mu, \bar{A}_3) \end{bmatrix} \quad \textit{for } \mu \in \rho(\mathcal{U}).$$

Thus, in order to analyze the spectrum of $\mathcal{U}$, we have to discuss the invertibility of $\bar{A}_1(\mu)$. Clearly, the following lemma can be obtained.

LEMMA 2.7. *The operator $\bar{A}_1(\mu)$, $\mu \in \rho(\bar{A}_2) \cap \rho(\bar{A}_3)$ is invertible in $L^2(\Omega)$ if and only if zero is not an eigenvalue of $\bar{A}_1(\mu)$.*

Set $K^* = \max\{\operatorname{Re} \lambda,\ \lambda \in \sigma(\bar{A}_2) \cup \sigma(\bar{A}_3)\}$.

LEMMA 2.8. *Assume $K^* < 0$; then there exist $\theta^* \in (0, (\pi/2))$ and $\gamma^* > 0$ such that (see Fig. 1)*

$$\mathcal{B} \supset S^* = \left\{\mu \,\middle|\, |\arg(\mu + \gamma^*)| \leq \frac{\pi}{2} + \theta^*\right\}.$$

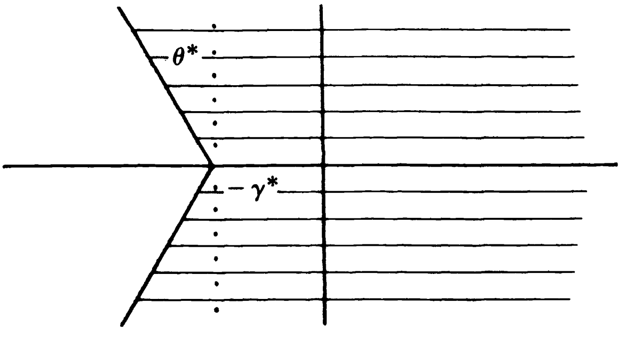


FIG. 1

Proof. Due to Lemma 2.7, it suffices to show that there exist $\theta^* \in (0, (\pi/2))$ and $\gamma^* > 0$ such that $\bar{A}_1(\mu)$ does not have zero as an eigenvalue whenever μ belongs to S^*.

Denote an arbitrary eigenvalue of $\bar{A}_1(\mu)$ by $\zeta(\mu)$ and any eigenfunction belonging to the corresponding eigenspace $\bar{W}_\mu$ by $W(\mu)$, normalized to 1 in the L^2-norm. Let us put $\mu_1 = \operatorname{Re} \mu$, $u_2 = I_m\mu$, $\zeta_1(\mu_1, \mu_2) = \operatorname{Re} \zeta(\mu)$, $\zeta_2(\mu_1, \mu_2) = \operatorname{Im} \zeta(\mu)$. Due to

$$\bar{A}_1(\mu) W(\mu) = \zeta(\mu) W(\mu),$$
$$\langle \bar{A}_1(\mu) W(\mu), W(\mu) \rangle = \zeta(\mu),$$
$$\langle \bar{A}_1(\mu) W(\mu), W(\mu) \rangle = \left\langle \left(A_1 - \frac{\gamma}{K} S^* \right) W(\mu), W(\mu) \right\rangle + \left\langle -\frac{\gamma}{K} S^* W(\mu), W(\mu) \right\rangle - \mu,$$

it turns out that

$$\zeta_1(\mu_1, \mu_2) = \left\langle \left(A_1 - \frac{\gamma}{K} S^* \right) W(\mu), W(\mu) \right\rangle + \left\langle -\frac{\gamma}{K} S^* W(\mu), W(\mu) \right\rangle - \mu_1,$$
$$\xi_2(\mu_1, \mu_2) = -\mu_2.$$

Since $S^*(x) > 0$, $x \in \Omega$ satisfying $(A_1 - (\gamma/K) S^*) S^* = 0$, $S^*|_{\partial\Omega} = 0$, which means that the second order elliptic operator $A_1 - (\gamma/K) S^*$ has $S^* \in H_0^1(\Omega)$ as positive eigenfunction with zero eigenvalue, it follows that the pure point spectrum of $A_1 - (\gamma/K) S^*$ is localized on the nonpositive half-axis, and

$$\left\langle \left(A_1 - \frac{\gamma}{K} S^* \right) u, u \right\rangle \leqq 0 \quad \text{for any } u \in H_0^1(\Omega).$$

Therefore, $\zeta_1(\mu_1, \mu_2) \leqq -\mu_1 - \langle (\gamma/K) S^* W(\mu), W(\mu) \rangle < 0$ for any $\mu_1 \geqq 0$. In view of the inequality $\zeta_1(0, \mu_2) < 0$ and the continuous dependence of $\zeta(\mu)$ on μ, it turns out that there exist $c > 0$, $\alpha > 0$, $-\alpha \in (k^*, 0)$ such that

$$\zeta_1(\mu_1, \mu_2) < 0 \quad \text{for any } \mu \in \rho(\bar{A}_2) \cap \rho(\bar{A}_3) \quad \text{with } \mu_1 \in (-\alpha, 0),$$

$|\mu_2| \leqq C$, while $|\zeta_2(\mu_1, \mu_2)| = |\mu_2| > C > 0$ when $|\mu_2| > C$. These facts show that $\bar{A}_1(\mu)$, $\mu \in \rho(\bar{A}_2) \cap \rho(\bar{A}_3)$ has possibly zero as an eigenvalue only for such kind of μ belonging to the subset $\{\mu \in \rho(\bar{A}_2) \cap \rho(\bar{A}_3)$ with $\operatorname{Re} \mu < -\alpha$, $|\operatorname{Im} \mu| < C\}$. Then the lemma is proved for any $\theta^*(0, \operatorname{arctg}(\alpha - \gamma^*/C))$, $\gamma^* \in (0, \alpha)$. (See Fig. 2.)

It is easily seen from Lemmas 2.6–2.8, that [10]

LEMMA 2.9.

$$\sigma(\mathcal{U}) \subset \{\lambda \in \mathbb{C}, \operatorname{Re} \lambda < -\gamma^*\}.$$

This implies, by Proposition 2.4, the next theorem.

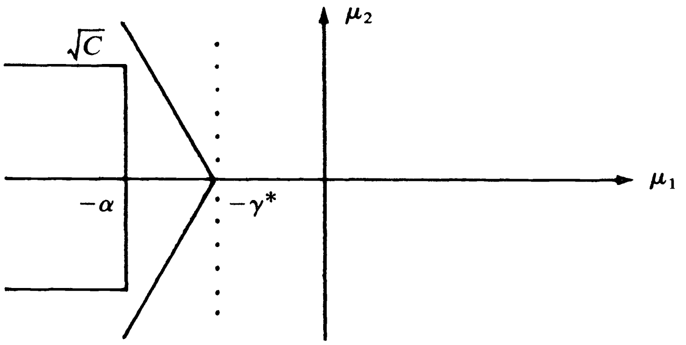


FIG. 2

THEOREM 2.6. *Assume $k_1\lambda_0+\gamma\geqq 0$ and $k^*<0$. Then the equilibrium solution $U^*=(S^*(x),0,0)$ is uniformly asymptotically stable in X^{α_0}. Namely, there exist $\rho>0$, $M\geqq 1$ such that if $\|U_0-U^*\|_{\alpha_0}\leqq \rho/2M$, $U_0\geqq 0$, then there exists a unique solution $U(t)$ of (2.3) on $0\leqq t<+\infty$ satisfying*

$$\|U(t)-U^*\|_{\alpha_0}\leqq 2Me^{-\gamma^*t}\|U_0-U^*\|_{\alpha_0}\to 0 \quad as\ t\to+\infty.$$

3. The case of Neumann conditions. We deal with problems (1.1), (1.2) and $(1.3)_{\text{II}}$ in this section. First, consider the following general form.

$$\begin{aligned} &u_{1t}-k_1\Delta u_1=f_1(u_1,u_2,u_3),\\ &u_{2t}-k_2\Delta u_2=f_2(u_1,u_2), \qquad \text{in } R^+\times\Omega,\\ &u_{3t}-k_3\Delta u_3=f_3(u_1,u_3), \end{aligned} \tag{3.1}$$

together with the boundary and initial conditions

$$B_i[u_i]\equiv\alpha_i(x)\frac{\partial u_i}{\partial n}+\beta_i(x)u_i=0 \quad \text{on } \partial\Omega\times R^+, \tag{3.2}$$

$$u_i(0,x)=u_{0i}(x) \quad \text{on } \Omega \tag{3.3}$$

where Ω is a bounded domain in R^n $(n=1,2,\cdots)$, $\partial\Omega$ is the boundary of Ω, $\alpha_i\geqq 0$, $\beta_i\geqq 0$ with $\alpha_i+\beta_i>0$ on $\partial\Omega$, $i=1,2,3$. $\partial/\partial n$ represents the outward normal derivative at the boundary.

Assume that α_i, β_i, u_{0i} are smooth nonnegative functions with $u_{0i}\not\equiv 0$, f_1 is continuously differentiable in $R^+\times R^+\times R^+$ and f_2, f_3 is continuously differentiable in $R^+\times R^+$. Ω is smooth, $\partial\Omega$ belongs to $C^{1+\alpha}$ $(\alpha>0)$. In addition, in order to employ the monotone argument to establish an existence-comparison theorem in terms of upper and lower solutions for (3.1)-(3.3), we assume

$$\begin{aligned} &\frac{\partial f_1}{\partial u_2}(u_1,u_2,u_3)\leqq 0, \qquad \frac{\partial f_1}{\partial u_3}(u_1,u_2,u_3)\leqq 0,\\ &\frac{\partial f_2}{\partial u_1}\geqq 0, \quad \frac{\partial f_3}{\partial u_1}\geqq 0 \quad \text{for } u_i\geqq 0, \quad i=1,2,3. \end{aligned} \tag{H}$$

Obviously, the reaction terms in our model (1.1) satisfy (H). The precise definition of upper and lower solutions are given as follows.

DEFINITION 3.1. Let $\bar U=(\bar u_1,\bar u_2,\bar u_3)$, $\underline{U}=(\underline{u}_1,\underline{u}_2,\underline{u}_3)$ be an ordered pair of smooth functions in D_T satisfying the following inequalities (3.4), (3.5), (3.6). Then $\bar U$, $\underline{U}$ are called upper and lower solutions of (3.1)-(3.3), respectively.

$$\begin{aligned} &\bar u_{1t}-k_1\Delta\bar u_1\geqq f_1(\bar u_1,\underline{u}_2,\underline{u}_3),\\ &\bar u_{2t}-k_2\Delta\bar u_2\geqq f_2(\bar u_1,\bar u_2), \qquad (t,x)\in D_T,\\ &\bar u_{3t}-k_3\Delta\bar u_3\geqq f_3(\bar u_1,\bar u_3), \end{aligned} \tag{3.4}$$

$$\begin{aligned} &\underline{u}_{1t}-k_1\Delta\underline{u}_1\leqq f_1(\underline{u}_1,\bar u_2,\bar u_3),\\ &\underline{u}_{2t}-k_2\Delta\underline{u}_2\leqq f_2(\underline{u}_1,\underline{u}_2), \qquad (t,x)\in D_T,\\ &\underline{u}_{3t}-k_3\Delta\underline{u}_3\leqq f_3(\underline{u}_1,\underline{u}_3), \end{aligned} \tag{3.5}$$

$$(3.6)\qquad \begin{aligned} &B_i[\bar{u}_i] \geqq 0 \geqq B_i[\underline{u}_i], \quad (t,x)\in S_T,\\ &\bar{u}_i(0,x) \geqq u_{i0}(x) \geqq \underline{u}_i(0,x), \quad x\in\Omega, \quad i=1,2,3, \end{aligned}$$

where $D_T=(0,T]\times\Omega$, $S_T=(0,T]\times\partial\Omega$, $T<\infty$ but can be arbitrarily large.

Suppose that for given reaction functions f_1, f_2, f_3 satisfying the above assumptions there exists an ordered pair of upper and lower solutions $\bar{U}=(\bar{u}_1,\bar{u}_2,\bar{u}_3)$, $\underline{U}=(\underline{u}_1,\underline{u}_2,\underline{u}_3)$. Define

$$\mathscr{L}(D_T)=\{(u_1,u_2,u_3)\colon u_i\in C(\bar{D}_T),\ \underline{u}_i\leqq u_i\leqq\bar{u}_i \text{ on } \bar{D}_T,\ i=1,2,3\}$$

and set

$$(3.7)\qquad l_i=\sup_{(u_1,u_2,u_3)\in\mathscr{L}(D_T)}\left\{\left|\frac{\partial f_i}{\partial u_i}\right|\right\}.$$

To establish an existence-comparison theorem we consider the sequence $\{U^{(k)}\}=\{u_1^{(k)}, u_2^{(k)}, u_3^{(k)}\}$ obtained from the linear problem.

$$(3.8)\qquad (u_i^{(k)})_+ - k_i\Delta u_i^{(k)} + l_i u_i^{(k)} = l_i u_i^{(k-1)} + f_i(u_1^{(k-1)}, u_2^{(k-1)}, u_3^{(k-1)}), \quad (t,x)\in D_T,$$

$$(3.9)\qquad B_i[u_i^{(k)}]=0, \quad (t,x)\in S_T,$$

$$(3.10)\qquad u_i^{(k)}(0,x)=u_{i0}(x), \quad x\in\Omega$$

where $i=1,2,3$ and $k=1,2,\cdots$.

For each k, the above system consists of three linear uncoupled initial-boundary value problems. Therefore the existence of $\{u_1^{(k)}, u_2^{(k)}, u_3^{(k)}\}$ follows from the standard existence theorem for scalar systems. To ensure that $\{u_1^{(k)}, u_2^{(k)}, u_3^{(k)}\}$ is a monotone sequence and converges to a unique solution of (3.1)-(3.3) we use the initial iteration $(\bar{u}_1^{(0)}, \underline{u}_2^{(0)}, \underline{u}_3^{(0)})=(\bar{u}_1, \underline{u}_2, \underline{u}_3)$ to construct the sequence $\{\bar{u}_1^{(k)}, \underline{u}_2^{(k)}, \underline{u}_3^{(k)}\}$ from the equations

$$(3.11)\qquad \begin{aligned} &\bar{u}_{1t}^{(k)} - k_1\Delta\bar{u}_1^{(k)} + l_1\bar{u}_1^{(k)} = l_1\bar{u}_1^{(k-1)} + f_1(\bar{u}_1^{(k-1)}, \underline{u}_2^{(k-1)}, \underline{u}_3^{(k-1)}),\\ &\underline{u}_{2t}^{(k)} - k_2\Delta\underline{u}_2^{(k)} + l_2\underline{u}_2^{(k)} = l_2\underline{u}_2^{(k-1)} + f_2(\underline{u}_1^{(k-1)}, \underline{u}_2^{(k-1)}),\\ &\underline{u}_{3t}^{(k)} - k_3\Delta\underline{u}_3^{(k)} + l_3\underline{u}_3^{(k)} = l_3\underline{u}_3^{(k-1)} + f_3(\underline{u}_1^{(k-1)}, \underline{u}_3^{(k-1)}), \end{aligned}$$

while the sequence $\{\underline{u}_1^{(k)}, \bar{u}_2^{(k)}, \bar{u}_3^{(k)}\}$ with $(\underline{u}_1^{(0)}, \bar{u}_2^{(0)}, \bar{u}_3^{(0)})=(\underline{u}_1, \bar{u}_2, \bar{u}_3)$ is determined from the equations

$$(3.12)\qquad \begin{aligned} &\underline{u}_{1t}^{(k)} - k_1\Delta\underline{u}_1^{(k)} + l_1\underline{u}_1^{(k)} = l_1\underline{u}_1^{(k-1)} + f_1(\underline{u}_1^{(k-1)}, \bar{u}_2^{(k-1)}, \bar{u}_3^{(k-1)}),\\ &\bar{u}_{2t}^{(k)} - k_2\Delta\bar{u}_2^{(k)} + l_2\bar{u}_2^{(k)} = l_2\bar{u}_2^{(k-1)} + f_2(\bar{u}_1^{(k-1)}, \bar{u}_2^{(k-1)}),\\ &\bar{u}_{3t}^{(k)} - k_3\Delta\bar{u}_3^{(k)} + l_3\bar{u}_3^{(k)} = l_3\bar{u}_3^{(k-1)} + f_3(\bar{u}_1^{(k-1)}, \bar{u}_3^{(k-1)}). \end{aligned}$$

In each system, the boundary and initial conditions are (3.9), (3.10). These two systems are interrelated. With this construction we prove our existence-comparison theorem.

THEOREM 3.1. *Suppose the pair of upper and lower solution* $(\bar{u}_1, \bar{u}_2, \bar{u}_3)$, $(\underline{u}_1, \underline{u}_2, \underline{u}_3)$ *can be chosen. Then the sequences*

$$\{\bar{U}^{(k)}\}=\{\bar{u}_1^{(k)}, \bar{u}_2^{(k)}, \bar{u}_3^{(k)}\}, \qquad \{\underline{U}^{(k)}\}=\{\underline{u}_1^{(k)}, \underline{u}_2^{(k)}, \underline{u}_3^{(k)}\}$$

obtained from (3.11), (3.12), (3.9) *and* (3.10) *converge monotonically from above and below, respectively, to a unique solution* $\{u_1, u_2, u_3\}$ *of* (3.1)-(3.3) *such that*

$$\underline{u}_i(t,x)\leqq u_i(t,x)\leqq\bar{u}_i(t,x), \quad (t,x)\in\bar{D}_T, \quad i=1,2,3.$$

Proof. Let $u_1^* = \bar{u}_1 - \bar{u}_1^{(1)}$, $u_2^* = \underline{u}_2^{(1)} - \underline{u}_2$, $u_3^* = \underline{u}_3^{(1)} - \underline{u}_3$. Then by (3.4), (3.5), (3.6), (3.11) and (3.12), it follows that

$$u_{it}^* - k_i \Delta u_i^* + l_i u_i^* \geqq 0, \tag{3.13}$$

$$B_i[u_i^*] \geqq 0, \tag{3.14}$$

$$u_i^*(0, x) \geqq 0, \qquad i = 1, 2, 3. \tag{3.15}$$

By the maximum principle, the above inequalities imply that $u_i^* \geqq 0$ on D_T, $i = 1, 2, 3$; namely, $\bar{u}_1 \geqq \bar{u}_1^{(1)}$, $\underline{u}_2 \leqq \underline{u}_2^{(1)}$, $\underline{u}_3 \leqq \underline{u}_3^{(1)}$. Similarly, it can be shown that $\underline{u}_1 \leqq \underline{u}_1^{(1)}$, $\bar{u}_2 \geqq \bar{u}_2^{(1)}$, $\bar{u}_3 \geqq \bar{u}_3^{(1)}$.

Now let $u_i^* = \bar{u}_i^{(1)} - \underline{u}_i^{(1)}$; then the assumptions (H), (3.7) and the relations in (3.11) and (3.12) imply that

$$u_{it}^* - k_i \Delta u_i^* + l_i u_i^* \geqq 0, \qquad i = 1, 2, 3.$$

Since $B_i[u_i^*] = 0$, $u_i^*(0, x) = 0$ by using the maximum principle we obtain $u_i^* \geqq 0$, i.e., $\underline{u}_i^{(1)} \leqq \bar{u}_i^{(1)}$, $i = 1, 2, 3$. The above conclusions lead to the relation

$$\underline{u}_i = \underline{u}_i^{(0)} \leqq \underline{u}_i^{(1)} \leqq \bar{u}_i^{(1)} \leqq \bar{u}_i^{(0)} = \bar{u}_i, \qquad i = 1, 2, 3.$$

Then, by induction, it is not difficult to show that

$$\underline{u}_i^{(k-1)} \leqq \underline{u}_i^{(k)} \leqq \bar{u}_i^{(k)} \leqq \bar{u}_i^{(k-1)}, \quad i = 1, 2, 3, \quad k = 1, 2, \cdots.$$

It follows from this monotone property that the pointwise limits $\lim_{k\to\infty} \bar{u}_i^{(k)}(t, x) = \tilde{u}_i(t, x)$, $\lim_{k\to\infty} \underline{u}_i^{(k)}(t, x) = \underset{\sim}{u}_i(t, x)$, $i = 1, 2, 3$, exist and $\underset{\sim}{u}_i \leqq \tilde{u}_i$ on $\bar{D}_T$.

By a standard regularity argument and the similar approach used by Pao in [11], it can be shown that $(\underset{\sim}{u}_1, \underset{\sim}{u}_2, \underset{\sim}{u}_3) = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3) = (u_1, u_2, u_3)$ is the unique solution of (3.1)-(3.3). Obviously, the solution (u_1, u_2, u_3) satisfies $\underline{u}_i \leqq u_i \leqq \bar{u}_i$, $i = 1, 2, 3$. Theorem 3.1 is proved.

For our problem (1.1)-(1.3)$_{\text{II}}$, we choose the pair of upper and lower solution as follows:

Set

$$\hat{S} = \sup_{\bar{\Omega}} S_0(x), \qquad \hat{V}_i = \sup_{\bar{\Omega}} V_{i0}(x), \qquad i = 1, 2.$$

Then solving the following problem:

$$\begin{aligned} &\dot{S}(t) = \gamma S - \frac{\gamma}{K} S^2, && S(0) = \hat{s}, \\ &\dot{V}_1(t) = \left(\frac{M_1 S}{a_1 + S} - D_1\right) V_1, && V_1(0) = \hat{V}_1, \\ &\dot{V}_2(t) = \left(\frac{M_2 S}{a_2 + S} - D_2\right) V_2, && V_2(0) = \hat{V}_2, \end{aligned} \tag{3.16}$$

we obtain

$$\begin{aligned} &S(t) = \frac{K}{1 + \left(\dfrac{K - \hat{s}}{\hat{s}}\right) e^{-\gamma t}}, \\ &V_i(t) = \hat{V}_i \exp\left[\int_0^t \left(\frac{M_i S(\eta)}{a_i + S(\eta)} - D_i\right) d\eta\right], \qquad i = 1, 2. \end{aligned} \tag{3.17}$$

Clearly, $(0,0,0)$ and $(S(t), V_1(t), V_2(t))$ are a pair of lower and upper solutions of (1.1)-$(1.3)_{\text{II}}$. Thus, Theorem 3.1 implies the following.

THEOREM 3.2. *There exists a unique solution of* (1.1)-$(1.3)_{\text{II}}$ *for* $(t,x)\in R^+\times\Omega$, *such that*

$$\begin{aligned} &0\leqq S(t,x)\leqq K\left[1+\frac{K-\hat{s}}{\hat{s}}\exp(-\gamma t)\right]^{-1}=S(t),\\ &0\leqq V_i(t,x)\leqq \hat{V}_i\exp[B_i(t)] \end{aligned}\tag{3.18}$$

where

$$B_i(t)=\int_0^t\left[\frac{M_iS(\eta)}{a_i+S(\eta)}-D_i\right]d\eta,\qquad i=1,2.$$

We next investigate the asymptotic properties of the solution.

From (3.16), (3.17) it is easy to see that $S(t)$ is increasing with $S(t)<K$ and $S(t)\to K$ as $t\to+\infty$ if $\hat{s}<K$, while $S(t)$ is decreasing with $S(t)>K$ and $S(t)\to K$ as $t\to+\infty$ if $\hat{s}>K$.

We consider the case $\hat{s}\leqq K$ first.

THEOREM 3.3. *Let* $M_iK/(a_i+K)<D_i$ $(i=1,2)$ *and* $0<S_0(x)\leqq K$ *for* $x\in\bar{\Omega}$. *Then the unique global solution* (S, V_1, V_2) *to the Neumann problem* (1.1)-$(1.3)_{\text{II}}$ *satisfies*

$$\lim_{t\to\infty}S(t,x)=K,\quad \lim_{t\to\infty}V_i(t,x)=0,\quad i=1,2.\tag{3.19}$$

Proof. Instead of (3.16) we consider

$$\begin{aligned} &\dot{S}(t)=0, \qquad S(0)=K,\\ &\dot{\bar{V}}_i(t)=\left(\frac{M_iK}{a_i+K}-D_i\right)\bar{V}_i,\quad \bar{V}_i(0)=\hat{V}_i,\quad i=1,2. \end{aligned}\tag{3.20}$$

In view of $S_0(x)\leqq k$, the solution $\{K, \bar{V}_1(t), \bar{V}_2(t)\}$ of (3.20) is also an upper solution of (1.1), (1.2), $(1.3)_{\text{II}}$. This implies

$$\lim_{t\to\infty}V_i(t,x)=0,\qquad i=1,2.$$

Let $\underline{S}_0=\inf_{\bar{\Omega}}S_0(x)>0$ and define $\underline{S}(t)$ as follows

$$\begin{aligned} &\frac{d\underline{S}(t)}{dt}=\gamma\underline{S}-\frac{\gamma}{K}\underline{S}^2-\frac{M_1}{y_1}\frac{\bar{V}_1(t)\underline{S}}{a_1+\underline{S}}-\frac{M_2}{y_2}\frac{\bar{V}_2(t)\underline{S}}{a_2+\underline{S}},\\ &\underline{S}(0)=\underline{S}_0 \end{aligned}\tag{3.21}$$

where $\bar{V}_1(t)$ and $\bar{V}_2(t)$ are determined by (3.20).

To show the first relation in (3.19) we recall the following definition and a theorem of Markus [9].

DEFINITION (Markus). Let A: $x_i'=f_i(x,t)$ and A_∞: $x_i'=f_i(x)$ $(i=1,2,\cdots,n)$ be a first order system of ordinary differential equations. The real-valued functions $f_i(x,t)$ and $f_i(x)$ are continuous in (x,t) for $x\in G$, where G is an open subset of $\mathbb{R}^n$, and for $t>t_0$ and they satisfy a local Lipschitz condition in x. A is said to be asymptotic to $A_\infty(A\to A_\infty)$ in G if for each compact set $K\subset G$ and for each $\varepsilon>0$, there is a $T=T(k,\varepsilon)>t_0$ such that $|f_i(x,t)-f_i(x)|<\varepsilon$ for all $i=1,2,\cdots,n$, all $x\in K$, and all $t>T$.

THEOREM (Markus). *Let* $A\to A_\infty$ *in* G *and let* p *be an asymptotically stable critical point of* A_∞. *Then there is a neighborhood* N *of* p *and a time* T *such that the omega limit set for every solution* $x(t)$ *of* A *which intersects* N *at a time later than* T *is equal to* p.

Now let us turn to (3.21). Since $\bar{V}_i(t)\to 0$ as $t\to\infty$, $(i=1,2)$ and $S\equiv K$ is an asymptotically stable solution of

$$\frac{dS}{dt}=\gamma S-\frac{\gamma}{K}S^2, \tag{3.22}$$

regarding the equations in (3.21) and (3.22) as A and A_∞ in Markus theorem respectively, it follows that $\underline{S}(t)\to K$ as $t\to+\infty$. On the other hand, it can be shown that $(\underline{S}(t), 0, 0)$ is a lower solution, which, together with the upper solution $(K, \bar{V}_1(t), \bar{V}_2(t))$, leads to

$$\lim_{t\to\infty} S(t, x)=K.$$

Now we consider the general case without the restriction $\hat{S}\leqq K$.

THEOREM 3.4. *Let* $M_iK/(a_i+K)<D_i$ $(i=1,2)$ *and* $\inf_{\bar{\Omega}} s_0(x)>0$. *Then the solution of* (S, V_1, V_2) *to the Neumann problem* (1.1), (1.2) *and* $(1.3)_{\text{II}}$ *satisfies* $\lim_{t\to\infty} S(x, t)=K$, $\lim_{t\to\infty} V_i(x, t)=0$.

Proof. Take a lower solution $(\underline{S}(t), 0, 0)$ and the upper solution $(\bar{S}(t), V_1(t), V_2(t))$ as follows:

$$\tag{3.23}\begin{aligned}
&\frac{d\bar{S}}{dt}=\gamma\bar{S}-\frac{\gamma}{K}\bar{S}^2, && \bar{S}(0)=\hat{S},\\
&\frac{d\underline{S}}{dt}=\gamma\underline{S}-\frac{\gamma}{K}\underline{S}^2-\frac{M_1}{y_1}\frac{V_1\underline{S}}{(a_1+\underline{S})}-\frac{M_2}{y_2}\frac{V_2\underline{S}}{(a_2+\underline{S})}, && \underline{S}(0)=\underline{S}_0,\\
&\frac{dV_1}{dt}=\left(\frac{M_2\bar{S}}{a_1+\bar{S}}-D_1\right)V_1, && V_1(0)=\hat{V}_1,\\
&\frac{dV_2}{dt}=\left(\frac{M_2\bar{S}}{a_2+\bar{S}}-D_2\right)V_2, && V_2(0)=\hat{V}_2
\end{aligned}$$

where the definitions of $\hat{s}$, $\underline{s}_0$, $\hat{V}_i$ are the same as before. Namely

$$\hat{S}=\sup_{\bar{\Omega}} S_0(x)<+\infty,\quad \underline{S}_0=\inf_{\bar{\Omega}} S_0(x)>0,\quad \hat{V}_i=\sup_{\bar{\Omega}} V_{i0}(x)>0,\quad i=1,2.$$

Since $M_iK/(a_i+K)<D_i$, there exists $\varepsilon>0$ such that

$$\frac{M_i(K+\varepsilon)}{a_i+(K+\varepsilon)}<D_i,\qquad i=1,2. \tag{3.24}$$

On the other hand, in view of the property of the solution of $(3.23)_1$, there exists $T_0>0$ such that

$$\bar{S}(t)<k+\varepsilon\quad\text{for } t>T_0. \tag{3.25}$$

Denote

$$\exp\left\{\int_0^{T_0}\left[\frac{M_i\bar{S}(\eta)}{a_i+\bar{S}(\eta)}-D_i\right]d\eta\right\}\text{ by } A(T_0),$$

it follows by (3.24) and (3.25) that

$$V_i(t)\leqq V_i(0)A(T_0)\exp\left[\frac{M_i(K+\varepsilon)}{a_i+(K+\varepsilon)}-D_i\right](t-T_0)\quad\text{for } t>T_0.$$

Hence, $V_i(t)\to 0$ as $t\to+\infty$. Then, the theorem of Markus leads to $\lim_{t\to\infty}\underline{S}(t)=k$. By Theorem 3.1, it is known that

$$0\leqq V_i(t, x)\leqq V_i(t),\qquad \underline{S}(t)\leqq S(t, x)\leqq\bar{S}(t).$$

Therefore

$$\lim_{t\to+\infty} S(t,x)=K, \quad \lim_{t\to+\infty} V_i(t,x)=0 \quad (i=1,2) \text{ uniformly for } x\in\Omega.$$

Next, we consider the case when one of the assumptions $M_iK/(a_i+K)<D_i$ does not hold, namely, either

$$\frac{M_1K}{a_1+K}\geqq D_1, \quad \frac{M_2K}{a_2+K}<D_2 \tag{3.26}$$

or

$$\frac{M_1K}{a_1+K}<D_1, \quad \frac{M_2K}{a_2+K}\geqq D_2. \tag{3.27}$$

For (3.26), the same argument used in the proof of Theorem 3.4 shows that $V_2(x; t)\to 0$ as $t\to+\infty$. Then it is natural to investigate the stability of $(\lambda_1, V_1^*, 0)$, which is a solution of (1.1), $(1.3)_{\text{II}}$. Where $\lambda_1=a_1D_1/(M_1-D_1)>0$ (this is guaranteed by $M_1K/(a_1+K)\geqq D_1$), $V_1^*=(y_1\gamma/M_1)(1-\lambda_1/K)(a_1+\lambda_1)\geqq 0$, which becomes equality if and only if $K=\lambda_1$, i.e. $M_1K/(a_1+K)=D_1$.

To analyze the stability of $(\lambda_1, V_1^*, 0)$, we recall linearized stability.

Consider the Neumann type initial-boundary problem

$$\begin{aligned} &\frac{\partial u}{\partial t}=DAu+F(u) && \text{in } R^+\times\Omega,\\ &\frac{\partial u}{\partial n}=0 && \text{on } R^+\times\partial\Omega,\\ &u(0,x)=u_0(x), && x\in\Omega. \end{aligned} \tag{3.28}$$

Suppose $u=\varphi(x)$ is a stationary solution of (3.28), namely,

$$\begin{aligned} &-D\Delta\varphi(x)=F(\varphi) \quad \text{in } \Omega,\\ &\left.\frac{\partial\varphi}{\partial n}\right|_{\partial\Omega}=0. \end{aligned} \tag{3.29}$$

Let S denote the differential operator obtained by formally linearizing the right-hand side of (3.28) about the given solution φ:

$$S(\bar u)\equiv D\Delta\bar u+\frac{\partial F}{\partial u}(\varphi(x))\bar u$$

where $\partial F/\partial u$ denotes the Jacobian matrix.

We consider S as an operator on the space $C^0(\bar\Omega)$ of bounded vector functions, continuous on $\bar\Omega$, with domain consisting of functions in $C^2(\bar\Omega)$ and satisfying the given Neumann boundary conditions. Let $\sigma(S)$ denote the spectrum of S.

Definition 3.2. φ is stable according to the linearized criterion (l.c.) or called linearly stable if $\sigma(S)$ is in the negative half-plane and is bounded away from the imaginary axis. Namely, there exists a negative number $\alpha<0$ such that $\operatorname{Re} z\leqq\alpha$ for any $z\in\sigma(S)$. φ is marginally stable (l.c.) if $\operatorname{Re} z\leqq 0$ for any $z\in\sigma(S)$ and there exists a z_0 such that $\operatorname{Re} z_0=0$. φ is unstable (l.c.) if $\sigma(S)$ contains a point in the right open half-plane.

DEFINITION 3.3. Let $\bar{u}$ be a solution of (3.29). Then $\bar{u}$ is stable if there is a neighborhood N of $\bar{u}$ and positive numbers C, α such that if u is a solution of (3.28) with $u(0)\in N$, then u exists for all $t>0$ and

$$\|u(t)-\bar{u}\|\leqq Ce^{-\alpha t}\|u(0,x)-\bar{u}\|, \qquad t>0$$

where $\|\cdot\|$ denotes the usual L_2-norm on $C(\bar{\Omega})$.

It is known that (cf. Smoller [12]) if $\bar{u}$ is a linearly stable solution of (3.28), then $\bar{u}$ is stable.

Now we analyze the stability of $(\lambda_1, V_1^*, 0)$. Let T be the linearization of the right-hand side of (1.1) about $(\lambda_1, V_1^*, 0)$.

$$T\begin{pmatrix} S \\ V_1 \\ V_2 \end{pmatrix}=\begin{bmatrix} k_1\Delta+\gamma-\frac{\gamma}{K}2\lambda_1-\frac{M_1a_1V_1^*}{y_1(a_1+\lambda_1)^2} & -\frac{D_1}{y_1} & -\frac{M_2\lambda_1}{y_2(a_2+\lambda_1)} \\ \frac{M_1a_1V_1^*}{(a_1+\lambda_1)^2} & k_2\Delta & 0 \\ 0 & 0 & k_3\Delta-D_2+\frac{M_2\lambda_1}{a_2+\lambda_1} \end{bmatrix}\begin{pmatrix} S \\ V_1 \\ V_2 \end{pmatrix}.$$

We need to investigate the spectrum $\sigma(T)$. Let $\sigma(\Delta)$ denote the spectrum of the Laplace operator as acting on $C^0(\bar{\Omega})$ with domain consisting of functions in $C^2(\bar{\Omega})$ and satisfying no-flux boundary conditions.

It is well known that $0\in\sigma(\Delta)$ and $\sigma(\Delta)$ consists of an infinite but discrete set of simple real eigenvalues, bounded from above, namely, $\cdots<\tau_n<\cdots<\tau_2<\tau_1<\tau_0=0$. Define

$$C=\frac{M_1a_1V_1^*}{(a_1+\lambda_1)^2}, \qquad E=\gamma-\frac{\gamma}{K}2\lambda_1-\frac{C}{y_1},$$

$$P(\xi,\mu)=P_1(\xi,\mu)\cdot P_2(\xi,\mu),$$

$$P_1(\xi,\mu)=\left(k_3\mu-D_2+\frac{M_2\lambda_1}{a_2+\lambda_1}-\xi\right),$$

$$P_2(\xi,\mu)=\xi^2-(k_1\mu+k_2\mu+E)\xi+\frac{D_1}{y_1}C+(k_1\mu+E)k_2\mu,$$

and

$$\Lambda=\{\eta\colon P(\eta,\mu)=0 \text{ for some } \mu\in\sigma(\Delta)\}.$$

Our first result is that

$$\sigma(T)\subset\Lambda. \tag{3.30}$$

To prove this, suppose that $\eta\notin\Lambda$. Then $P(\eta,\mu)\neq 0$ for all $\mu\in\sigma(\Delta)$, so $0\notin p(\eta,\sigma(\Delta))$. Now consider the sixth order differential operator $p(\eta,\Delta)$. It is known that for such polynomials p, $\sigma(p(\Delta))=p(\sigma(\Delta))$ [2], so that in our case we deduce $0\notin\sigma(p(\eta,\Delta))$. Therefore $p(\eta,\Delta)^{-1}$ exists as a bounded operator on $C^0(\bar{\Omega})$, which implies that both $p_1(\eta,\Delta)^{-1}$ and $p_2(\eta,\Delta)^{-1}$ exist as bounded operators on $C^0(\bar{\Omega})$. For any given f, g, $h\in C^0(\bar{\Omega})$, we can solve the equations

$$(T-\eta I)\begin{pmatrix} S \\ V_1 \\ V_2 \end{pmatrix}=\begin{pmatrix} f \\ g \\ h \end{pmatrix}$$

explicitly as follows:

$$S=\frac{D_1}{y_1}p_2(\eta,\Delta)^{-1}g+(k_2\Delta-\eta)p_2(\eta,\Delta)^{-1}\left[\frac{M_2\lambda_1}{y_2(a_2+\lambda_1)}p_1(\eta,\Delta)^{-1}h+f\right],$$

$$V_1=(k_1\Delta+E-\eta)p_2(\eta,\Delta)^{-1}g-Cp_2(\eta,\Delta)^{-1}\left[\frac{M_2\lambda_1}{y_2(a_2+\lambda_1)}p_1(\eta,\Delta)^{-1}h+f\right],$$

$$V_2=p_1(\eta,\Delta)^{-1}h.$$

This shows that $\eta\notin\sigma(T)$, and this conclusion in turn proves (3.30).

Our problem has been reduced to analyzing the property of Λ. Suppose $\eta\in\Lambda$; then there exists $\mu\in\sigma(\Delta)$ such that

$$\left(k_3\mu-D_2-\frac{M_2\lambda_1}{a_2+\lambda_1}-\eta\right)\left[\eta^2-(k_1\mu+k_2\mu+E)\eta+\frac{D_1}{y_1}C+(k_1\mu+E)k_2\mu\right]=0.$$

Let η_i $(i=1,2,3)$ denote the roots of the above equation. Clearly,

$$\eta_1=k_3\mu-D_2+\frac{M_2\lambda_1}{a_2+\lambda_1},$$

$$\eta_{2,3}=\frac{k_1\mu+k_2\mu+E}{2}\mp\frac{1}{2}\left[(k_1\mu+k_2\mu+E)^2-4\left(\frac{D_1C}{y_1}+(k_1\mu+E)k_2\mu\right)\right]^{1/2}.$$

Due to $M_1K/(a_1+K)\geqq D_1$, $M_2K/(a_2+K)<D_2$, it can be shown that $-D_2+((M_2\lambda_1)/(a_2+\lambda_1))<0$, which implies, since $\mu\leqq 0$, that $\eta_1<0$.

Assume $K<a_1+2\lambda_1$, which shows $E<0$; then it turns out that Re $\eta_{2,3}<0$. Thus, it ends up that $(\lambda_1, V_1^*, 0)$ is a linearly stable solution of (1.1)-(1.3)$_{\rm II}$. We obtain the asymptotic property of the solution (s, v_1, v_2) in this case as follows.

THEOREM 3.5. *Suppose* $M_1K/(a_1+K)\geqq D_1$, $M_2K/(a_2+K)<D_2$ *and* $K<a_1+2\lambda_1$ (*namely,* $a_1D_1/(M_1-D_1)\leqq K<(a_1D_1+a_1M_1)/(M_1-D_1)$). *Then there exists* $\varepsilon>0$ *such that*

$$\lim_{t\to\infty}\|S(t,x)-\lambda_1\|=0,\qquad \lim_{t\to\infty}\|V_1(t,x)-V_1^*\|=0,$$

if $\|S_0(x)-\lambda_1\|<\varepsilon$, $\|V_{10}(x)-V_1^*\|<\varepsilon$. *Moreover,* $\lim_{t\to\infty}V_2(t,x)=0$, *uniformly for* $x\in\Omega$.

Similarly, we obtain the following result for case (3.27).

THEOREM 3.6. *Suppose* $M_1K/(a_1+K)<D_1$, $M_2K/(a_2+K)\geqq D_2$ *and* $K<a_2+2\lambda_2$ (*namely,* $a_2D_2/(M_2-D_2)\leqq K<(a_2D_2+a_2M_2)/(M_2-D_2)$). *Then there exists* $\varepsilon>0$ *such that*

$$\lim_{t\to\infty}\|S(t,x)-\lambda_2\|=0,\qquad \lim_{t\to\infty}\|V_2(t,x)-V_2^*\|=0$$

if $\|S_0(x)-\lambda_2\|<\varepsilon$, $\|V_{20}(x)-V_2^*\|<\varepsilon$. *Moreover,* $\lim_{t\to\infty}V_1(t,x)=0$, *uniformly for* $x\in\Omega$, *where* $(\lambda_2, 0, V_2^*)$ *is a solution of* (1.1), (1.3)$_{\rm II}$, $\lambda_2=a_2D_2/(M_2-D_2)>0$, $V_2^*=(y_2\gamma/M_2)(1-\lambda_2/K)(a_2+\lambda_2)\geqq 0$ *which becomes equality if and only if* $K=\lambda_2$, *i.e.,* $M_2K/(a_2+K)=D_2$.

Finally, we consider the case when both of the assumptions $M_iK/(a_i+K)<D_i$ do not hold, namely,

$$\frac{M_iK}{a_i+K}\geqq D_i,\qquad i=1,2. \tag{3.31}$$

Clearly, (3.31) is equivalent to $0<\lambda_i\leqq K$, $i=1,2$. By the same argument as used before, we obtain the following results.

THEOREM 3.7. *Suppose* $0<\lambda_1<\lambda_2<K<a_1+2\lambda_1$; *then there exists* $\varepsilon>0$ *such that*

$$\lim_{t\to\infty}\|S(t,x)-\lambda_1\|=0,\quad \lim_{t\to\infty}\|V_1(t,x)-V_1^*\|=0,\quad \lim_{t\to\infty}\|V_2(t,x)\|=0$$

if $\|S_0(x)-\lambda_1\|<\varepsilon$, $\|V_{10}(x)-V_1^*\|<\varepsilon$, $\|V_{20}(x)\|<\varepsilon$.

The last theorem is similar.

THEOREM 3.8. *Suppose* $0<\lambda_2<\lambda_1<K<a_2+2\lambda_2$; *then there exists* $\varepsilon>0$ *such that*

$$\lim_{t\to\infty}\|S(t,x)-\lambda_2\|=0,\quad \lim_{t\to\infty}\|V_1(t,x)\|=0,\quad \lim_{t\to\infty}\|V_2(t,x)-V_2^*\|=0$$

if $\|S_0(x)-\lambda_2\|<\varepsilon$, $\|V_{10}(x)\|<\varepsilon$, $\|V_{20}(x)-V_2^*\|<\varepsilon$, *where* λ_1, V_1^* *and* λ_2, V_2^* *are the same as in Theorems* 3.5 *and* 3.6, *respectively.*

REFERENCES

[1] G. J. BUTLER AND P. WALTMAN, *Bifurcation from a limit cycle in a two predator-one prey ecosystem modeled on a chemestat*, J. Math. Biol., 12 (1981), pp. 295-310.

[2] P. FIFE, *Mathematical aspects of reacting and diffusing systems*, Lecture Notes in Biomathematics 28, Springer, Berlin-New York, 1979.

[3] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[4] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 840, Springer, Berlin-New York, 1981.

[5] C. S. HOLLING, *The functional response of predators to prey density and its role in mimicry and population regulation*, Mem. Entomol. Soc. Canada, 45 (1965), pp. 5-60.

[6] ———, *The functional response of invertebrate predators to prey density*, Mem. Entomol. Soc. Canada, 48 (1966), pp. 1-85.

[7] S. B. HSU, S. P. HUBBELL AND O. WALTMAN, *A contribution to the theory of competing predators*, Ecological Monographs, 48 (1978), pp. 337-349.

[8] ———, *Competing predators*, SIAM J. Appl. Math., 35 (1978), pp. 617-625.

[9] L. MARKUS, *Asymptotically autonomous differential systems*, in Contributions to the Theory of Nonlinear Oscillation, Vol. 3, Princeton Univ. Press, Princeton, NJ, 1956, pp. 17-29.

[10] P. DE MOTTONI AND A. TESEI, *Asymptotic stability results for a system of quasilinear parabolic equations*, Applicable Anal., 9 (1979), pp. 7-21.

[11] C. V. PAO, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165-198.

[12] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[13] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton Univ. Press, Princeton, NJ, 1970.

[14] F. TREVES, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

007

PERSISTENCE AND SMOOTHNESS OF HYPERBOLIC INVARIANT MANIFOLDS FOR FUNCTIONAL DIFFERENTIAL EQUATIONS*

LUIS T. MAGALHÃES†

Abstract. The persistence and smoothness of hyperbolic invariant manifolds is established for small perturbations of retarded functional differential equations. Properties of exponential dichotomies and of spectra of invariant manifolds, which are established here for semiflows, form the basis for proving the results. The analysis uses a moving system of coordinates, around the original hyperbolic invariant manifold, consisting of coordinates along the tangential, unstable and stable manifolds of the linearized equation along solutions of the unperturbed equation which lie on the original hyperbolic manifold.

Key words. invariant sets, hyperbolicity, functional differential equations

AMS(MOS) subject classifications. 34K05, 34K15, 58F10

1. Introduction. The study of invariant sets is of fundamental importance in the geometric theory of dynamical systems. Since, in general, these sets can have a wild topological structure, it is of interest to identify situations where they have the particularly simple structure of differentiable manifolds. Especially important in this context are those manifolds that persist under small perturbations, a situation that leads naturally to the study of hyperbolic invariant manifolds.

There is an extensive literature on hyperbolic invariant manifolds for ordinary differential equations, cf. [1]-[3], [6], [8], [9], [13], [14]. In the context of infinite-dimensional dynamical systems, the study of hyperbolic invariant manifolds was pursued for certain parabolic partial differential equations [7] and for particular cases of functional differential equations [10]-[12].

The studies of Kurzweil [10], [11] on hyperbolic invariant manifolds for functional differential equations (FDE) rely on establishing fixed points for maps which correspond to discrete dynamical systems obtained by discretization of the semiflow induced by the equations. The approach in this paper is, in contrast, based directly on the semiflow and uses techniques of exponential dichotomies for obtaining bounded solutions, following the work of Hale, cf. [3], for hyperbolic invariant manifolds of ordinary differential equations (ODE). It is believed that this approach is simpler than methods based on discretization.

The paper begins by establishing the concept of exponential dichotomies in the context of skew-product semiflows in Banach vector bundles, and the concept of hyperbolicity using the notion of spectrum of an invariant manifold, developed by Sacker and Sell for flows [15]-[19]. Here, the numerous discussions with George Sell, while this work was in progress, played an important role, in particular in connection to work of his in collaboration with Sacker [18]. A short reference is then made to FDEs on manifolds, based on the work of Oliva, cf. [5], referring the reduction of the general situation to the case of FDEs on euclidean space R^q. These are basically the contents of §§ 2, 3 and 4.

* Received by the editors July 29, 1985; accepted for publication March 7, 1986.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455 and Departamento de Matematica, Instituto de Superior Técnico, Universidade Técnica de Lisboa, 1096 Lisboa Codex, Portugal. This research was partially supported by the Institute for Mathematics and Its Applications (IMA) at the University of Minnesota, by the Centro de Matematica e Aplicações Fundamentais (CMAF) of the Instituto Nacional de Investigação Cientifica (INIC) at Lisbon, Portugal, and by the Fundação Calouste Gulbenkian at Lisbon, Portugal.

The main difficulty in establishing the persistence and properties of hyperbolic manifolds for FDEs is related to the impossibility of extending backwards in time all the solutions. It is useful to linearize the equation around the original invariant manifold and to consider the corresponding tangential, unstable and stable manifolds, which are themselves invariant under the linearized semiflow. As it turns out, one only needs to extend the solutions backwards in time along the tangential and unstable directions. That involves no difficulties because both the tangential and unstable manifolds are finite-dimensional and, therefore, the semiflow for the linearized equation can be extended on them to a flow of an ODE. This leaves out only the evolution on the stable manifold to be considered as an FDE, but we do not need to extend the solutions backwards there. These ideas are pursued in § 5 by the introduction of a moving system of coordinates centered at solutions of the unperturbed equation lying on the original invariant manifold. The perturbed equation in these coordinates is given by a family of systems of two perturbed ODEs describing the evolution of the tangential and unstable variables, and an FDE for the stable variable. These systems are written as perturbations of the linear variational equation around the given hyperbolic manifold, using the variation of constants formula introduced by Hale for FDEs, cf. [4]. The analysis is actually based on the existence of exponential dichotomies for the linear variational equation which split the solutions with initial conditions in the tangential, unstable, or stable spaces.

In § 6, we study the persistence and the smoothness of integral manifolds for FDEs of the form obtained by application of the system of coordinates around the original invariant manifold, and we also study the structure of the semiflow around the integral manifolds, under general assumptions of global Lipschitzian nonlinearities with sufficiently small Lipschitz constants. The persistence of the integral manifolds is established using techniques of exponential dichotomies for the determination of bounded solutions, following an approach introduced by Hale, cf. [3], and also applied by Sell [19] in the context of ODEs. The smoothness properties of the integral manifolds are established by a modification of the method developed by Fenichel [2] also in the context of ODEs.

Finally, the results obtained for systems in coordinate form are applied, in § 7, to prove the persistence and smoothness of hyperbolic invariant manifolds for FDEs, and to study the local geometric structure of the orbits around them. Using "cut off" functions around the original invariant manifold, in a similar way as it is usually done for center manifold theory, one can get a system that agrees with the original system inside a neighborhood of the invariant manifold, and whose nonlinear terms satisfy the global Lipschitz conditions assumed in § 6. Thus, finding an invariant manifold for the perturbed equation, in such a neighborhood of the original manifold, amounts to finding an invariant manifold for the auxiliary system which is obtained through the application of the "cut off" functions mentioned above. The results obtained for systems in coordinate form are only good to get patches of the invariant manifolds which occur under perturbations. These pieces have to be patched up in order to obtain the perturbed manifold, but this can be easily done using the uniqueness properties established in § 6 for the integral manifolds of systems in coordinate form.

The system of coordinates which is introduced in this paper, around the original hyperbolic invariant manifold, is redundant, in the sense that each point close to the manifold can be represented by infinitely many combinations of the coordinates. In fact, the system of coordinates is centered at points of the invariant manifold, which therefore account for some of the coordinates used, and also involves coordinates along the tangential, unstable and stable spaces of the linearized equation around the manifold. The redundancy comes from the fact that the coordinates giving the point on the manifold where the coordinate frame is centered and the tangential coordinates eventually play against each other when describing neighborhoods of a particular point on the manifold. This situation is not different from many other cases where moving systems of coordinates have been used with the intent of simplifying the analysis, as for instance in certain situations describing the movement of bodies in the context of Newtonian mechanics. It is, of course, true that a nonredundant system of coordinates could be introduced leaving out the tangential components. However, the use of the redundant coordinates described above simplifies the analysis for two reasons: (i) it allows the separation of the dynamics of the linearized equations around the original manifold from the dynamics on the manifold itself, in particular the tangent bundle is invariant under the linearized equation, and (ii) it avoids the introduction of abstract equations to describe the evolution of the different coordinates, because, in this way, they can be given by FDEs in euclidean space.

Some of the ideas introduced in this paper are new and useful even in the case of ODEs. In particular, the use of the redundant system of coordinates described above, although most useful in the case of FDEs, is also natural in the context of ODEs since it simplifies the description of the linearized equation around the original invariant manifold and considerably simplifies the analysis.

2. Spectrum of a linear skew-product semiflow. Let W be a topological space. A *flow* on W is a continuous mapping $\pi: R\times W\to W$ such that $\pi(0,w)=w$ and $\pi(t,\pi(s,w))=\pi(t+s,w)$ for all $w\in W$ and $t,s\in R$. A *semiflow* on W is a continuous mapping $\pi:[0,\infty)\times W\to W$ satisfying the preceding conditions for $t,s\geqq 0$.

Let X be a smooth Banach manifold without boundary and let E be a Banach vector bundle over X with fiber projection $p: E\to X$, i.e., E is a vector bundle over X with each fiber $E(x)=p^{-1}(x)$, $x\in X$, being a Banach space. Points in E can be represented by ordered pairs (x,z) with $x\in X$, and z a vector in the fiber $E(x)$. A semiflow π on E is said to be a *skew-product semiflow* on E if there is a flow ϕ on X such that the fiber projection p commutes with π and ϕ, i.e., π can be represented as

$$\pi(t,x,z)=(\phi(t,x),\psi(t,x,z)),\qquad t\geqq 0$$

and $\psi(t,x,z)$ is in the fiber $E(\phi(t,x))$. Such a skew-product semiflow π is a *linear skew-product semiflow* if the mapping $z\to\Psi(t,x)z=\psi(t,x,z)$ is a linear mapping from the fiber $E(x)$ to the fiber $E(\phi(t,x))$. One defines analogously *skew-product flow* and *linear skew-product flow*. When π is a semiflow on E, $\pi(t,x,z)$, $\psi(t,x,z)$ and $\Psi(t,x)$ are only defined for $t\geqq 0$. However, they can be extended for $t\leqq 0$ at those points (x,z) through which there is a backwards continuation defined for all $t\leqq 0$. Let us define the set B by

$B=\{(x,z)\in E$: there is exactly one continuous function $(u,v):(-\infty,0]\to E$
such that $u(0)=x$, $v(0)=z$ and $\pi(t,u(s),v(s))=(u(t+s),v(t+s))$
for all $s\leqq 0$ and all $t\in[0,-s]\}$.

The set

$$S=\{(x,z)\in E: |\psi(t,x,z)|\to 0 \text{ as } t\to+\infty\}$$

is called the *stable set* of X under π, and the set

$$U=\{(x,z)\in B: |\psi(t,x,z)|\to 0 \text{ as } t\to-\infty, \text{for the continuation of } \psi(t,x,z) \text{ for } t\leqq 0\}$$

is called the *unstable set* of X under π. A set $I\subset E$ is said to be *positively invariant* under π if $\pi(t,x,z)\in I$ for all $t\geqq 0$ and $(x,z)\in I$, and I is said to be *invariant* under

π if $I \subset B$ and the preceding condition holds for all $t \in R$, for the continuation of $\pi(t, x, z)$. The sets S and U are both positively invariant under π, and the sets B, $S \cap B$ and U are all invariant under π. It is easy to see thai these sets are vector subbundles of E.

The linear skew-product semiflow $\pi = (\phi, \psi)$ on E is said to admit an *exponential dichotomy* on X if there exist linear projections $P(x)$ defined on $E(x)$ and depending continuously on $x \in X$, and there exist constants $K, \alpha > 0$ such that

(i) $N(P) = \{(x, z) \in E: P(x)z = 0\} \subset B$,

(ii) $|\Psi(t, x)P(x)| \leqq K e^{-\alpha t}, \quad t \geqq 0, \quad x \in X$,

(iii) $|\Psi(t, x)[I - P(x)]| \leqq K e^{+\alpha t}, \quad t \leqq 0$,
for the continuation of $\Psi(t, x)z$ for $t \leqq 0$ with $(x, z) \in B$.

We note that condition (iii) makes sense because (i) implies that the range of the mapping $(x, z) \to (x, [I - P(x)]z)$ is contained in B. Whenever π admits an exponential dichotomy on X we have $U = N(P) = \{(x, z) \in E: P(x)z = 0\}$ and $S = R(P) = \{(x, z) \in E: z = P(x)z'$ for some $z' \in E(x)\}$. Then, the stable and unstable sets are complementary subbundles of E.

Given a linear skew-product semiflow $\pi = (\phi, \psi)$ on a vector bundle E, and a real number λ, we define a mapping π_λ by

$$\pi_\lambda(t, x, z) = (\phi(t, x), e^{-\lambda t}\psi(t, x, z)).$$

It is easily seen that π_λ is also a linear skew-product semiflow on E and that the invariant sets under π and under π_λ agree for all $\lambda \in R$. We also define

$$\Psi_\lambda(t, x)z = e^{-\lambda t}\psi(t, x, z) = e^{-\lambda t}\Psi(t, x)z.$$

The stable and the unstable sets of X under π_λ are denoted by S_λ and U_λ, respectively. Clearly, if $\mu < \lambda$ then $S_\mu \subset S_\lambda$ and $U_\mu \supset U_\lambda$. The set of all $\lambda \in R$ for which π_λ admits an exponential dichotomy on X is called the *resolvent set* of π on E and is denoted by $\rho(E, \pi)$. The complement of the resolvent set on R is called the *spectrum* of π on E and is denoted by $\Sigma(E, \pi)$.

A skew-product semiflow $\pi = (\phi, \psi)$ on the vector bundle E is said to be *uniformly completely continuous* if for each $x \in X$ there is a neighborhood V_x of x in X and a real number $r_x \geqq 0$ such that, for all $t \geqq r_x$, the mapping $(y, z) \to \pi(t, y, z)$ maps bounded subsets of $E_{V_x} = \{(y, z) \in E: y \in V_x\}$ into relatively compact subsets of E.

The following theorem contains the properties of the spectrum $\Sigma(E, \pi)$ which are used in this paper.

THEOREM 2.1. *Let $\pi = (\phi, \psi)$ be a uniformly completely continuous linear skew-product semiflow on a Banach vector bundle E defined over a compact, connected, smooth Banach manifold X. Then the spectrum $\Sigma(E, \pi)$ is a closed set of real numbers which is bounded above, and, consequently, it is a union of closed intervals, the spectral intervals (an interval $[a, b]$ is allowed to degenerate to a point when $a = b$).*

Associated with each spectral interval there is a spectral subbundle V of E, which satisfies the following properties:

(1) *If $\mu, \lambda \in \rho(E, \pi)$ and $(\mu, \lambda) \cap \Sigma(E, \pi) = [a, b]$, then the spectral subbundle V associated with $[a, b]$ has finite dimension, satisfies $V = U_\mu \cap S_\lambda$, and is invariant under π;*

(2) *If $\lambda \in \rho(E, \pi)$ and $(-\infty, \lambda) \cap \Sigma(E, \pi) = (-\infty, b]$, then the spectral subbundle V associated with $(-\infty, b]$ satisfies $V = S_\lambda$ and is positively invariant under π.*

Moreover, if $\lambda \in \rho(E, \pi)$, then the number of spectral intervals included in $(\lambda, +\infty)$ is finite.

Proof. Let $\lambda \in \rho(E, \pi)$. Then π_λ admits an exponential dichotomy on X. There exist a linear projection $P_\lambda(x)$ on $E(x)$ and constants $K, \alpha > 0$ such that $N(P_\lambda) \subset B$, and

$$|\Psi_\lambda(t, x)P_\lambda(x)| \leqq K e^{-\alpha t}, \qquad t \geqq 0, \quad x \in X,$$

$$|\Psi_\lambda(t, x)[I - P_\lambda(x)]| \leqq K e^{+\alpha t}, \qquad t \leqq 0, \quad x \in X,$$

where the last inequality holds for some backwards extension of $\Psi_\lambda(t, x)$. If μ satisfies $|\lambda - \mu| < \beta = \alpha/2$, then

$$|\Psi_\mu(t, x)P_\lambda(x)| \leqq K e^{-\beta t}, \qquad t \geqq 0, \quad x \in X,$$

$$|\Psi_\mu(t, x)[I - P_\lambda(x)]| \leqq K e^{+\beta t}, \qquad t \leqq 0, \quad x \in X.$$

Therefore, with $P_\mu = P_\lambda$, π_μ admits an exponential dichotomy on X. It follows that $\mu \in \rho(E, \pi)$, $U_\mu = U_\lambda$ and $S_\mu = S_\lambda$, for all μ such that $|\lambda - \mu| < \alpha/2$. This implies that $\rho(E, \pi)$ and $\{\mu \in R\colon U_\mu = U_\lambda, S_\mu = S_\lambda\}$, for any $\lambda \in \rho(E, \pi)$, are open sets. Thus $\Sigma(E, \pi)$ is a closed set.

The fact that $\Sigma(E, \pi)$ is bounded above results from the compactness of X and the semigroup property of π. In fact, let $k = \sup\{|\Psi(t, x)|\colon x \in X \text{ and } 0 \leqq t \leqq 1\}$. Since Ψ is continuous and the supremum is taken over a compact set, the constant k is finite. Fix $t \geqq 0$ and let m be the largest integer smaller or equal to t. Then, with $x_i = \phi(i, x)$, we have

$$\Psi(t, x) = \Psi(t - m, x_m)\Psi(1, x_{m-1}) \cdots \Psi(1, x)$$

and, consequently,

$$|\Psi(t, x)| \leqq k^{m+1} \leqq kk^t = k\, e^{at}, \qquad t \geqq 0,$$

where $a = \log k$. Therefore, π_λ admits an exponential dichotomy on X for $\lambda > a$, with $\alpha = \lambda - a$ and $P_\lambda(x)$ equal to the identity on $E(x)$. This proves that $(a, +\infty) \subset \rho(E, \pi)$ and $\Sigma(E, \pi) \subset (-\infty, a]$.

Since $\Sigma(E, \pi)$ is a closed set of real numbers which is bounded above, it is a union of compact intervals with, possibly an interval of the form $(-\infty, b]$. Let $\mu < \lambda$ be real numbers in $\rho(E, \pi)$ which separate one of the closed intervals that make up $\rho(E, \pi)$ from the others, i.e., the intersection of the interval (μ, λ) with $\rho(E, \pi)$ is precisely one spectral interval $[a, b]$. With this spectral interval we can associate the spectral bundle $V = U_\mu \cap S_\lambda$. In order to prove the properties (1) and (2) we need to show that V is independent of the points μ, λ, provided they satisfy the properties indicated. More precisely, we need to show that for $\mu, \lambda \in \rho(E, \pi)$ with $\mu < \lambda$ we have $U_\mu = U_\lambda$ and $S_\mu = S_\lambda$ if and only if $(\mu, \lambda) \subset \rho(E, \pi)$. Assume that $U_\mu \neq U_\lambda$ or $S_\mu \neq S_\lambda$, with $(\mu, \lambda) \subset \rho(E, \pi)$, and define $\bar{\sigma} = \sup\{\sigma \in \rho(E, \pi)\colon U_\sigma = U_\mu, S_\sigma = S_\mu\}$. Because the set $\{\mu \in R\colon U_\mu = U_\lambda, S_\mu = S_\lambda\}$, for any $\lambda \in \rho(E, \pi)$, is an open set, and because U_σ depends monotonically on σ, we get a contradiction. This shows that $U_\mu = U_\lambda$ and $S_\mu = S_\lambda$. To prove the converse, let $\mu, \lambda \in \rho(E, \pi)$ and assume that $U_\mu = U_\lambda$ and $S_\mu = S_\lambda$. Then π_μ and π_λ admit exponential dichotomies on X, with projections P_μ, P_λ and constants K_μ, K_λ and α_μ, α_λ, respectively. Let $K = \max\{K_\mu, K_\lambda\}$ and $\alpha = \min\{\alpha_\mu, \alpha_\lambda\}$. For either $\sigma = \mu$ or $\sigma = \lambda$, we have

$$|\Psi_\sigma(t, x)P_\sigma(x)| \leqq K e^{-\alpha t}, \qquad t \geqq 0, \quad x \in X,$$

$$|\Psi_\sigma(t, x)[I - P_\sigma(x)]| \leqq K e^{+\alpha t}, \qquad t \leqq 0, \quad x \in X.$$

Since $U_\mu = U_\lambda$ and $S_\mu = S_\lambda$, we have $P_\mu = P_\lambda = P$. Consequently

$$e^{-\sigma t}|\Psi(t,x)P(x)| \leqq K e^{-\alpha t}, \qquad t \geqq 0, \quad x \in X,$$
$$e^{-\sigma t}|\Psi(t,x)[I - P(x)]| \leqq K e^{+\alpha t}, \qquad t \leqq 0, \quad x \in X,$$

for σ equal to either μ or λ. This implies that these inequalities must also hold for all $\sigma \in [\mu, \lambda]$, proving that each one of the π_σ, for $\sigma \in [\mu, \lambda]$, admits an exponential dichotomy, and, therefore, $[\mu, \lambda] \subset \rho(E, \pi)$.

The invariance properties of the spectral subbundles follow from the positive invariance of the S_λ and the invariance of the U_λ and $S_\lambda \cap B$, for every $\lambda \in \rho(E, \pi)$.

It remains to prove that the spectral bundles associated with compact spectral intervals have finite dimension. For this we use the complete continuity of the semiflow. The semiflow π_λ is uniformly completely continuous for each $\lambda \in \rho(E, \pi)$. Thus, for each $x \in X$ there is a neighborhood V_x of x in X and a real number $r_x \geqq 0$ such that, for $t \geqq r_x$, the mapping $(x, z) \to \pi(t, x, z)$ maps bounded subsets of $E_{V_x} = \{(y, z) \in E: y \in V_x\}$ into relatively compact subsets of E. Since X is compact, we can extract a finite subcovering V_{x_i} of X and define $r = \max\{r_{x_i}\}$. Then, for all $x \in X$, $t \geqq r$, the mapping $z \to \Psi_\lambda(t, x)z$ maps bounded subsets of $E(x)$ into relatively compact subsets of $E(\phi(t, x))$. Since $\lambda \in \rho(E, \pi)$, π_λ admits an exponential dichotomy on X with projection P and constants $K, \alpha > 0$. Then $U_\lambda(x) = N(P(x))$ is a closed linear subspace of $E(x)$. Let us denote $S = \{z \in U_\lambda(x): |z| \leqq 1\}$. For each $z \in S$, the mapping $\Psi_\lambda(t, x)z$ has one backwards extension defined for all $t \leqq 0$ and such that

$$|\Psi_\lambda(t, x)z| \leqq K e^{+\alpha t}, \qquad t \leqq 0.$$

For each $z \in S$, define the set $S' = \{z' \in E(\phi(-r, z)): z' = \Psi_\lambda(-r, x)z \text{ with } z \in S\}$. For each $z' \in S'$ we have $|z'| \leqq K e^{-\alpha r}$ and, consequently, the set S' is bounded. Therefore, the mapping $z' \to \Psi_\lambda(r, \phi(-r, x))z'$ maps S' into a compact subset of $E(x)$. Since S is the image of S' under this map, it follows that S is a compact subset of the Banach space $U_\lambda(x)$. The only Banach spaces which have the closed unit ball compact, are the finite dimensional spaces. Consequently, $\dim U_\lambda(x) < \infty$. Clearly, if $\mu, \lambda \in \rho(E, \pi)$ then $V(x) = U_\mu(x) \cap S_\lambda(x)$ is also finite dimensional and, because X is connected, the dimensions of all fibers $V(x)$, for $x \in X$, are the same.

Finally, if $\lambda \in \rho(E, \pi)$, then $\dim U_\lambda$ is finite and the preceding properties of U_λ, S_λ, with the monotone dependence of S_λ on λ, imply that the union of the spectral subbundles associated with spectral intervals contained in the interval $(\lambda, +\infty)$ is equal to U_λ and, consequently, the number of such spectral intervals is finite. QED

The evolution on the spectral subbundles associated with compact spectral intervals can be given by ordinary differential equations (ODE).

PROPOSITION 2.2. *Let π be a skew-product semiflow on a Banach vector bundle E defined over a connected, smooth Banach manifold X. If V is a finite-dimensional subbundle of E which is invariant under π, then the restriction of π to V can be extended to a flow on V.*

Moreover, if the mapping $t \to \pi(t, x, z)$ is differentiable at $t = 0$, for every $(x, z) \in V$, and its derivative at $t = 0$ is locally Lipschitzian in $(x, z) \in V$, then the flow of π on V can be given by an ODE.

Proof. Since V is invariant under π, through every point of V there is one backwards continuation of π. Consequently, $\pi(t, x, z)$ is well defined and belongs to V, for all $t \in R$, $(x, z) \in V$. It is clear from the definition of backwards continuation that π is a flow on V.

If the mapping $t \to \pi(t, x, z)$ is differentiable at $t = 0$ with the derivative being locally Lipschitzian in $(x, z) \in V$, then we can define a vector field on V by assigning to each point $(x, z) \in V$ the derivative $\partial\pi(t, x, z)/\partial t|_{t=0}$. Initial value problems for the ODE defined by this vector field have unique solutions and define a flow which coincides with π. QED

3. Hyperbolic invariant manifolds. Let X be a smooth Banach manifold without boundary and let ϕ be a continuously differentiable semiflow on X. Let Y be a smooth, compact, connected submanifold of X and assume that Y is positively invariant under the semiflow ϕ. Denote by E the subset of the tangent bundle TX defined by $E = \bigcup_{y \in Y} T_yX$ and suppose that there exists a subbundle N of E which is complementary to the tangent bundle $T = TY$. Then E, T and N are vector bundles over Y. Since ϕ is continuously differentiable, we can define

$$\psi(t, y, z) = \Psi(t, y)z = D_2\phi(t, \phi(t, y))z, \qquad t \geqq 0$$

for all $y \in Y$ and $z \in E(y)$, and

$$\pi(t, y, z) = (\phi(t, y), \psi(t, y, z)), \qquad t \geqq 0.$$

Then π is a linear skew-product semiflow on E. It is called the *linearized skew-product semiflow around Y induced by the semiflow ϕ*. The vector bundle T is positively invariant under the semiflow $\pi = (\phi, \psi)$. The semiflow π induces, by restriction, a semiflow on T which is denoted by π^T and is called the *tangential flow induced by π* on T

$$\pi^T(t, y, z) = (\phi(t, y), \psi^T(t, y, z)), \qquad t \geqq 0$$

defined for $(y, z) \in T$. Analogously, π induces a semiflow π^N on N. In fact, if $P(y)$ denote projections on $E(y)$ which depend continuously on y and are such that $T(y) =$ null space of $P(y)$ and $N(y) =$ range space of $P(y)$, we can set for $(y, z) \in N$

$$\pi^N(t, y, z) = (\phi(t, y), P(\phi(t, y))\psi(t, y, z)), \qquad t \geqq 0.$$

Since T is a positively invariant set for π, the mapping π^N is a linear skew-product semiflow on N. It is called the *normal flow induced by π* on N. Let $\Sigma(E, \pi)$, $\Sigma(T, \pi^T)$ and $\Sigma(N, \pi^N)$ denote the spectra of the semiflows π, π^T and π^N on the vector bundles E, T and N, respectively. We say that Y is a *k-hyperbolic invariant manifold* under ϕ if there exists an $\alpha > 0$ such that $\Sigma(T, \pi^T) \subset (-\alpha, \alpha)$ and $\Sigma(N, \pi^N) \cap (-k\alpha, k\alpha) = \emptyset$. A *hyperbolic invariant manifold* under ϕ is simply a 1-hyperbolic invariant manifold.

THEOREM 3.1. *Let X be a smooth Banach manifold without boundary and let ϕ be a continuously differentiable semiflow on X. If $Y \subset X$ is a connected k-hyperbolic invariant manifold under ϕ and $\pi = (\phi, \psi)$ is the linearized skew-product semiflow around Y induced by the semiflow ϕ, then the tangent bundle of Y, $T = TY$, the unstable set U and the stable set S of Y under π decompose the tangent bundle of X as a Whitney sum $TX = T + U + S$, and the restrictions of π to U, of π to T and of ϕ to Y can be extended to flows on U, T and Y, respectively. Furthermore, if $\pi^U = (\phi, \psi^U)$, $\pi^T = (\phi, \psi^T)$ denote the skew-product flows on U, T obtained by extension of the restrictions of π to U, T, respectively, and if $\pi^S = (\phi, \psi^S)$ denotes the restriction of π to S, then there exist $K, \alpha > 0$ such that*

$$\begin{aligned} &|\psi^U(t, x, z)| \leqq K e^{k\alpha t}|z|, && t \leqq 0, \quad (x, z) \in U, \\ &|\psi^S(t, x, z)| \leqq K e^{-k\alpha t}|z|, && t \geqq 0, \quad (x, z) \in S, \\ &|\psi^T(t, x, z)| \leqq K e^{\alpha|t|}|z|, && t \in R, \quad (x, z) \in T. \end{aligned} \tag{3.1}$$

Proof. We can write

$$\pi(t, y, z) = (\phi(t, y), \psi(t, y, P(y)z)) + (\phi(t, y), \psi(t, y, [I - P(y)]z)).$$

Consequently, if $\psi(t, y, z) = \Psi(t, y)z$, $\psi^T(t, y, z) = \Psi^T(t, y)z$ and $\psi^N(t, y, z) = \Psi^N(t, y)z$, we have

$$\Psi(t, y)z = \Psi^N(t, y)P(y)z + \Psi^T(t, y)[I - P(y)]z.$$

Since $\Sigma(N, \pi^N) \cap \Sigma(T, \pi^T) = \varnothing$, it follows directly from the definitions of spectrum and dichotomy that

$$\Sigma(E, \pi) = \Sigma(N, \pi^N) \cup \Sigma(T, \pi^T).$$

If we denote $V_0 = U_{-\alpha} \cap S_\alpha$, $V_+ = U_{k\alpha}$ and $V_- = S_{-k\alpha}$ it follows from Theorem 2.1 that V_0 and V_+ are unions of a finite number of compact spectral subbundles and V_- is a countable union of spectral subbundles such that $TX = V_0 + V_- + V_+$, as a Whitney sum. Since $\Sigma(T, \pi^T) \subset (-\alpha, \alpha)$, we have $V_0 = T$, and it is also clear that $V_- = S$ and $V_+ = U$. Thus, T and U are finite-dimensional, and Proposition 2.2 implies that the restrictions of π to T and U can be extended to flows on T and U, respectively. It is clear that $\Sigma(U, \pi^U) \subset (k\alpha, +\infty)$, $\Sigma(T, \pi^T) \subset (-\alpha, \alpha)$, $\Sigma(S, \pi^S) \subset (-\infty, -k\alpha)$, and, therefore, the inequalities (3.1) are valid.

The dimensions of Y and $TY = T$ are the same. Consequently, Y is a finite-dimensional manifold invariant under ϕ. It follows that the restriction of ϕ to Y can be extended to a flow on Y. QED

4. Functional differential equations on manifolds. Let M be a separable smooth finite-dimensional connected manifold without boundary, and let TM be the tangent bundle of M, that is, TM is the union of the tangent spaces $T_yM = TM(y)$ of points $y \in M$, with $p_M : TM \to M$ denoting the projection that maps each $TM(y)$ onto the base point y. If I denotes the closed interval $I = [-r, 0]$ for $r > 0$, $C^0(I, M)$ denotes the set of continuous functions from I to M and $\rho : C^0(I, M) \to M$ is the evaluation map $\rho(\phi) = \phi(0)$, then a *retarded functional differential equation* (RFDE) *on* M is a continuous function $F: C^0(I, M) \to TM$ such that $p_M \circ F = \rho$.

The tangent bundle TM can be identified with $M \times R^m$ where $m = \dim M$. Then, for any RFDE F, there exists a function $f: C^0(I, M) \to R^m$, such that $F(\phi)$ can be identified with $(\xi(0), f(\xi))$, for all $\xi \in C^0(I, M)$. The RFDE (F) is frequently represented as $(x(t), \dot{x}(t)) = F(x_t) = (x(t), f(x_t))$ or, simply, $\dot{x}(t) = f(x_t)$, where, given a function x of a real variable and with values in the manifold M, we denote $x_t(\theta) = x(t+\theta)$, $\theta \in I$, whenever the right-hand side is defined.

Given a locally Lipschitzian RFDE (F) on M, its maximal solution $x(t)$ satisfying the initial condition ξ at $t = t_0$ (which necessarily exists and is unique) is sometimes denoted by $x(t; t_0, \xi, F)$ and x_t is denoted $x_t(t_0, \xi, F)$. The *solution map* or *semiflow* of F is then defined by $\phi(t, \xi) = x_t(0, \xi, F)$. The arguments ξ, F are dropped when confusion may not arise, and t_0 is dropped when it is equal to zero. If F is bounded and has bounded continuous derivative, then the solution map is a smoothing operator, in the sense that if it is uniformly bounded for t in compact sets of $[0, \infty)$, then for $t \geqq r$, the function $\phi(t, \cdot): C^0(I, M) \to C^0(I, M)$ maps bounded sets into relatively compact sets.

We denote by BC^k the set of bounded continuous functions from $C^0(I, M)$ into TM which have bounded continuous derivatives up to order $k \geqq 1$. The RFDEs on M we consider in the sequel will always be taken from BC^k for $k \geqq 1$. Each such RFDE (F) induces, by linearization, another RFDE (L) on the tangent bundle TM, which is called the *linear variational equation*. Being an RFDE on TM, the linear variational equation is a map $L: C^0(I, TM) \to T^2M$. The double tangent bundle T^2M can be identified with $M \times R^m \times R^m \times R^m$, and, therefore, if the given RFDE (F) on M is represented as $\dot{x}(t) = f(x_t)$, as done before, then the linear variational equation of F

can be represented, in an analogous fashion, as $(x(t), y(t), \dot{x}(t), \dot{y}(t)) = L(x_t, y_t) = (x(t), y(t), f(x_t), Df(x_t)y_t)$ or, simply, as a system of the two equations $\dot{x}(t) = f(x_t)$ and $\dot{y}(t) = Df(x_t)y_t$, where Df denotes the derivative of f. The solution maps ϕ of F and λ of the linear variational equation L are then related by $\lambda(t, \cdot) = D\phi(t, \cdot)$.

It is clear from the preceding discussion that we can, without loss of generality, restrict the study of the persistence of hyperbolic invariant manifolds, under small perturbations of RFDEs on a manifold M, to the particular case where the manifold is euclidean, i.e., $M = R^q$, for some integer q.

5. System of coordinates around hyperbolic invariant manifolds for FDEs. For a fixed real number $r > 0$, let $C = C([-r, 0]; R^q)$ denote the Banach space of continuous functions from the interval $[-r, 0]$ to R^q, taken with the uniform norm. Given a Banach space B and positive integers k, p, the set

$$BC^k(B; R^p) = \{f: B \to R^p, f \text{ is continuously differentiable and has bounded derivatives up to order } k\},$$

taken with the usual addition and multiplication by scalars and the uniform C^k-norm, is a Banach space. By uniform C^k-norm we mean

$$\|f\|_k = \sup\{|D^i f(\xi)|: i = 1, \cdots, k \text{ and } \xi \in B\}.$$

We are interested in discussing functional differential equations (FDE) defined by functions $f \in BC^k(C; R^n)$ with $k \geqq 1$, as

$$\dot{u}(t) = f(u_t), \tag{5.1}$$

where u_t denotes the segment of the function u defined over the interval $[t - r, t]$, i.e., $u_t(\theta) = u(t + \theta)$ for $\theta \in [-r, 0]$. The solutions of (5.1) define a semiflow $(t, \xi) \to u_t(\xi)$ on C, with $u_0 = \xi$. The mapping $u_t(\cdot): C \to C$ is C^k for all $t \geqq 0$ and is completely continuous for $t \geqq r$.

Let $M \subset C$ be a compact, connected, C^k-manifold which is k-hyperbolic under the semiflow defined by the solutions of (5.1). The vector bundle $E = \bigcup_{\omega \in M} T_\omega C$ can be identified with $M \times C$, since C is infinite-dimensional and M is a finite-dimensional manifold.

We will also consider the *linear variational equation* around M

$$\dot{v}(t) = Df(u_t(\omega))v_t \tag{5.2}$$

for each $\omega \in M$. The linearized semiflow around M which is induced by (5.1) is the linear skew-product semiflow defined, for $(\omega, \xi) \in E$, by

$$\pi(t, \omega, \xi) = (u_t(\omega), v_t(\omega, \xi)), \qquad t \geqq 0,$$

where $v_t(\omega, \xi) \in C$ denotes points in the orbit of (5.2), which passes through the point ξ at $t = 0$.

Because M is a k-hyperbolic manifold under (5.1), the vector bundle TM is invariant under the skew-product semiflow π, and TM has a complementary subbundle N of E, i.e., $E = TM + N$. Let U, S denote, respectively, the unstable and stable subbundles of N. The fibers $T_\omega = T_\omega M$, U_ω and S_ω can be identified with linear subspaces of C, and one can write $C = T_\omega + U_\omega + S_\omega$. Because M is connected, the dimensions of these fibers are independent of $\omega \in M$, and because M is hyperbolic and the semiflow π is completely continuous for $t \geqq r$, it follows that both T_ω and U_ω are finite-dimensional, with dimensions that we denote $d_T = \dim T_\omega$ and $d_U = \dim U_\omega$. We can choose bases for T_ω and U_ω consisting of vectors of unit length that depend on ω in a C^k fashion. These bases are arranged as columns of $q \times d_T$ matrices Φ_ω^T and

$q\times d_U$ matrices Φ^U_ω. To each point $(\omega, \xi)\in E$, we can associate coordinates $x\in R^{d_T}$, $y\in R^{d_U}$, $\zeta\in C$ by the relations $\xi=\xi^T_\omega+\xi^U_\omega+\xi^S_\omega$ where $\xi^T_\omega\in T_\omega$, $\xi^U_\omega\in U_\omega$, $\xi^S_\omega\in S_\omega$, $\xi^T=\Phi^T_\omega x$, $\xi^U_\omega=\Phi^U_\omega y$ and $\zeta=\xi^S_\omega$. These relations associate a unique quadruple $(\omega, x, y, \zeta)\in M\times R^{d_T}\times R^{d_U}\times C$ to each point $(\omega, \xi)\in E$. This system of coordinates (ω, x, y, ζ) around the hyperbolic invariant manifold M is redundant. In fact, the same point in a neighborhood of M can be represented in several ways in these coordinates, according to which point $\omega\in M$ is taken as origin of the coordinate system. In spite of this redundancy, the use of these coordinates facilitates the study of the persistence of hyperbolic invariant manifolds under perturbations.

Given a FDE

$$\dot{w}(t)=f(w_t)+g(w_t), \tag{5.3}$$

where $g\in BC^k$, and defining $v(t)=w(t)-u(t)$, we can write it as a perturbation of the linear variational equation (5.2) in the form

$$\dot{v}(t)=Df(u_t(\omega))v_t+G(u_t(\omega), v_t), \tag{5.4}$$

with $\omega\in M$, by defining

$$G(\phi, \psi)=f(\phi+\psi)-Df(\phi)\psi-f(\phi)+g(\phi+\psi).$$

Equation (5.4) defines a skew-product semiflow on E by

$$\tilde{\pi}(t, \omega, \xi)=(u_t(\omega), v_t(\omega, \xi)),$$

where $v_t(\omega, \xi)$ denotes points on the orbit of (5.4), which satisfies the initial condition $v_0=\xi$. The variation of constants formula for (5.4) can be written (see [4]) as

$$v_t=T_\omega(t, s)v_\sigma+\int_\sigma^t T_\omega(t, s)X_0G(u_s(\omega), v_s)\,ds, \qquad t\geqq\sigma, \tag{5.5}$$

where $T_\omega(t, \sigma)$ denotes the solution operator of the linear variational equation (5.2) and $X_0(\theta)$ is defined to be the identity I_q at $\theta=0$ and to be zero for $\theta\in[-r, 0]$ (notice that the columns of X_0 do not belong to C, but the formula still makes sense if interpreted as suggested by Hale in [4]).

As M is an hyperbolic compact manifold under the semiflow defined by (5.1), which is completely continuous for $t\geqq r$, it follows that (5.1) defines an ODE on M. The points $v_t(\omega)$, which must satisfy (5.5), can be represented in the system of coordinates introduced above as $(u_t(\omega), x(t), y(t), z_t)$, where $x(t)$, $y(t)$, z_t satisfy the variation of constants formulas obtained by projecting both sides of (5.5) along the "coordinate directions." More precisely, we have the following result.

THEOREM 5.1. *Let $f\in BC^k(C, R^q)$, $k\geqq 1$, and assume $M\subset C$ is a compact connected C^k-manifold that is k-hyperbolic under the semiflow defined by the solutions of*

$$\dot{u}(t)=f(u_t). \tag{5.6}$$

Then there exists a system of local coordinates around M, $(\omega, x, y, \zeta)\in M\times R^{d_T}\times R^{d_U}\times C$, where $d_T=\dim M$ and d_U is the dimension of the unstable bundle associated with the linear variational equation (5.2), *such that, for each $\omega\in M$, there exist two matrix-valued functions $\Psi^T_\omega(t, \sigma)$ and $\Psi^U_\omega(t, \sigma)$, which are continuously differentiable in $t, \sigma\in R$, a linear subspace L of C with codimension d_T+d_U, a linear operator $T^S_\omega(t, \sigma)$ acting on L which is continuously differentiable in $t\geqq\sigma$, a $q\times q$ matrix-valued function $X_0^{S,\omega,T}$ defined* *on* $[-r, 0]$ *which is continuous on* τ, *and functions* n, u, s *defined from* $M \times R^{d_T} \times R^{d_U} \times C \times BC^k(C, R)$ *into, respectively,* R^{d_T}, R^{d_U}, R^q *which are bounded and, for each fixed* $\omega \in M$, *are of class* BC^k *in the remaining variables, such that*

(i) $\Psi^T_\omega(t, t) = I_{d_T}$, $\Psi^U_\omega(t, t) = I_{d_U}$, $T^S_\omega(t, t) = I_L$ *for all* $t \in R$;

(ii) *there exist* K, α, $\alpha_0 > 0$ *with* $\alpha > k\alpha_0$ *such that*

$$e^{-\alpha_0|t|}|\Psi^T_\omega(t, \tau)| \to 0 \quad \text{as } |t| \to \infty \quad \text{for all } \tau \in R,$$
$$|\Psi^U_\omega(t, \tau)| \leqq K e^{\alpha(t-\tau)}, \qquad t \leqq \tau,$$
$$|T_\omega(t, \tau)\phi| \leqq K e^{-\alpha(t-\tau)}|\phi|, \qquad t \geqq \tau, \quad \phi \in S_{u_0(\omega)},$$

where S *denotes the stable bundle associated with the linear variational equation of* (5.6) *around* M;

(iii) $|X_0^{S,\omega,\tau}| \leqq 1$, $\omega \in M$, $\tau \in R$;

(iv) *the functions* n, u, s *and their partial derivatives relative to* x, y, ζ *vanish at the points* $(\omega, 0, 0, 0, 0) \in M \times R^{d_T} \times R^{d_U} \times L \times BC^k(C; R^q)$, *and each one of them assumes related values at all points* (ω, x, y, ζ) *that represent the same point of* C;

(v) *for each* $g \in BC^k(C; R^q)$, *the perturbed equation*

$$\dot{w}(t) = f(w_t) + g(w_t) \tag{5.7}$$

is, in the new coordinates and for $t \geqq \sigma$, *equivalent to the system*

$$\begin{aligned} x(t) &= \Psi^T_\omega(t, \sigma)x(\sigma) + \int_\sigma^t \Psi^T_\omega(t, \tau)n(u_\tau(\omega), x(\tau), y(\tau), z_\tau, g)\, d\tau, \\ y(t) &= \Psi^U_\omega(t, \sigma)y(\sigma) + \int_\sigma^t \Psi^U_\omega(t, \tau)u(u_\tau(\omega), x(\tau), y(\tau), z_\tau, g)\, d\tau, \\ z_t &= T^S_\omega(t, \sigma)z_\sigma + \int_\sigma^t T^S_\omega(t, \tau)X_0^{S,\omega,\tau}s(u_\tau(\omega), x(\tau), y(\tau), z_\tau, g)\, d\tau. \end{aligned} \tag{5.8}$$

Proof. We need to project both sides of the variation of constants formula (5.5) in the tangential, unstable and stable directions along the points $u_t(\omega) \in M$, as indicated in the discussion preceding the theorem. Forgetting, for the moment, the differentiability properties of the functions involved, and recalling that T and U are invariant under the skew-product semiflow associated with the linear variational equation, we see that the introduction of bases for the finite-dimensional fibers $T_{u_t(\omega)}$ and $U_{u_t(\omega)}$ and the representation of the projected equations in terms of these bases lead to the first two equations in system (5.8).

The linear spaces $S_{u_t(\omega)}$ have codimension $d_T + d_U$ in C and can be one-to-one mapped onto a fixed subspace L of C of the same codimension. Projecting (5.5) onto S and representing this projection in the subspace L, we obtain an equation of the form of the last equation in system (5.8). The term $T_\omega(t, \tau)X_0^{S,\omega,\tau}$ needs some explanation. First, we notice that $T_\omega(\tau + r, \tau)X_0$ is a matrix with columns in C because of the smoothing action of $T_\omega(t, \tau)$. In fact, though $X_0(\theta)$ is a matrix valued function defined for $\theta \in [-r, 0]$ and discontinuous at $\theta = 0$, the solutions of the FDE with initial conditions equal to each one of the columns of X_0 are continuous for $t \geqq 0$ and, consequently, after $t = r$ units of time all the segment of the solution from $t - r$ to t is continuous, showing that the columns of $T_\omega(\tau + r, \tau)$ do, indeed, belong to C. This matrix can be projected onto $T_{u_{\tau+r}(\omega)}$ and $U_{u_{\tau+r}(\omega)}$ to give components $[T_\omega(\tau + r, \tau)X_0]^T_{u_{\tau+r}(\omega)}$ and $[T_\omega(\tau + r, \tau)X_0]^U_{u_{\tau+r}(\omega)}$, respectively. Since $T_\omega(\tau + r, \tau)$ is a homeomorphism from $T_{u_\tau(\omega)}$ to $T_{u_{\tau+r}(\omega)}$ and from $U_{u_\tau(\omega)}$ to $U_{u_{\tau+r}(\omega)}$, because T and U are invariant under the semiflow, it follows that there exist unique matrix-valued functions

$X_0^{T,\omega,\tau}$ and $X_0^{U,\omega,\tau}$ whose columns belong to $T_{u_\tau(\omega)}$ and $U_{u_\tau(\omega)}$, respectively, and are such that

$$T_\omega(\tau+r,\tau)X_0^{T,\omega,\tau}=[T_\omega(\tau+r,\tau)X_0]^T_{u_{\tau+r}(\omega)},$$
$$T_\omega(\tau+r,\tau)X_0^{U,\omega,\tau}=[T_\omega(\tau+r,\tau)X_0]^U_{u_{\tau+r}(\omega)}.$$

Now, we can define $X_0^{S,\omega,\tau}=X_0-X_0^{T,\omega,\tau}-X_0^{U,\omega,\tau}$, and it becomes clear that the last equation in system (5.8) is correct, provided $T^S_\omega(t,\tau)X_0^{S,\omega,\tau}$ is understood in the same sense as $T_\omega(t,\tau)X_0$ was (notice the $X_0^{S,\omega,\tau}$ does not belong to L, as X_0 does not belong to C; for an explanation of this notation refer to [4]).

Properties (i) and (iii) are easy to verify, property (ii) is a consequence of the hyperbolicity of M through Theorem 3.1, and property (iv) results from the positive invariance of T, U, L under the skew-product semiflow associated with the linear variational equation around M and the invariance of M under (5.6).

It remains to establish the smoothness properties of the functions n, u, s. For this we need to show that the normal bundle N, the projections associated with the decomposition $C=T_{u_t(\omega)}+U_{u_t(\omega)}+S_{u_t(\omega)}$, the vectors forming the bases for $T_{u_t(\omega)}$ and $U_{u_t(\omega)}$ and the one-to-one mapping from $S_{u_t(\omega)}$ onto L can all be chosen to be C^k-smooth in t. The possibility of choosing a C^k-smooth normal bundle N can be proved by a slight modification of the proof given by Whitney [20] for the case when the manifold M is modeled in a finite-dimensional euclidean space. In fact, a "natural" choice of the normal bundle would only be C^{k-1} smooth, but the procedure introduced by Whitney in the cited paper can be used to smooth it to be of class C^k. It follows that the projections associated with the decomposition $C=T_{u_t(\omega)}+U_{u_t(\omega)}+L_{u_t(\omega)}$ are of class C^k in t, provided $U_{u_t(\omega)}$ and $S_{u_t(\omega)}$ are C^k in t. These are defined in terms of the null space and the range, respectively, of the linear projections $P(u_t(\omega))$, defined on $N_{u_t(\omega)}$, which are associated with the dichotomy of the linearized skew-product semiflow around M induced by the given equation. Although these projections are, at the outset, only required to depend continuously on the points in the manifold M, they are in fact of class C^k in t because their null spaces are related, forwards and backwards in time, by a semiflow of class C^k. More precisely, the null space of $P(\omega)$ is mapped onto the null space of $P(u_t(\omega))$ by the map $\xi\to v_t(\omega,\xi)$ given by the solutions of equation (5.2). Since this map is of class C^k in t, due to the general results on smoothness of solutions of FDEs (see [4]), and N is a C^k vector bundle, it follows that $P(u_t(\omega))$ is C^k in t. The possibility of choosing the one-to-one mapping from $S_{u_t(\omega)}$ onto L to be C^k-smooth in t is a direct consequence of the C^k-smoothness of $P(u_t(\omega))$. In order to get the C^k-smoothness in t for the bases taken for $T_{u_t(\omega)}$ and $U_{u_t(\omega)}$, we only need to choose them to be mapped one to each other by the flows on these bundles, since these flows are of class C^k in t. QED

6. Functional differential equations in coordinate form. Under certain general conditions discussed in the preceding section, the linearization of a given FDE around a hyperbolic compact manifold M, and the introduction of local coordinates around the manifold lead to a family of systems parametrized by $\omega\in M$ and of the form

$$\dot x(t)=N(t)x(t)+n(t,x(t),y(t),z_t,\lambda), \tag{6.1}$$

$$y(t)=\Psi(t,\sigma)y(\sigma)+\int_\sigma^t \Psi(t,\tau)u(\tau,x(\tau),y(\tau),z_\tau,\lambda)\,d\tau, \tag{6.2}$$

$$z_t=T(t,\sigma)z_\sigma+\int_\sigma^t T(t,\tau)X_0^{s,\tau}s(\tau,x(\tau),y(\tau),z_\tau,\lambda)\,d\tau, \tag{6.3}$$

where λ is a parameter in a Banach space Λ, $x(t)\in R^{d_N}$, $y(t)\in R^{d_U}$, $z_t\in L$, L is a linear subspace of $C = C([-r, 0]; R^q)$ of codimension $(d_N + d_U)$, d_N, d_U, q are nonnegative integers with $d_N \geqq 1$, $X_0^{S,\tau}$ is a $q \times q$ matrix-valued function defined on $[-r, 0]$ and continuous in τ with its columns belonging to L and satisfying $|X_0^{S,\tau}| \leqq 1$ for all $\tau \in R$, $N(t)$ and $\Psi(t, \tau)$ are matrix-valued functions defined for $t, \tau \in R$, $T(t, \tau)$ are linear operators acting on L for $t \geqq \tau$, and the following hypotheses are satisfied:

(H_1) The function N of t is bounded and continuous for all $t \in R$;

(H_2) The functions n, u, s of $(t, x, y, \zeta, \lambda)$ are bounded, continuously differentiable in x, y, ζ and their partial derivatives relative to x, y, ζ as well as the functions n, u, x themselves are all bounded by some $B(\mu, \varepsilon) > 0$ over the region $t \in R$, $|x|, |y|, |\zeta| \leqq \mu \leqq \mu_0$, $|\lambda| \leqq \varepsilon \leqq \varepsilon_0$, with the function $B(\mu, \varepsilon)$ being nondecreasing in μ and ε, and approaching zero as $\mu, \varepsilon \to 0$;

(H_3) The matrix-valued function $\Psi(t, \tau)$ is continuously differentiable in $t, \tau \in R$, satisfies $\Psi(t, \tau)\Psi(\tau, \sigma) = \Psi(t, \sigma)$ for all $t, \tau, \sigma \in R$ and $\Psi(t, t)$ is the identity matrix for all $t \in R$. The linear operators $T(t, \tau)$ defined on L are continuously differentiable in t, τ such that $t \geqq \tau$, satisfy $T(t, \tau)T(\tau, \sigma) = T(t, \sigma)$ for all $t \geqq \tau \geqq \sigma$, and $T(t, t)$ is the identity operator on L for all $t \in R$;

(H_4) There exist $K \geqq 1$ and $\alpha > \alpha_0 > 0$, such that

$$|\Psi(t, \tau)| \leqq K e^{\alpha(t-\tau)}, \qquad t \leqq \tau,$$

$$|T(t, \tau)\phi| \leqq K e^{-\alpha(t-\tau)}|\phi|, \qquad t \geqq \tau, \quad \phi \in L$$

and the principal matrix solution $\Phi(t, \tau)$ of $\dot{x} = N(t)x$ satisfies

$$|\Phi(t, \tau)| \leqq K e^{\alpha_0|t-\tau|} \quad \text{for all } t, \tau \in R.$$

In this section we consider a more restricted situation, which will be used later on to establish the general result on the persistence of hyperbolic manifolds. More precisely, the hypothesis (H_2) is replaced by

(H_2') The functions n, u, s of $(t, x, y, \zeta, \lambda)$ are bounded and continuous, vanish at all points where x, y, ζ are simultaneously zero, and are globally Lipschitzian in the coordinates x, y, ζ, in the sense that there exists a $D > 0$ such that

$$|n(t, x, y, \zeta, \lambda) - n(\theta, \bar{x}, \bar{y}, \bar{\zeta}, \lambda)| \leqq D(|x - \bar{x}| + |y - \bar{y}| + |\zeta - \bar{\zeta}|)$$

for all $t \in R$, $x, \bar{x} \in R^{d_N}$, $y, \bar{y} \in R^{d_U}$, $\zeta, \bar{\zeta} \in L$, $\lambda \in \Lambda$, and similarly for u and s.

In the proofs of the results of this section on the persistence of hyperbolic invariant manifolds for system (6.1)-(6.3) with $|\lambda|$ small, we use the following property of solutions $(x(t), y(t), z_t)$, $t \in R$, which have $y(t)$ and z_t bounded.

LEMMA 6.1. *Assume the hypotheses* (H_1)-(H_4) *hold. Then* $(x(t), y(t), z_t)$, $t \in R$, *is a solution of the system* (6.1)-(6.3) *with* $y(t)$ *and* z_t *bounded if and only if for some* γ *belonging to the interval* (α_0, α) *the function*

$$w(t, x(0)) = e^{-\gamma|t|}(x(t), y(t), z_t), \qquad t \in R \tag{6.4}$$

agrees, when $x(0) = b \in R^{d_N}$, *with a fixed point of the transformation* T *defined on the set of bounded continuous functions* $w: R \times R^{d_N} \to R^{d_N} \times R^{d_U} \times L$ *by*

$$\begin{aligned} Tw(t, b) = \Big(& e^{-\gamma|t|}\Phi(t, 0)b + e^{-\gamma|t|} \int_0^t \Phi(t, \tau)n(\tau, e^{\gamma|\tau|}w(\tau, b), \lambda)\, d\tau, \\ & e^{-\gamma|t|} \int_{+\infty}^t \Psi(t, \tau)u(\tau, e^{\gamma|\tau|}w(\tau, b), \lambda)\, d\tau, \\ & e^{-\gamma|t|} \int_{-\infty}^t T(t, \tau)X_0^{S,\tau}s(\tau, e^{\gamma|\tau|}w(\tau, b), \lambda)\, d\tau \Big). \end{aligned} \tag{6.5}$$

If (H_2) *is replaced by* (H_2') *the same holds.*

Proof. If $(x(t), y(t), z_t)$ is a solution of (6.1)–(6.3) which is defined for all $t \in R$ and has $y(t)$ bounded, it follows from hypothesis (H_4) that for $\sigma \geqq t$

$$\left| y(t) - \int_\sigma^t \Psi(t, \tau) u(\tau, x(\tau), y(\tau), z_\tau, \lambda)\, d\tau \right| = |\Psi(t, \sigma) y(\sigma)| \leqq K e^{\alpha(t-\sigma)} |y(\sigma)|$$

and, letting $\sigma \to +\infty$, we get

$$y(t) = \int_{+\infty}^t \Psi(t, \tau) u(\tau, x(\tau), y(\tau), z_\tau, \lambda)\, d\tau, \tag{6.6}$$

where the improper integral converges because u is bounded and Ψ satisfies the exponential estimates in assumption (H_4). Analogously, if $(x(t), y(t), z_t)$, $t \in R$, is a solution of (6.1)–(6.3) with z_t bounded, then

$$z_t = \int_{-\infty}^t T(t, \tau) X_0^{S,\tau} s(\tau, x(\tau), y(\tau), z_\tau, \lambda)\, d\tau. \tag{6.7}$$

Conversely, if $(x(t), y(t), z_t)$, $t \in R$, is a given continuous function satisfying equation (6.1) and (6.6)–(6.7) then it is a solution of the system (6.1)–(6.3), with $y(t)$ and z_t bounded, because of hypotheses (H_1) and (H_4).

The variation of constants formula for (6.1) gives

$$x(t) = \Phi(t, 0) x(0) + \int_0^t \Phi(t, \tau) n(\tau, x(\tau), y(\tau), z_\tau, \lambda)\, d\tau. \tag{6.8}$$

From hypothesis (H_2) or (H_2') there exists $B > 0$ which bounds n, and from hypothesis (H_4) we get

$$|x(t)| \leqq K e^{\alpha_0 |t|} |x(0)| + B \left| \int_0^t K e^{\alpha_0 |t-\tau|}\, d\tau \right| \leqq \left(|x(0)| + \frac{B}{\alpha_0} \right) K e^{\alpha_0 |t|}.$$

Consequently (6.1) is equivalent to (6.8) and $e^{-\gamma |t|} |x(t)|$, $t \in R$, is bounded. This finishes the proof of the statement. QED

THEOREM 6.2. *If the hypotheses* (H_1), (H_2'), (H_3), (H_4) *are satisfied and the constants* α_0, α, K *of hypothesis* (H_4) *and* D *of hypothesis* (H_2') *satisfy the inequality*

$$DK(K+1) \left[\frac{1}{\gamma - \alpha_0} + \frac{2}{\alpha - \gamma} + \frac{1}{\alpha + \gamma} \right] < 1 \tag{6.9}$$

for some γ *in the interval* (α_0, α), *then there exist continuous functions* h_1, h_2 *defined on* $R \times R^{d_N} \times \Lambda$ *and with values in* R^{d_U} *and* L, *respectively, which are bounded and such that the function* $x \to (h_1(t, x, \lambda), h_2(t, x, \lambda))$ *is Lipschitzian with Lipschitz constant* $(K+1)$ *and with* $h_1(t, 0, 0) = 0$, $h_2(t, 0, 0) = 0$ *for all* $t \in R$, $x \in R^{d_N}$, $\lambda \in \Lambda$, *such that the set*

$$M_\lambda = \{(t, x, y, \zeta) \in R \times R^{d_N} \times R^{d_U} \times L \colon y = h_1(t, x, \lambda), \zeta = h_2(t, x, \lambda)\}$$

is an integral manifold for system (6.1)–(6.3), *in the sense that if* $(t_0, x(t_0), y(t_0), z_{t_0}) \in M_\lambda$ *then the solution of* (6.1)–(6.3) *with this initial data stays in* M_λ *for all time.*

Furthermore, M_λ *is the maximal integral manifold for system* (6.1)–(6.3) *contained in* $R \times R^{d_N} \times V$, *for any bounded neighborhood of zero* $V \subset R^{d_U} \times L$.

Proof. The preceding lemma indicates that finding integral manifolds M_λ for (6.1)–(6.3), which belong to $R \times R^{d_N} \times V$ for some neighborhood of zero $V \subset R^{d_U} \times L$, amounts to finding fixed points of the mapping T in the lemma. These fixed points are studied, in the present proof, by an application of the contraction mapping principle to a specific set of continuous functions $w \colon R \times R^{d_N} \to R^{d_N} \times R^{d_U} \times L$ taken with a metric generated by a suitable family of pseudonorms.

Let us denote by W the set of bounded continuous functions $w: R\times R^{d_N}\to R^{d_N}\times R^{d_U}\times L$ that satisfy

$$|w(t,b)-w(t,\bar{b})|\leqq (K+1)|b-\bar{b}| \quad \text{for all } b,\ \bar{b}\in R^{d_N},\quad t\in R,$$

where $K\geqq 1$ is the constant in the hypothesis (H$_4$). The set W is a complete metric space with the topology generated by the family of pseudonorms

$$\text{(6.10)}\qquad \|w\|_n=\sup\{|w(t,b)|: t\in R, |b|\leqq n\},\quad n=1,2,\cdots.$$

It is clear that, for each $w\in W$, Tw is a continuous function from $R\times R^{d_N}$ into $R^{d_N}\times R^{d_U}\times L$. Using the hypotheses (H$_2'$) and (H$_4$) we get

$$\begin{aligned}|Tw(t,b)|\leqq \Bigg[& e^{-\gamma|t|}\, e^{\alpha_0|t|}K|b|+e^{-\gamma|t|}\int_0^t K\, e^{\alpha_0|t-\tau|}D\, e^{\gamma|\tau|}\,d\tau \\ & +e^{-\gamma|t|}\int_t^{+\infty} K\, e^{\alpha(t-\tau)}D\, e^{\gamma|\tau|}\,d\tau \\ & +e^{-\gamma|t|}\int_{-\infty}^{t} K\, e^{-\alpha(t-\tau)}D\, e^{\gamma|\tau|}\,d\tau\Bigg]\sup_{\tau\in R}|w(\tau,b)|.\end{aligned}$$

Consequently, as $\gamma\in(\alpha_0,\alpha)$, we obtain

$$|Tw(t,b)|\leqq K|b|+DK\left[\frac{1}{\gamma-\alpha_0}+\frac{2}{\alpha-\gamma}+\frac{1}{\alpha+\gamma}\right]\sup_{\tau\in R}|w(\tau,b)|.$$

This shows that Tw is bounded for each $w\in W$. In a similar way, and using (6.9), one obtains

$$\begin{aligned}|Tw(t,b)-Tw(t,\bar{b})| & \leqq K|b-\bar{b}|+DK\left[\frac{1}{\gamma-\alpha_0}+\frac{2}{\alpha-\gamma}+\frac{1}{\alpha+\gamma}\right]\sup_{\tau\in R}|w(\tau,\bar{b})-w(\tau,\bar{b})| \\ & <(K+1)|b-\bar{b}|.\end{aligned}$$

On the other hand, if w, $\bar{w}\in W$, we have

$$|Tw(t,b)-T\bar{w}(t,b)|\leqq DK\left[\frac{1}{\gamma-\alpha_0}+\frac{2}{\alpha-\gamma}+\frac{1}{\alpha+\gamma}\right]\sup_{\tau\in R}|w(\tau,b)-\bar{w}(\tau,b)|,$$

which implies that

$$\|Tw-T\bar{w}\|_n\leqq DK\left[\frac{1}{\gamma-\alpha_0}+\frac{2}{\alpha-\gamma}+\frac{1}{\alpha+\gamma}\right]\|w-\bar{w}\|_n.$$

Since Condition (6.9) is satisfied, then T is a uniform contraction from W into itself, in the given family of pseudonorms. The contraction mapping principle implies that there exists a unique fixed point of T in the set W and that this fixed point depends continuously on λ.

Let $w^*=Tw^*$ be the fixed point of T in W and define the functions h_1 and h_2 by

$$(b, h_1(t,b,\lambda), h_2(t,b,\lambda))=w^*(t,b).$$

It is clear that h_1, h_2 are defined for $t\in R$, $b\in R^{d_N}$, $\lambda\in\Lambda$, that they are continuous bounded functions which are Lipschitzian in the variable b with Lipschitz constant $(K+1)$ and that they vanish at the points $(t,0,0)$.

Suppose $(x(t), y(t), z_t)$, $t\in R$, is any solution of (6.1)-(6.3) with $y(t)$ and z_t bounded, and denote $b=x(0)$. Lemma 6.1 implies that $w(t)=e^{-\gamma|t|}(x(t),y(t),z_t)$ is a fixed point of the map T_b defined on the set $B(R)$ of the continuous bounded functions $w: R\to R^{d_N}\times R^{d_U}\times L$ by the same formula (6.5) as in the definition of T, but replacing

$w(t, b)$ by $w(t)$. The argument used above for T also shows that T_b maps $B(R)$ into itself and is a contraction in the supremum norm on $B(R)$. Therefore T_b has a unique fixed point in $B(R)$ which must satisfy $w(t) = w^*(t, b)$. It follows that M_λ is the maximal integral manifold in $R \times R^{d_N} \times V$, for any $V \subset R^{d_U} \times L$ which is a bounded neighborhood of zero. QED

Remark. The functions h_1, h_2 of the previous theorem do not depend on the particular value of $\gamma \in (\alpha_0, \alpha)$, provided it satisfies (6.9). In fact, if $\alpha_0 < \gamma_1 < \gamma_2 < \alpha$ and the function $w_1(t, x(0))$ is bounded and given as in (6.4) with $\gamma = \gamma_1$, then the function $w_2(t, x(0))$ also given as in (6.4) but with $\gamma = \gamma_2$ is also bounded. The uniqueness of the fixed point of the mapping T implies that $w_1 = w_2$.

The structure of the solutions around M_λ is preserved under small perturbations, as is illustrated by the following result.

THEOREM 6.3. *Assume the same hypotheses as for Theorem* 6.2. *Then the manifold* M_λ *is the intersection of two manifolds* S_λ, $U_\lambda \subset R \times R^{d_N} \times R^{d_U} \times L$ *which are positive integral manifolds for* (6.1)-(6.3), *and are such that solutions with initial data in* S_λ *approach* M_λ *as* $t \to +\infty$ *and solutions with initial data in* U_λ *are globally defined and approach* M_λ *as* $t \to -\infty$. *Moreover* S_λ, U_λ *are homeomorphic to* $R \times L$, $R \times R^{d_U}$, *respectively, and have the forms*

$$S_\lambda = \{(t, x, y, \zeta) \in R \times R^{d_N} \times R^{d_U} \times L \colon y = h^S(t, x, \zeta, \lambda)\}$$

and

$$U_\lambda = \{(t, x, y, \zeta) \in R \times R^{d_N} \times R^{d_U} \times L \colon \zeta = h^U(t, x, y, \lambda)\},$$

where the functions $(x, \zeta) \to h^S(t, x, \zeta, \lambda)$ *and* $(x, y) \to h^U(t, x, y, \lambda)$ *are Lipschitzian homeomorphisms from, respectively,* $R^{d_N} \times L$ *to* S_λ *and* $R^{d_N} \times R^{d_U}$ *to* U_λ, *with Lipschitz constant* $(K+1)$ *and satisfy* $h^S(t, x, \zeta, 0) = 0$, $h^U(t, x, y, 0) = 0$. *In addition, there exist* σ, $C > 0$ *such that, if* $(x(t), y(t), z_t)$ *is a solution of* (6.1)-(6.3) *with initial condition in* S_λ, *then*

$$|(x(t), y(t), z_t)| \leqq C(|x(0)| + |z_0|)\, e^{\sigma t}, \qquad t \geqq 0$$

and, if $(x(t), y(t), z_t)$ *is a solution of* (6.1)-(6.3) *with initial condition in* U_λ, *then*

$$|(x(t), y(t), z_t)| \leqq C(|x(0)| + |y(0)|)\, e^{-\sigma t}, \qquad t \leqq 0,$$

where $\sigma \in (\alpha_0, \gamma)$.

Proof. If $(x(t), y(t), z_t)$ is an arbitrary solution of (6.1)-(6.3) which is defined for all $t \geqq 0$ and has $y(t)$ bounded for $t \geqq 0$, we find, as in the proof of Lemma 6.1, that

$$y(t) = \int_{+\infty}^{t} \Psi(t, \tau) u(\tau, x(\tau), y(\tau), z_\tau, \lambda)\, d\tau.$$

Consequently, based on the discussion in Lemma 6.1 and Theorem 6.2, we expect that looking for the set S_λ will amount to finding fixed points of the transformation T^S defined on the set of bounded continuous functions $w \colon R^+ \times R^{d_N} \times L \to R^{d_N} \times R^{d_U} \times L$, by

$$\begin{aligned} T^S w(t, b, \zeta) = \Big(& e^{-\gamma t}\Phi(t, 0) b + e^{-\gamma t} \int_0^t \Phi(t, \tau) n(\tau, e^{\gamma \tau} w(\tau, b, \zeta), \lambda)\, d\tau, \\ & e^{-\gamma t} \int_{+\infty}^{t} \Psi(t, \tau) u(\tau, e^{\gamma\tau} w(\tau, b, \zeta), \lambda)\, d\tau, \qquad (6.11) \\ & e^{-\gamma t} T(t, 0)\zeta + e^{-\gamma t} \int_0^t T(t, \tau) X_0^{S,\tau} s(\tau, e^{\gamma\tau} w(\tau, b, \zeta), \lambda)\, d\tau \Big), \end{aligned}$$

where $\gamma \in (\alpha_0, \alpha)$.

We denote by W^S the set of bounded continuous function $w: R^+ \times R^{d_U} \times L \to R^{d_N} \times R^{d_U} \times L$ which satisfy

$$|w(t, b, \zeta) - w(t, \bar{b}, \bar{\zeta})| \leqq (K+1)(|b-\bar{b}| + |\zeta - \bar{\zeta}|)$$

for all $b, \bar{b} \in R^{d_N}$, $\zeta, \bar{\zeta} \in L$, $t \in R^+$, where $K \geqq 1$ is the constant in the hypothesis (H_4). The set W^S is a complete metric space with the topology generated by the family of pseudonorms

$$\|w\|_n = \sup \{|w(t, b, \zeta)| : t \in R^+, |b| \leqq n, |\zeta| \leqq n\}, \qquad n = 1, 2, \cdots.$$

Similar to what was done in the proof of Theorem 6.2 for the mapping T on W, it can be shown that T^S is a uniform contraction on W^S and, therefore, there exists a unique fixed point of T^S in W^S and it depends continuously on λ.

If we let $W^S = T^S w^S$ be the fixed point of T^S on W^S and define the function h^S by

$$(b, h^S(t, b, \zeta, \lambda), \zeta) = w^S(t, b, \zeta),$$

it is clear that the set S_λ, defined as in the statement of the theorem, is a positive integral manifold for system (6.1)-(6.3) and is homeomorphic to $R^{d_N} \times L$.

Next, we will show that $|w^S(t, b, \zeta)| \to 0$ as $t \to +\infty$. Let $\mu = \limsup_{t \to +\infty} |w^S(t, b, \zeta)|$. Because of (6.9), we can choose $\delta > 1$ so that

$$KD\left[\frac{1}{\gamma - \alpha_0} + \frac{1}{\gamma - \alpha} + \frac{1}{\gamma + \alpha}\right]\delta < 1.$$

If $\mu > 0$, then there is a $\sigma > 0$ so that $|w^S(t, b, \zeta)| \leqq \mu\delta$ for $t \geqq \sigma$. Then, using formula (6.11) and the estimates available for its terms, we get for $t \geqq \sigma$

$$\begin{aligned} |w^S(t, b, \zeta)| = |T^S w^S(t, b, \zeta)| &\leqq K e^{-(\gamma - \alpha_0)t}|b| + K e^{-(\alpha+\gamma)t}|\zeta| \\ &\quad + KD\left[\frac{e^{-(\gamma-\alpha_0)(t-\sigma)}}{\gamma - \alpha_0} + \frac{e^{-(\gamma+\alpha)(t-\sigma)}}{\gamma+\alpha}\right] \sup_{t \geqq 0} |w^S(\tau, b, \zeta)| \\ &\quad + KD\left[\frac{1}{\gamma-\alpha_0} + \frac{1}{\gamma-\alpha} + \frac{1}{\gamma+\alpha}\right]\mu\delta. \end{aligned}$$

Letting $t \to +\infty$, we get

$$\mu \leqq KD\left[\frac{1}{\gamma-\alpha_0} + \frac{1}{\gamma-\alpha} + \frac{1}{\gamma+\alpha}\right]\delta\mu < \mu,$$

which is a contradiction. Hence $\mu = 0$. This proves that $|w^S(t, b, \zeta)| \to 0$ as $t \to +\infty$.

Now we derive the exponential rate of decay of w^S as $t \to +\infty$. Let $v(t, b, \zeta) = \sup_{\tau \geqq t} |w^S(\tau, b, \zeta)|$. Since $|w^S(\tau, b, \zeta)| \to 0$ as $\tau \to +\infty$, for every $t \geqq 0$ there is a $\sigma \geqq t$ such that

$$v(\tau, b, \zeta) = v(\sigma, b, \zeta) = |w^S(\sigma, b, \zeta)|, \qquad t \leqq \tau \leqq \sigma.$$

On the other hand, estimates using formula (6.11) give

$$\begin{aligned} |w^S(t, b, \zeta)| \leqq K e^{-(\gamma-\alpha_0)t}|b| + K e^{-(\gamma+\alpha)t}|\zeta| &+ \int_0^t K e^{-(\gamma-\alpha_0)(t-\tau)} D |w^S(\tau, b, \zeta)|\, d\tau \\ &+ \int_t^{+\infty} K e^{-(\gamma-\alpha)(t-\tau)} D |w^S(\tau, b, \zeta)|\, d\tau \\ &+ \int_0^t K e^{-(\gamma+\alpha)(t-\tau)} D |w^S(\tau, b, \zeta)|\, d\tau. \end{aligned}$$

Consequently,

$$\begin{aligned} v(t, b, \zeta) = v(\sigma, b, \zeta) \leqq K e^{-(\gamma-\alpha_0)t}|b| + K e^{-(\gamma+\alpha)t}|\zeta| + \int_0^t K e^{-(\gamma-\alpha_0)(t-\tau)} Dv(\tau, b, \zeta) \\ + \int_t^\sigma K e^{-(\gamma-\alpha_0)(\sigma-\tau)} Dv(\tau, b, \zeta)\, d\tau \\ + \int_t^{-\infty} K e^{(\alpha-\gamma)(t-\tau)} Dv(\tau, b, \zeta)\, d\tau \\ + \int_0^t K e^{-(\gamma+\alpha)(t-\tau)} Dv(\tau, b, \zeta)\, d\tau \\ + \int_t^\sigma K e^{-(\gamma+\alpha)(\sigma-\tau)} Dv(\tau, b, \zeta)\, d\tau \end{aligned}$$

and, therefore,

$$\begin{aligned} v(t, b, \zeta) \leqq K e^{-(\gamma-\alpha_0)t}(|b| + |\zeta|) + \int_0^t K e^{-(\gamma-\alpha_0)(t-\tau)} Dv(\tau, b, \zeta)\, d\tau \\ + KD\left(\frac{1}{\gamma-\alpha_0} + \frac{1}{\alpha-\gamma} + \frac{1}{\gamma+\alpha}\right) v(t, b, \zeta). \end{aligned}$$

Due to (6.9), we can write

$$\begin{aligned} e^{(\gamma-\alpha_0)t} v(t, b, \zeta) \leqq \left[1 - KD\left(\frac{1}{\gamma-\alpha_0} + \frac{1}{\alpha-\gamma} + \frac{1}{\gamma+\alpha}\right)\right]^{-1} \\ \cdot K\left[|b| + |\zeta| + \int_0^t D e^{(\gamma-\alpha_0)\tau} v(\tau, b, \zeta)\, d\tau\right]. \end{aligned}$$

Applying the Gronwall inequality we obtain

$$|v(t, b, \zeta)| \leqq C(|b| + |\zeta|)\, e^{-Bt},$$

where

$$C = K \Big/ \left[1 - KD\left(\frac{1}{\gamma-\alpha_0} + \frac{1}{\alpha-\gamma} + \frac{1}{\alpha+\gamma}\right)\right] \quad \text{and} \quad B = \gamma - \alpha_0 - CD.$$

From (6.9), noting that $K \geqq 1$, it is easy to verify that C, B are positive. It follows that there exist $C, B > 0$ such that

$$|(x(t), y(t), z_t)| \leqq C(|x(0)| + |z_0|)\, e^{(\gamma-B)t}, \qquad t \geqq 0 \tag{6.12}$$

for every solution $(x(t), y(t), z_t)$ of (6.1)-(6.3) which has initial data on S_λ.

In a similar way, we obtain the manifold U_λ. Now, we look for fixed points of the mapping defined on the set of bounded continuous functions $w: R^- \times R^{d_N} \times R^{d_U} \to R^{d_N} \times R^{d_U} \times L$ by

$$T^U w(t, b, c) = \Big(e^{\gamma t}\Phi(t, 0)b + e^{\gamma t} \int_0^t \Phi(t, \tau) n(\tau, e^{-\gamma\tau} w(\tau, b, c), \lambda)\, d\tau,$$

$$e^{\gamma t}\Psi(t,0)c+e^{\gamma t}\int_0^t \Psi(t,\tau)u(\tau,e^{-\gamma\tau}w(\tau,b,c),\lambda)\,d\tau,$$

$$e^{\gamma t}\int_{-\infty}^t T(t,\tau)X_0^{S,\tau}s(\tau,e^{-\gamma\tau}w(\tau,b,c),\lambda)\,d\tau\Big).$$

Similarly, we obtain for some constants $C, B>0$

$$|(x(t),y(t),z_t)|\leqq C(|x(0)|+|y(0)|)\,e^{-(\gamma-B)t},\qquad t\leqq 0 \tag{6.13}$$

for every solution $(x(t),y(t),z_t)$ of system (6.1)-(6.3) which has initial data on U_λ.

It is clear, from the above construction and the proofs of Lemma 6.1 and Theorem 6.2, that $M_\lambda=S_\lambda\cap U_\lambda$ and the trajectory of any solution $(x(t),y(t),z_t)$ for which $e^{-\gamma|t|}|(x(t),y(t),z_t)|$ is bounded for $t\in R$ lies necessarily on M_λ. Consequently, (6.12) and (6.13) imply the exponential estimates in the statement of the theorem. QED

The following result establishes the smoothness properties of M_λ, S_λ and U_λ.

THEOREM 6.4. *If the hypotheses* (H_1), (H_2'), (H_3), (H_4) *are satisfied, the functions n, u, s are continuously differentiable with bounded derivatives up to order $k\geqq 1$, relative to x, y, ζ, and the constants α_0, α, K of hypothesis* (H_4), *and D of hypothesis* (H_2') *satisfy the inequalities*

$$\alpha>k\alpha_0 \tag{6.14}$$

and

$$DK(K+1)\left[\frac{1}{\gamma-\alpha_0}+\frac{2}{\gamma-\alpha}+\frac{1}{\gamma+\alpha}\right]<1 \tag{6.15}$$

for some γ in the interval $(k\alpha_0,\alpha)$, then the function $x\to(h_1(t,x,\lambda),h_2(t,x,\lambda))$, defined as in Theorem 6.2, *is of class C^k, and the same is true for the functions $(x,\zeta)\to h^S(t,x,\zeta,\lambda)$ and $(x,y)\to h^U(t,x,y,\lambda)$ of Theorem* 6.3.

Proof. We recall from the proof of Theorem 6.2 that h_1, h_2 were defined in terms of the fixed point of the transformation T in the set W. Consequently, if w denotes this fixed point, in order to prove that $h_1(t,x,\lambda)$, $h_2(t,x,\lambda)$ are C^k functions of x, we only need to show that the function $b\to w(t,b)$ is of class C^k. This will be done by induction.

Let us assume the hypothesis of the theorem is satisfied for $k=1$. If the derivative $\partial w(t,b)/\partial b$ exists, it must satisfy the equation obtained by formally differentiating $w=Tw$ relative to b. In particular, $\partial w/\partial b$ must be a fixed point of the mapping F defined for functions taking $R\times R^{d_N}$ to linear transformations of R^{d_N} into $R^{d_N}\times R^{d_U}\times L$

$$Fv(t,b)=\Big(e^{-\gamma|t|}\Phi(t,0)+e^{-\gamma|t|}\int_0^t\Phi(t,\tau)\frac{\partial n}{\partial w}(\tau,e^{\gamma|\tau|}w(\tau,b),\lambda)\,e^{\gamma|\tau|}v(\tau,b)\,d\tau,$$

$$e^{-\gamma|t|}\int_{+\infty}^t\Psi(t,\tau)\frac{\partial u}{\partial w}(\tau,e^{\gamma|\tau|}w(\tau,b),\lambda)\,e^{\gamma|\tau|}v(\tau,b)\,d\tau, \tag{6.16}$$

$$e^{-\gamma|t|}\int_{-\infty}^t T(t,\tau)X_0^{s,\tau}\frac{\partial s}{\partial w}(\tau,e^{\gamma|\tau|}w(\tau,b),\lambda)\,e^{\gamma|\tau|}v(\tau,b)\,d\tau\Big).$$

We remark that the improper integrals converge because of hypothesis (H_4), the boundedness of the partial derivatives of u, s and the assumption $\alpha>\alpha_0$. Let Z denote the set of all functions continuously differentiable in the second argument and defined from $R \times R^{d_N}$ to the set of all linear transformations of R^{d_N} into $R^{d_N} \times R^{d_U} \times L$ which satisfy $|v(t, b)| \leqq K+1$, for all $t \in R$, $b \in R^{d_N}$. If $v \in Z$ then

$$\begin{aligned} |Fv(t, b)| &\leqq K e^{-(\gamma-\alpha_0)|t|} + e^{-\gamma|t|} \left| \int_0^t K e^{\alpha_0|t-\tau|} D e^{\gamma|\tau|}(K+1)\, d\tau \right| \\ &\quad + e^{-\gamma|\tau|} \int_t^{+\infty} K e^{\alpha(t-\tau)} D e^{\gamma|\tau|}(K+1)\, d\tau \\ &\quad + e^{-\gamma|\tau|} \int_{-\infty}^t K e^{-\alpha(t-\tau)} D e^{\gamma|\tau|}(K+1)\, d\tau \\ &\leqq K + DK(K+1)\left[\frac{1}{\gamma-\alpha_0} + \frac{2}{\alpha-\gamma} + \frac{1}{\alpha+\gamma}\right] < K+1. \end{aligned} \tag{6.17}$$

Therefore, F maps the set Z into itself. Let us consider the sequence $\{v_i\}$ of functions, taking $R \times R^{d_N}$ into the linear transformations of R^{d_N} into $R^{d_N} \times R^{d_U} \times L$ which is defined recursively by

$$v_1 = 0 \quad \text{and} \quad v_{i+1} = F(v_i), \qquad i \geqq 1.$$

Since F maps Z into itself, we have $\{v_i\} \subset Z$. The set Z with the metric inherited from the usual uniform norm is a complete metric space. As in (6.17), we get

$$\|v_{i+1} - v_i\| = \|Fv_i - Fv_{i-1}\| \leqq DK\left[\frac{1}{\gamma-\alpha_0} + \frac{2}{\alpha-\gamma} + \frac{1}{\alpha+\gamma}\right] \|v_i - v_{i-1}\|. \tag{6.18}$$

It follows from (6.9) that $\{v_i\}$ is a Cauchy sequence in Z and, consequently, converges to some $v \in Z$ as $i \to \infty$. Clearly v is a fixed point of F.

Now, we can prove that v is, indeed, the derivative $\partial w/\partial b$. Fix $t \in R$, $b \in R^{d_N}$ and let σ be a function defined for small values of $\varepsilon > 0$ by

$$\sigma(\varepsilon) = \sup_{\substack{t \in R \\ |h| \leqq \varepsilon}} \frac{|w(t, b+h) - w(t, b) - v(t, b)h|}{|h|}. \tag{6.19}$$

In order to prove that $\partial w/\partial b$ exists and is equal to v we need to show that $\sigma(\varepsilon) \to 0$ as $\varepsilon \to 0$. From formulas (6.5) and (6.16), using the first order Taylor expansion and hypotheses (H_2') and (H_4) in a similar way as for inequalities (6.17) and (6.18), we get

$$\begin{aligned} &|w(t, b+h) - w(t, b) - v(t, b)h| \\ &\quad \leqq DK\left[\frac{1}{\gamma-\alpha_0} + \frac{2}{\alpha-\gamma} + \frac{1}{\alpha+\gamma}\right] \\ &\qquad \cdot \sup_{\tau \in R} [|w(\tau, b+h) - w(\tau, b) - v(\tau, b)h| + o(|w(\tau, b+h) - w(\tau, b)|)]. \end{aligned}$$

Recalling from Theorem 6.2 that $w(t, b)$ is Lipschitzian in b with Lipschitz constant equal to $(K+1)$, we get

$$\sigma(\varepsilon) \leqq DK\left[\frac{1}{\gamma-\alpha_0} + \frac{2}{\alpha-\gamma} + \frac{1}{\alpha+\gamma}\right] \sigma(\varepsilon) + o(\varepsilon)$$

as $\varepsilon \to 0$. Applying (6.9) we get $\sigma(\varepsilon) = o(\varepsilon)$, proving that $w(t, b)$ is differentiable in b and the derivative $\partial w(t, b)/\partial b$ is the continuous function v defined above. The preceding reasoning shows that the function $b \to w(t, b)$ is continuously differentiable and, consequently, also $b \to (h_1(t, b, \lambda), h_2(t, b, \lambda))$ is.

Now, let us assume that $b \to w(t, b)$ is of class C^j for a certain $j \geqq 1$ and the hypothesis of the theorem is satisfied for $k = j+1$. Let $v = \partial w^j/\partial b^j$. Differentiating j times both sides of equation $w = Tw$, we get

$$
\begin{aligned}
v(t, b) = \Bigg(& e^{-\gamma|t|} \int_0^t \Phi(t, \tau) \frac{\partial n}{\partial w}(\tau, e^{\gamma|\tau|} w(\tau, b), \lambda)\, e^{\gamma|\tau|} v(\tau, b)\, d\tau, \\
& e^{-\gamma|t|} \int_{+\infty}^t \Psi(t, \tau) \frac{\partial u}{\partial w}(\tau, e^{\gamma|\tau|} w(\tau, b), \lambda)\, e^{\gamma|\tau|} v(\tau, b)\, d\tau, \\
& e^{-\gamma|t|} \int_{-\infty}^t T(t, \tau) X_0^{S,\tau} \frac{\partial s}{\partial w}(\tau, e^{\gamma|\tau|} w(\tau, b), \lambda)\, e^{\gamma|\tau|} v(\tau, b)\, d\tau \Bigg) \\
& + (\text{terms not involving } v).
\end{aligned} \tag{6.20}
$$

The terms not involving v contain derivatives of n, u, s relative to w up to order j, derivatives of w up to order $j-1$ and exponential factors of the form $e^{i\gamma|\tau|}$ for $i = 1, 2, \cdots, j$. The improper integrals in these terms converge because of hypothesis (H_4), the boundedness of all the partial derivatives of n, u, s up to order j, and the assumption $\alpha > k\alpha_0$. We can let T be the mapping transforming v to the function $T(v)$ of (t, b) according to the right-hand side of (6.20) and define recursively the sequence $\{v_i\}$ by

$$v_i = 0 \quad \text{and} \quad v_{i+1} = T(v_i) \quad \text{for } i \geqq 1.$$

As in the first part of the proof we have that $\{v_i\}$ is a Cauchy sequence and, consequently, it converges to a fixed point of T which must be $v = \partial^j w/\partial b^j$. Clearly, the functions $v_i(t, b)$ are differentiable in b, and $\partial v_{i+1}/\partial b$ is given by the right-hand side of (6.20) with v replaced by $\partial v_i/\partial b$. Proceeding as for (6.18), we get

$$\left\| \frac{\partial v_{i+1}}{\partial b} - \frac{\partial v_i}{\partial b} \right\| \leqq DK \left[\frac{1}{\gamma - \alpha_0} + \frac{2}{\alpha - \gamma} + \frac{1}{\alpha + \gamma} \right] \left\| \frac{\partial v_i}{\partial b} - \frac{\partial v_{i-1}}{\partial b} \right\|.$$

From (6.9), we get that $\{\partial v_i/\partial b\}$ is a Cauchy sequence. Arguing as in the first part of the proof, where the first derivative was handled, we can show that $\partial v/\partial b = \partial w^{j+1}/\partial b^{j+1}$ exists and is equal to the limit of $\{\partial v_i/\partial b\}$ as $i \to \infty$. This completes the induction.

The smoothness of h^U and h^S can be handled in a similar way. QED

7. Functional differential equations with hyperbolic invariant manifolds. As before, consider $r > 0$ and let $C = C([-r, 0]; R^q)$ denote the Banach space of continuous functions from the interval $[-r, 0]$ into R^q, where q is a positive integer and C is taken with the uniform norm. Let us consider a FDE

$$\dot{u}(t) = f(u_t), \tag{7.1}$$

where $f \in BC^k(C, R^n)$ with $k \geqq 1$, and suppose that $M \subset C$ is a compact, connected, C^k-manifold which is k-hyperbolic under the semiflow defined by the solutions of (7.1). It is known from § 5 that the equation can be linearized around the manifold M and a system of coordinates can be introduced around M so that the equation becomes of the form discussed in § 6. The aim of the present section is to show how the results on the persistence and smoothness of integral manifolds for systems in coordinate form, as presented in the previous section, can be applied to (7.1), yielding the persistence of an hyperbolic invariant manifold close to M, under small perturbations of (7.1).

THEOREM 7.1. *Let $f \in BC^k(C, R^q)$, $k \geqq 1$, and assume that $M_0 \subset C$ is a compact, connected, C^k-manifold which is k-hyperbolic under the semiflow defined by the solutions of the equation*

$$\dot{u}(t) = f(u_t). \tag{7.2}$$

If $g \in BC^k(C, R^q)$ and $\|g\|_1 = \sup\{|g(\phi)|, \|g'(\phi)\|: \phi \in C\}$ is sufficiently small, then there exists a C^k-manifold $M_g \subset C$ that is invariant under the perturbed equation

$$\dot{u}(t) = f(u_t) + g(u_t). \tag{7.3}$$

There exists a neighborhood $O \subset C$ of M_0 such that, for $\|g\|_1$ sufficiently small, the manifold M_g is the maximal invariant set for (7.3) *which is contained in O. The manifold M_g depends continuously in g, in the sense that M_g can be made arbitrarily close to M_0 in the Hausdorff metric by choosing $\|g\|_1$ sufficiently close to zero. Furthermore, there exist C^k-manifolds U_g, S_g with $U_g \cap O$ negatively invariant and $S_g \cap O$ positively invariant under* (7.3) *such that $M_g = S_g \cap U_g \cap O$ and*

$$|u_t(\phi, g)| \leqq C|\phi|\, e^{\sigma t}, \qquad t \geqq 0, \quad \text{for } \phi \in (S_g \cap O),$$
$$|u_t(\phi, g)| \leqq C|\phi|\, e^{-\sigma t}, \qquad t \leqq 0, \quad \text{for } \phi \in (U_g \cap O)$$

for some constants C, $\sigma > 0$.

Proof. First, we introduce a system of coordinates around M_0 as indicated in § 5. For each fixed $\omega \in M_0$, the system in coordinate form (5.8) can be written as an equation (6.1)-(6.3), where we take for Λ the Banach space of bounded continuously differentiable functions from C into R^q which have bounded first derivative, taken with the uniform C^1-norm and take $\lambda = g$. As a consequence of Theorem 5.1, the hypotheses (H_1) to (H_4) of § 6 are all satisfied with $\alpha > k\alpha_0$. The only hypothesis which is necessary for applicability of the results in § 6 and is not necessarily fulfilled is that contained in (H_2'), namely that the functions n, u, s of $(t, x, y, \zeta, \lambda)$ are globally Lipschitzian in x, y, ζ, and the requirement that they admit a Lipschitz constant D satisfying (6.9). Consequently these functions have to be "cut-off" and replaced by functions $\bar{n}$, $\bar{u}$, $\bar{s}$ which agree with n, u, s for $|x|$, $|y|$, $|\zeta|$ sufficiently small and are globally Lipschitzian in x, y, ζ with a Lipschitz constant satisfying (6.9).

Let α_0, α, K be as in Theorem 5.1 and let $D > 0$ be chosen to satisfy (6.9). Assume $0 < \mu \leqq \mu_0$, $0 < \varepsilon \leqq \varepsilon_0$ and $B(\mu, \varepsilon)$ is as in hypothesis (H_2) of § 6. Let us consider a C^∞ function $\nu: R^+ \to [0, 1]$ such that

$$\nu(\rho) \in \begin{cases} \{1\} & \text{if } \rho/[\mu^2(2+r)] \leqq 1/4, \\ (0,1) & \text{if } \frac{1}{4} < \rho/[\mu^2(2+r)] < 1, \\ \{0\} & \text{if } 1 \leqq \rho/[\mu^2(2+r)] \end{cases}$$

and $0 \leqq -\nu'(\rho) \leqq 2/[\mu^2(2+r)]$ for $\rho \geqq 0$, and a C^∞ function $.R^+ \to [0, 1]$ such that

$$\sigma(\rho) \in \begin{cases} \{1\} & \text{if } \rho \leqq \varepsilon/2, \\ (0,1) & \text{if } \varepsilon/2 < \rho < \varepsilon, \\ \{0\} & \text{if } \varepsilon \leqq \rho. \end{cases}$$

We define the function $\bar{n}$: $R \times R^{d_N} \times R^{d_U} \times L \times \{\lambda \in \Lambda: \|\lambda\| \leqq \varepsilon_0\}$ so that it satisfies

$$\bar{n}(t, x, y, \zeta, \lambda) = \sigma(\varepsilon)\nu\left(|x|^2 + |y|^2 + \int_{-r}^{0} |\zeta(\tau)|^2\, d\tau\right) n(t, x, y, \zeta, \lambda)$$

for $t \in R$, $|x|$, $|y|$, $|\zeta| \leqq \mu$, $\|\lambda\| \leqq \varepsilon$, vanishing outside this region, and define $\bar{u}$, $\bar{s}$ in a similar way. Then, over the region $t \in R$, $|x|$, $|y|$, $|\zeta| \leqq \mu/2$, $\|\lambda\| \leqq \varepsilon/2$, we have $\bar{n} = n$,

$\bar{u}=u$, $\bar{s}=s$. It remains to show that $\bar{n}$, $\bar{u}$, $\bar{s}$ are globally Lipschitzian in x, y, ζ with Lipschitz constant D.

Let $t\in R$, $\|\lambda\|\leqq \varepsilon$ and $V=\{(x,y,\zeta)\in R^{d_N}\times R^{d_U}\times L\colon |x|,|y|,|\zeta|\leqq \mu\}$. If (x,y,ζ), $(\bar{x},\bar{y},\bar{\zeta})\in V$, then

$$\begin{aligned}
&|\bar{n}(t,x,y,\zeta,\lambda)-\bar{n}(t,\bar{x},\bar{y},\bar{\zeta},\lambda)|\\
&\leqq\left|\nu\left(|x|^2+|y|^2+\int_{-r}^{0}|\zeta(\tau)|^2\,d\tau\right)-\nu\left(|\bar{x}|^2+|\bar{y}|^2+\int_{-r}^{0}|\bar{\zeta}(\tau)|^2\,d\tau\right)\right||n(t,x,y,\zeta,\lambda)|\\
&\qquad+\nu\left(|\bar{x}|^2+|\bar{y}|^2+\int_{-r}^{0}|\bar{\zeta}(\tau)|^2\,d\tau\right)|n(t,x,y,\zeta,\lambda)-n(t,\bar{x},\bar{y},\bar{\zeta},\lambda)|\\
&\leqq\frac{2}{\mu^2(2+r)}2\mu(2+r)(|x-\bar{x}|+|y-\bar{y}|+|\zeta-\bar{\zeta}|)B(\mu,\varepsilon)3\mu\\
&\qquad+B(\mu,\varepsilon)(|x-\bar{x}|+|y-\bar{y}|+|\zeta-\bar{\zeta}|)\\
&=13B(\mu,\varepsilon)(|x-\bar{x}|+|y-\bar{y}|+|\zeta-\bar{\zeta}|).
\end{aligned}$$

If $(x,y,\zeta)\in V$ and $(\bar{x},\bar{y},\bar{\zeta})\notin V$ then there exists a point (x^*,y^*,ζ^*) lying in the intersection of the boundary of V and the straight line joining the points (x,y,ζ) and $(\bar{x},\bar{y},\bar{\zeta})$. Thus $\bar{n}(t,\bar{x},\bar{y},\bar{\zeta},\lambda)=\bar{n}(t,x^*,y^*,\zeta^*,\lambda)=0$ and

$$\begin{aligned}
|\bar{n}(t,x,y,\zeta,\lambda)-\bar{n}(t,\bar{x},\bar{y},\bar{\zeta},\lambda)|&=|\bar{n}(t,x,y,\zeta,\lambda)-\bar{n}(t,x^*,y^*,\zeta^*,\lambda)|\\
&\leqq 13B(\mu,\varepsilon)(|x-\bar{x}|+|y-\bar{y}|+|\zeta-\bar{\zeta}|).
\end{aligned}$$

If $(x,y,\zeta)\notin V$ and $(\bar{x},\bar{y},\bar{\zeta})\notin V$, then $\bar{n}(t,x,y,\zeta,\lambda)=\bar{n}(t,\bar{x},\bar{y},\bar{\zeta},\lambda)=0$. It follows that $\bar{n}$ is globally Lipschitzian in x, y, ζ with Lipschitz constant D, provided μ and ε are taken so small that $13B(\mu,\varepsilon)<D$.

The preceding reasoning also applies to $\bar{u}$, $\bar{s}$. Consequently, the functions $\bar{n}$, $\bar{u}$, $\bar{s}$ satisfy the hypothesis (H_2') of § 6 with a global Lipschitz constant D which satisfies (6.9). We are now in a situation where the theorems of § 6 can be applied to the system in coordinate form, with n, u, s replaced by $\bar{n}$, $\bar{u}$, $\bar{s}$. It remains to see what these theorems imply for the system (7.3) in the phase space C. For this we need to take into account the redundancy built into the system of coordinates introduced around the manifold M_0.

Each point of C lying close to M_0 is represented, in the system of coordinates around M_0 which was introduced in § 5, by a set of points (ω,x,y,ζ) which contains exactly one element with the second coordinate equal to zero. Because the integral manifold introduced in Theorem 6.2,

$$M_\lambda=\{(t,x,y,\zeta)\in R\times R^{d_N}\times R^{d_U}\times L\colon y=h_1(t,x,\lambda),\ \zeta=h_2(t,x,\lambda)\},$$

is, for the system in coordinate form, the maximal integral manifold contained in sets with the y, ζ-coordinates bounded, and because the functions n, u, s and $\bar{n}$, $\bar{u}$, $\bar{s}$ agree for x, y, ζ sufficiently small, it follows that there exists a neighborhood of zero $V\subset R\times R^{d_N}\times R^{d_U}\times L$ such that $M_\lambda\cap V$ is the maximal integral manifold contained in V. Therefore $M_\lambda\cap V$ represents in coordinate form a patch of a submanifold M_g of C which is invariant under (7.3). We recall that the system (7.3) is represented in coordinate form by a family of systems of the form (6.1)-(6.3), one for each $\omega\in M$ which is taken as initial condition for the solution of $\dot{u}(t)=f(u_t)$ used as center of the moving coordinate system. Based on this and on the redundancy built in the system of coordinates used, we can consider a function H defined from $M_0\times\{g\in BC^k(C;R^n)\colon \|g\|\leqq\varepsilon\}$ into C so that $H(\omega,g)$ is the point of C represented in coordinate form by $(\omega, 0, h_1^\omega(0,0,g), h_2^\omega(0,0,g))$ where $h_1 = h_1^\omega$, $h_2 = h_2^\omega$ are the functions considered above for M_λ for the particular system in coordinate form which corresponds to take the moving coordinate system centered on the solution of $\dot{u}(t) = f(u_t)$, $u_0 = \omega$. Then $M_g = \{H(\omega, g)\colon \omega \in M_0\}$, and its properties stated in the theorem follow directly from the theorems of § 6 about the properties of the functions h_1, h_2, if we recall the redundancy built in the system of coordinates, namely that changes of $h_1^\omega(0,0,g)$, $h_2^\omega(0,0,g)$ with ω can be identified with changes of $h_1(0,x,g)$, $h_2(0,x,g)$ with ω fixed and x changing.

The manifolds U_g and S_g can be treated in a similar way. QED

Remark. The persistence of hyperbolic invariant manifolds for FDEs was studied above for the case of retarded FDEs on euclidean space R^q. As indicated in § 4, the same result for retarded FDEs on a smooth, finite dimensional, separable and connected manifold follows from the result in euclidean space.

Acknowledgment. I wish to thank Professor George Sell for his helpful remarks and criticisms while this work was in progress.

REFERENCES

[1] W. A. COPPEL AND K. J. PALMER, *Averaging and integral manifolds*, Bull. Austral. Math. Soc., 2 (1970), pp. 197-222.

[2] N. FENICHEL, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971/72), pp. 193-226.

[3] J. K. HALE, *Ordinary Differential Equations*, Krieger Publishing Co., New York, 1980.

[4] ———, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[5] J. K. HALE, L. T. MAGALHÃES AND W. M. OLIVA, *An Introduction to Infinite Dimensional Dynamical Systems—Geometric Theory*, Springer-Verlag, New York, 1984.

[6] P. HARTMAN, *Ordinary Differential Equations*, Hartman, Baltimore, 1973.

[7] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics no. 840, Springer-Verlag, New York, 1981.

[8] M. HIRSCH, C. PUGH AND M. SHUB, *Invariant Manifolds*, Lecture Notes in Mathematics no. 583, Springer-Verlag, New York, 1977.

[9] J. KURZWEIL, *Exponentially stable integral manifolds, averaging principle and continuous dependence on a manifold*, Czechoslovak Math. J., 16 (1966), pp. 380-423, pp. 463-491.

[10] ———, *Invariant manifolds* I, Comment. Math. Univ. Carolin., 11 (1970), pp. 309-336.

[11] ———, *Global Solutions of Functional Differential Equations*, Lecture Notes in Mathematics no. 144, Springer-Verlag, New York, 1970.

[12] L. T. MAGALHÃES, *Invariant manifolds for functional differential equations close to ordinary differential equations*, Funkcial. Ekvac., 28 (1985), pp. 57-82.

[13] R. MAÑÉ, *Persistent manifolds are normally hyperbolic*, Bull. Amer. Math. Soc., 80 (1974), pp. 90-91.

[14] R. J. SACKER, *A perturbation theorem for invariant manifolds and Hölder continuity*, J. Math. Mech., 18 (1969), pp. 705-762.

[15] R. J. SACKER AND G. R. SELL, *Existence of exponential dichotomies and invariant splittings* I, II, III, J. Differential Equations, 15 (1974), pp. 429-458; 22 (1976), pp. 478-496, pp. 497-522.

[16] ———, *A spectral theory for linear differential systems*, J. Differential Equations, 27 (1978), pp. 320-358.

[17] ———, *The spectrum of an invariant submanifold*, J. Differential Equations, 38 (1980), pp. 135-160.

[18] ———, personal communication.

[19] G. R. SELL, *The structure of a flow in the vicinity of an almost periodic motion*, J. Differential Equations, 27 (1978), pp. 359-393.

[20] H. WHITNEY, *Differentiable manifolds*, Ann. Math., 37 (1936), pp. 645-680.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

008

ON PERIODIC SOLUTIONS OF A THERMOSTAT EQUATION*

GUSTAF GRIPENBERG†

Abstract. An equation describing the regulation of, for example, temperature with the aid of a thermostat is studied with emphasis on the question whether there are periodic solutions and if there can be several solutions with different periods.

Key words. thermostat, periodic solutions, Volterra equation

AMS(MOS) subject classifications. 93C10, 45D05

1. Introduction. The purpose of this paper is to study periodic solutions of equations describing, for example, temperature regulation by a thermostat. Therefore one considers a situation where a heater is turned ON if the temperature at some fixed point drops to a level θ_{low} and is turned OFF if the temperature rises to a level θ_{high} where one has $\theta_{\text{low}} < \theta_{\text{high}}$. Thus it will be a basic assumption in this paper that the control-function can only take two values. We will not go into further details about how the heating process affects the temperature since one can show that under quite reasonable assumptions one obtains an equation of the form

$$y(t) = \int_{-\infty}^{t} a(t-s)u(s)\,ds, \qquad t \in \mathbf{R} \tag{1}$$

where $y(t)$ is the normalized temperature at the thermostat at time t, $u(t)=1$ if the heater is turned ON at time t; otherwise $u(t)=0$ and a is an integrable real function on $\mathbf{R}^+ = [0, \infty)$ so that $\int_{\mathbf{R}^+} a(t)\,dt = 1$. The temperature has here been normalized in such a way that without heating it approaches 0 and with uninterrupted heating it approaches one. If one considers the heating process known for time $t<0$ and also allows some additive external influence on the temperature, then (1) becomes

$$y(t) = \int_0^t a(t-s)u(s)\,ds + f(t), \qquad t \in \mathbf{R}^+, \quad u(0) = u_0 \in \{0, 1\}. \tag{2}$$

For some derivations of equations like (1) and (2) where diffusion and other effects are taken into account, see e.g. [1] and [2].

The equations (1) and (2) do not involve the crucial part of the formulation of the problem, i.e., the thermostat control or the dependence of u on y. In this paper the relation is taken to be the following one: If $u(t)=1$ and y reaches the value θ_{high} at time t, then $u(t+)=0$ and conversely at the lower limit. Below we will also use a weaker formulation of this condition.

Concerning the question of existence of solutions we take a much more straightforward approach than the one used in [1] and [2], since here we avoid multivalued functions completely. The main problem addressed is therefore whether there exist periodic solutions and if there can be periodic solutions with different periods. Concerning the first question it is intuitively clear that periodic solutions should exist and concerning the second one a reasonable guess (that turns out to be correct) is that it all depends on the function a. The question of existence of periodic solutions is also considered in [3] but there much stronger assumptions are made on the function a and the limits θ_{high} and θ_{low}. It should, however, be pointed out that here only the

* Received by the editors October 31, 1985; accepted for publication (in revised form) May 13, 1986.

† Helsinki University of Technology, Institute of Mathematics, SF-02150 Espoo 15, Finland.

existence of so-called weak solutions (to be defined below) is established in the general case.

The stability of the periodic solutions will not be studied in this paper.

2. Statements of results. First we define the concepts related to the thermostat regulation.

DEFINITION 1. If $I \subset \mathbf{R}$ is an interval and $y: I \to \mathbf{R}$ is a continuous function, then a function $u: I \to \{0, 1\}$ is *weakly thermostat controlled* by y with respect to the higher limit θ_{high} and the lower limit θ_{low} on the interval I provided that u is left-continuous with right-hand limits on I, and for each $t \in I$ the following conditions hold:

(i) $u(t) = 1$ if $y(t) < \theta_{\text{low}}$,

(ii) $u(t) = 0$ if $y(t) > \theta_{\text{high}}$,

(iii) if $u(t) - u(t+) = 1$, then $\theta_{\text{high}} \in \{y(s) \mid s \in I, s \leqq t \text{ and } u(r) = 1 \text{ for all } r \in (s, t]\}$,

and

(iv) if $u(t) - u(t+) = -1$, then $\theta_{\text{low}} \in \{y(s) \mid s \in I, s \leqq t \text{ and } u(r) = 0 \text{ for all } r \in (s, t]\}$.

The function $u: I \to \{0, 1\}$ is *strictly thermostat controlled* by y with respect to the limits θ_{high} and θ_{low} if it is weakly thermostat controlled by y with respect to these limits and in addition for every $t \in I$ the following conditions hold:

(iii′) If $y(t) = \theta_{\text{high}}$ and $u(t) = 1$, then $u(t+) = 0$, and

(iv′) if $y(t) = \theta_{\text{low}}$ and $u(t) = 0$, then $u(t+) = 1$.

From this definition one sees that if u is weakly but not strictly thermostat controlled by y, then y must at some time have a local maximum equal to θ_{high} (or minimum equal to θ_{low}).

The reason for introducing the concept of a weakly thermostat controlled function is that in the general case we only obtain the existence of periodic solutions where u is weakly controlled by y. But this seems to be more of a mathematical than a real-world problem.

For convenience we define exactly what we mean by solutions.

DEFINITION 2. A pair (y, u) is a strict/weak thermostat controlled solution of (1) with respect to the limits θ_{high} and θ_{low} if $y \in C(\mathbf{R}; \mathbf{R})$, u is strictly/weakly thermostat controlled by y on $\mathbf{R}$ with respect to these limits and (1) holds. Similarly a pair (y, u) is a *strict/weak thermostat controlled solution* of (2) with respect to θ_{high} and θ_{low} if $y \in C(\mathbf{R}^+; \mathbf{R}^+)$, u is strictly/weakly thermostat controlled by y on $\mathbf{R}^+$ with respect to these limits and (2) holds.

Now we can give our basic existence result formulated for (2).

THEOREM 1. *Assume that $\theta_{\text{low}} < \theta_{\text{high}}$ are real numbers, $a \in L^1_{\text{loc}}(\mathbf{R}^+; \mathbf{R})$, $f \in C(\mathbf{R}^+; \mathbf{R})$ and let $u_0 \in \{0, 1\}$ be such that $u_0 = 1$ if $f(0) < \theta_{\text{low}}$ and $u_0 = 0$ if $f(0) > \theta_{\text{high}}$. Then there exists unique functions $y \in C(\mathbf{R}^+; \mathbf{R})$ and $u: \mathbf{R}^+ \to \{0, 1\}$ such that (y, u) is a strict thermostat solution of (2) with respect to the limits θ_{high} and θ_{low}.*

Observe that the theorem says nothing about the uniqueness of weakly thermostat controlled solutions; in fact it is easy to see that if there exists such a solution that is not thermostat controlled in the strict sense, then this solution cannot be unique.

Next we consider the question of existence of periodic solutions and here we use the further normalization that the $0 < \theta_{\text{low}} < \theta_{\text{high}} < 1$ and that $\int_{\mathbf{R}^+} a(s)\, ds = 1$.

THEOREM 2. *Assume that $0 < \theta_{\text{low}} < \theta_{\text{high}} < 1$ and that $a \in L^1(\mathbf{R}^+; \mathbf{R})$ satisfies $\int_{\mathbf{R}^+} a(s)\, ds = 1$. Then there exists periodic functions $y \in C(\mathbf{R}; \mathbf{R})$ and $u: \mathbf{R} \to \{0, 1\}$ (having the same period) so that (y, u) is a weak thermostat controlled solution of (1) with respect to the limits θ_{high} and θ_{low}.*

Next we investigate further some simple cases where we can prove that the periodic function u is strictly controlled by y. Note, however, that we in each case claim that every weak solution is in fact a strict one (which is not true in general), and not that there always exists a strict thermostat controlled solution (which one would expect to be the case for all kernels a).

THEOREM 3. *Let the assumptions of Theorem 2 hold. Assume in addition that a is nonincreasing on $(0, \infty)$ and not equal to a nonzero constant on any open interval. Then every periodic weak solution (y, u) of equation (1) is a strict solution.*

The previous result is quite trivial. The next one is slightly harder to prove. Recall that a function c is completely monotone if it is infinitely many times differentiable and $(-1)^j c^{(j)}(t) \geqq 0$ for all $j \geqq 0$.

THEOREM 4. *Let the assumptions of Theorem 2 hold. Assume in addition that $a(t) = \int_0^t a_1(t-s)a_2(s)\,ds$, $t > 0$ where a_1 and a_2 are locally integrable on $\mathbf{R}^+$, and completely monotone on $(0, \infty)$. Then every periodic weak solution (y, u) of (1) is a strict solution.*

Unfortunately, the assumption above on the function a is apparently not satisfied in the example given in [2], but if in that example the thermostat measures the temperature at an endpoint of the rod studied there, then it will be satisfied.

Finally we consider the question of existence of several periodic solutions with different minimal periods. This will be done only in the symmetric case $1-\theta_{\text{high}} = \theta_{\text{low}}$ and for the very special kernels $a_m(t) = (1/m!)t^m e^{-t}$ but these results can of course be slightly extended with the aid of perturbation arguments. ($\lfloor t \rfloor$ is the largest integer $\leqq t$.)

THEOREM 5. *Let $m \geqq 0$ and $a(t) = (1/m!)t^m e^{-t}$, $t \geqq 0$. Then there exist numbers $\alpha > 0$ such that (1) has $\lfloor m/4 \rfloor + 1$ different strict periodic thermostat controlled solutions (y, u) with respect to the limits $\frac{1}{2} + \alpha$ and $\frac{1}{2} - \alpha$.*

3. Proof of Theorem 1. We take $u(0) = u_0$, $y(0) = f(0)$. Next we define two sequences of numbers $\{\alpha_j\}_{j=0}^{\infty}$ and $\{t_j\}_{j=0}^{\infty}$ as follows: If $y(0) = \theta_{\text{high}}$ then $\alpha_0 = 0$ and if $y(0) = \theta_{\text{low}}$, then $\alpha_0 = 1$ and else $\alpha_0 = u_0$. Let $t_0 = 0$ and assume that we have already defined the numbers α_j and t_j for $j = 0, 1, \cdots, k$. As one would expect, we define

$$u(t) = \alpha_j \qquad t \in (t_j, t_{j+1}], \quad j \geqq 0.$$

Now let the function v_k be defined by

$$v_k(t) = \begin{cases} \int_0^t a(t-s)u(s)\,ds + f(t) & \text{when } t \in [0, t_k]; \\ \int_0^{t_k} a(t-s)u(s)\,ds + \int_{t_k}^t a(t-s)\alpha_k\,ds + f(t) & \text{when } t > t_k. \end{cases}$$

We see that v_k is a continuous function and we define t_{k+1} by

$$t_{k+1} \overset{\text{def}}{=} \inf\{t > t_k \mid \alpha_k(v_k(t) - \theta_{\text{high}}) + (1-\alpha_k)(v_k(t) - \theta_{\text{low}}) = 0\}.$$

Thus t_{k+1} is the first time $v_k(t)$ reaches θ_{high} if $\alpha_k = 1$ and otherwise the first time $v(t)$ reaches θ_{low}. Finally we define $\alpha_{k+1} = 1 - \alpha_k$. If $t_{k+1} = \infty$, then we do not have to choose any further points.

We define y to be equal to the right-hand side of (2) and we immediately see that u is strictly thermostat controlled by y on its domain of definition. It remains to prove that $\lim_{j\to\infty} t_j = \infty$. To see this we observe that when $j > 1$, then $|v_{j-1}(t_j) - v_{j-1}(t_{j-1})| = \theta_{\text{high}} - \theta_{\text{low}}$. Furthermore, since f is continuous, a is locally integrable and u only takes the values 1 and 0 we see that on each bounded interval the functions $\{v_j\}_{j=1}^{\infty}$ are equicontinuous and as $\theta_{\text{high}} > \theta_{\text{low}}$ this implies that $\inf_{t_j < T, j > 1}(t_j - t_{j-1}) > 0$. Therefore $\lim_{j\to\infty} t_j = \infty$. This completes the proof. □

We remark that one can have another notion of strict solution if one defines the switching times to be the last time before the temperature rises above the upper limit if the heater is ON or drops below the lower limit if the heater is OFF, that is, one would define the numbers $\{t_j\}$ by

$$t_{k+1} \overset{\text{def}}{=} \inf\{t > t_k \mid \alpha_k(v_k(t) - \theta_{\text{high}}) + (1-\alpha_k)(\theta_{\text{low}} - v_k(t)) > 0\}.$$

4. Proof of Theorem 2. The proof will be based on the following idea: Assume that u is given on $\mathbf{R}^-$ so that it is periodic with period $T+S$, $u(t)=0$ when $t \in (-S, 0]$ and $u(t)=1$ when $t \in (-T-S, -S]$. Now take $u(0+)=1$ and solve equation (1) with u as given on $\mathbf{R}^-$. At some T' the function y reaches θ_{high}, and the value of u is switched to zero. At some later time $T'+S'$ the function reaches θ_{low} again. If we can find a fixed-point of the mapping $(S, T) \mapsto (S', T')$, then we have also found a periodic solution. Now the problem is that this mapping could conceivably be discontinuous and therefore an approximation argument must be used.

Let us for the moment assume that

$$a \in C^1(\mathbf{R}^+; \mathbf{R}) \quad \text{and} \quad a' \in L^1(\mathbf{R}^+; \mathbf{R}). \tag{3}$$

For each pair of nonnegative numbers T and S, define the set $E_{T,S}$ as follows:

$$E_{T,S} = \bigcup_{m=-\infty}^{\infty} (m(S+T), m(S+T)+T] \tag{4}$$

so that $\chi_{E_{T,S}}$ is a function that is zero on intervals of length S and one on intervals of length T. Now define the function $\psi_{T,S}$ by

$$\psi_{T,S}(t) = \int_{-\infty}^{0} a(t-s)\chi_{E_{T,S}}(s)\, ds, \qquad t \in \mathbf{R}^+.$$

It follows from (3) that $\psi_{T,S}$ is Lipschitz-continuous. In fact the Lipschitz-constant will always be less than the number $c \overset{\text{def}}{=} \int_{\mathbf{R}^+} |a'(s)|\, ds + 1$.

Let ε be a number so that $0 < \varepsilon < \frac{1}{2}(\theta_{\text{high}} - \theta_{\text{low}})$ and define

$$f_h(\varepsilon, s) = c \max\left\{0, 1 - \frac{\theta_{\text{high}} - s}{\varepsilon}\right\};$$

$$f_l(\varepsilon, s) = c \min\left\{0, -1 - \frac{\theta_{\text{low}} - s}{\varepsilon}\right\}.$$

Now let v_ε be the solution of the equation

$$v_\varepsilon(t) = \int_0^t f_h(\varepsilon, v_\varepsilon(s))\, ds + \int_0^t a(s)\, ds + \psi_{T,S}(t), \qquad t \in \mathbf{R}^+.$$

It follows from standard results that this equation has a unique solution. If $v_\varepsilon(0) \geqq \theta_{\text{high}}$ then we define $T'=0$; otherwise we let

$$T' \overset{\text{det}}{=} \inf\{t > 0 \mid v_\varepsilon(t) = \theta_{\text{high}}\}.$$

To see that we always have $T' < \infty$ we have only to recall that $\psi_{T,S}(t) \to 0$ as $t \to \infty$, $f_h(\varepsilon, v_\varepsilon(s)) \geqq 0$ and that $\int_{\mathbf{R}^+} a(s)\, ds = 1 > \theta_{\text{high}}$.

To obtain S' we let w_ε be the solution of the equation

$$w_\varepsilon(t) = \int_{T'}^t f_l(\varepsilon, w_\varepsilon(s))\, ds + \int_0^{T'} a(t-s)\, ds + \psi_{T,S}(t), \qquad t \geqq T'$$

and define $S'=0$ if $w_\varepsilon(T')\leqq\theta_{\text{low}}$ and otherwise

$$S'\overset{\text{def}}{=}\inf\{t>T'\,|\,w_\varepsilon(t)=\theta_{\text{low}}\}-T'.$$

Again we conclude that the number S' must be finite.

Now we have constructed a mapping $G_\varepsilon:(T,S)\mapsto(T',S')$. To see that the mapping G_ε is continuous we have only to check that $\psi_{T,S}(t)$ depends continuously on T and S for each t, use a standard result about the continuous dependence of solutions of ordinary differential equations upon data and finally note that it follows from our definition of the functions f_h and f_l that $v_\varepsilon'(T')\geqq 1$ and $w_\varepsilon'(S')\leqq -1$.

We claim that there exist positive numbers δ and τ (independent of ε) such that G_ε maps the set $\{(T,S)\,|\,0\leqq T, S\leqq\tau, T+S\geqq\delta\}$ into itself. Clearly it is sufficient to choose τ so large that $\int_\tau^\infty|a(s)|\,ds<\min\{1-\theta_{\text{high}},\theta_{\text{low}}\}$. Furthermore we always have

$$\operatorname{var}\left(\psi_{T,S}(t)+\int_0^t a(s)\,ds;[0,T']\right)$$
$$+\operatorname{var}\left(\psi_{T,S}(t)+\int_0^{T'} a(t-s)\,ds;[T',T'+S']\right)\geqq\theta_{\text{high}}-\theta_{\text{low}}-2\varepsilon,$$

and hence we can see that we can choose $\delta\leqq(\theta_{\text{high}}-\theta_{\text{low}}-2\varepsilon)/\int_{\mathbf{R}^+}|a'(s)|\,ds$.

Now it follows from Brouwer's Fixed-Point theorem that the mapping $G_\varepsilon:(T,S)\mapsto(T',S')$ has a fixed point that we call $(T_\varepsilon,S_\varepsilon)$. It is easy to check that neither of these numbers can be zero.

We define the function u_ε to be the characteristic function of the set $E_{T_\varepsilon,S_\varepsilon}$ and let y_ε be defined by

$$y_\varepsilon(t)=\int_{-\infty}^t a(t-s)u_\varepsilon(s)\,ds,\qquad t\in\mathbf{R}.$$

Observe that u_ε is not thermostat controlled by y_ε. We do however know that

$$(5)\qquad \begin{aligned} &u_\varepsilon(t)<\theta_{\text{high}}, && t\in[0,T_\varepsilon],\\ &y_\varepsilon(t)>\theta_{\text{low}}, && t\in[S_\varepsilon,S_\varepsilon+T_\varepsilon],\end{aligned}$$

as one can easily see from the definition of the mapping $(T,S)\mapsto(T',S')$ and the fact that S_ε and T_ε are positive.

Next we let $\varepsilon\downarrow 0$ and observe that due to the compactness we may choose a subsequence $\{\varepsilon_j\}$ such that the numbers T_{ε_j} and S_{ε_j} converge towards some positive numbers T^* and S^*. We define the function u by

$$u(t)=\chi_{E_{T^*,S^*}}(t),\qquad t\in\mathbf{R},$$

and

$$y(t)=\int_{-\infty}^t a(t-s)u(s)\,ds,\qquad t\in\mathbf{R}.$$

Now it is quite obvious that the functions u_{ε_j} converge in $L^1_{\text{loc}}(\mathbf{R};\mathbf{R})$ towards the function u and also that the functions y_{ε_j} converge uniformly on compact subsets towards the function y. It remains to prove that u is weakly thermostat controlled by y. It follows from inequalities (5) that we must have $y(t)\leqq\theta_{\text{high}}$ on $[0,T^*]$ and $y(t)\geqq\theta_{\text{low}}$ on $[T^*,T^*+S^*]$. If on the other hand there is no point $t\in[0,T]$ for which $y(t)=\theta_{\text{high}}$, then it follows that for sufficiently small numbers ε_j we have $y_{\varepsilon_j}(t)\leqq\theta_{\text{high}}-\varepsilon_j$ for all $t\in[0,T_{\varepsilon_j}]$. This implies that $v_{\varepsilon_j}(t)=y_{\varepsilon_j}(t)$ for all $t\in[0,T_{\varepsilon_j}]$ which is impossible. A similar argument for the lower bound θ_{low} shows that we actually do have a weakly controlled solution.

It remains to remove the differentiability assumption and this can be done by approximating the original kernel with differentiable ones. Thus one gets a sequence of weak solutions and an argument similar to the one used above shows that it must have a limit that is a weakly controlled periodic solution of equation (1) on **R**. □

5. Proof of Theorem 3. Let $T > S$ be two points such that $u(s) = 1$ when $s \in (S, T]$ and $u(S) = 0$. Let $\tau \in (S, T)$ be an arbitrary point. From (1) we have

$$y(T) - y(\tau) = \int_0^{T-\tau} a(s)\,ds - \int_{\mathbf{R}^+} (a(s) - a(T-\tau+s))u(\tau - s)\,ds. \tag{6}$$

Since $\int_{\mathbf{R}^+} (a(s) - a(T-\tau+s))\,ds = \int_0^{T-\tau} a(s)\,ds$ this equation shows that y is nondecreasing on the intervals where u is 1 (and by the same argument nonincreasing when u is 0). If $a(s) = 0$ when $s > \tau - S$, then $y(\tau) = 1$ which is impossible. Hence it follows from our assumptions that $a(s) - a(T-\tau+s) > 0$ if $s \in (\tau - S, \tau - S + \varepsilon)$ for some positive number ε. But this implies by (6) and the fact that $u(S) = 0$ that

$$y(T) - y(\tau) > \int_0^{T-\tau} a(s)\,ds + \int_{\mathbf{R}^+} (a(s) - a(T-\tau+s))\,ds = 0$$

and since τ was arbitrary it follows that u is strictly controlled by y at the upper limit. A similar argument can be applied to the lower limit to complete the proof. □

6. Proof of Theorem 4. We consider first the case when the functions a_1 and a_2 are of the form $a_1(t) = \sigma_1 e^{-\sigma_1 t}$ and $a_2(t) = \sigma_2 e^{-\sigma_2 t}$ for some positive numbers σ_1 and σ_2. Assume that u takes the value one on intervals of length T and is equal to zero on intervals of length S, that is $u(t) = \chi_{E_{T,S}}(t)$ where $E_{T,S}$ is given in (4). Let us denote

$$v(t) = \int_{\mathbf{R}^+} a_1(s)u(t-s)\,ds, \qquad t \in \mathbf{R}.$$

Now it is obvious that v is increasing on the intervals where u takes the value 1 and decreasing elsewhere. Differentiating both sides of (1) we have

$$y'(t) = \sigma_2(v(t) - y(t)), \qquad t \in \mathbf{R}.$$

We see that on the intervals where u is 1 the function y can have only one point where the derivative vanishes and that point must be a minimum. Hence we conclude that

$$y(t) < y(T), \qquad t \in [0, T], \tag{7}$$

provided we can show that $y(0) < y(T)$. Suppose for the moment that this has been done. Since completely monotone functions are limits of sums of exponentials we see that the relation (7) must hold in the general case too and hence the solution (y, u) is a strict solution.

Thus it remains to prove that $y(0) < y(T)$. We know that

$$v'(t) = \sigma_1(u(t) - v(t))$$

and since v is periodic with period $T + S$ one can easily solve the differential equation above and from the periodicity condition conclude that

$$v(0) = e^{-\sigma_1 S} \frac{1 - e^{-\sigma_1 T}}{1 - e^{-\sigma_1 (T+S)}},$$

$$v(T) = \frac{1 - e^{-\sigma_1 T}}{1 - e^{-\sigma_1 (T+S)}}.$$

Now we proceed to solve the function y and using the fact that this function is periodic too, we get

$$y(0)=(1-e^{-\sigma_2 T})+e^{-\sigma_2 T}y(0)-(1-v(0))\frac{\sigma_2}{\sigma_2-\sigma_1}(e^{-\sigma_1 T}-e^{-\sigma_2 T}),$$

$$y(T)=e^{-\sigma_2 S}y(T)+v(T)\frac{\sigma_2}{\sigma_2-\sigma_1}(e^{-\sigma_1 S}-e^{-\sigma_2 S}).$$

(Here we assume that $\sigma_2 \neq \sigma_1$, the case when they are equal is treated in the same manner.) If we solve $y(0)$ and $y(T)$ from these equations and insert the values of $v(0)$ and $v(T)$ obtained above, then we get the relation

$$\begin{aligned}(1-e^{-\sigma_1(T+S)})&(1-e^{-\sigma_2(T+S)})(y(T)-y(0))\\ &=(1-e^{-\sigma_2 T})(1-e^{-\sigma_2 S})(1-e^{-\sigma_1(T+S)})\\ &\quad-(1-e^{-\sigma_1 T})(1-e^{-\sigma_2 T})\frac{\sigma_2}{\sigma_2-\sigma_1}(e^{-\sigma_1 S}-e^{-\sigma_2 S})\\ &\quad-(1-e^{-\sigma_1 S})(1-e^{-\sigma_2 S})\frac{\sigma_2}{\sigma_2-\sigma_1}(e^{-\sigma_1 T}-e^{-\sigma_2 T}).\end{aligned} \tag{8}$$

Next write

$$1-e^{-\sigma_1(T+S)}=\tfrac{1}{2}(1-e^{-\sigma_1 T})(1+e^{-\sigma_1 S})+\tfrac{1}{2}(1-e^{-\sigma_1 S})(1+e^{-\sigma_1 T}) \tag{9}$$

and then observe that it is easy to verify that

$$\begin{aligned}&\frac{1}{2}(1-e^{-\sigma_2 S})(1+e^{-\sigma_1 S})-\frac{\sigma_2}{\sigma_2-\sigma_1}(e^{-\sigma_1 S}-e^{-\sigma_2 S})\\ &\quad=e^{-(\sigma_1+\sigma_2)S/2}\left(\sinh\left(\frac{(\sigma_1+\sigma_2)S}{2}\right)-\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}\sinh\left(\frac{(\sigma_1-\sigma_2)S}{2}\right)\right)>0.\end{aligned}$$

A similar result holds of course with S replaced by T and if these together with (9) are used in the right-hand side of the relation (8), then we obtain the desired conclusion that $y(T)>y(0)$. □

7. Proof of Theorem 5. Let T be positive, $a_m(t)=(1/m!)t^m e^{-t}$ and define

$$u_T(t)=\begin{cases}1 & \text{if } t\in(mT, mT+T],\ m \text{ even},\\ 0 & \text{otherwise.}\end{cases}$$

Let

$$y_{m,T}(t)\overset{\text{def}}{=}\int_{\mathbf{R}^+} a_m(s)u_T(t-s)\,ds, \qquad t\in\mathbf{R}$$

and

$$h_m(T)=\frac{1}{m!}(-1)^m T^{m+1}\frac{d^m}{dT^m}\left(\frac{1}{T}\frac{1}{e^T+1}\right), \qquad T>0.$$

A straightforward calculation shows that

$$y_{m,T}(0)=h_m(T), \qquad m\geqq 0, \quad T>0.$$

We need some further information about the function h_m.

LEMMA 1. *Let h_m be the function defined above. Then for each $m \geqq 0$, $h_m(0)=\frac{1}{2}$ and*

(i) $(-1)^{\lfloor m/2 \rfloor}(h_m(T)-\frac{1}{2})$ *is negative and decreasing when $T \in (0, \delta_m)$, $\delta_m > 0$,*

(ii) $h_{m+1}(T)=h_m(T)-(1/(m+1))Th'_m(T)$ *when $T>0$,*

(iii) *the function h_m has $\lfloor m/2 \rfloor$ local extrema on $(0, \infty)$ and no other points where the derivative vanishes.*

Proof. Let us first define the function $f: \mathbf{R}^+ \to \mathbf{R}^+$ by $f(T)=1/(e^T+1)$. A calculation shows that

$$h'_m(T)=\frac{1}{m!}(-1)^m T^m f^{(m+1)}(T), \qquad T \geqq 0$$

and from this relation we get the claim (ii) with the aid of a partial integration.

Next let us define

$$g_0(s) \overset{\text{def}}{=} s(1-s), \qquad s \in \mathbf{R},$$

$$g_m(s) \overset{\text{def}}{=} s(1-s)g'_{m-1}(s), \qquad m \geqq 1, \quad s \in \mathbf{R}.$$

Again it is straightforward to check that

$$h'_m(T)=-\frac{1}{m!}T^m g_m\left(\frac{1}{e^T+1}\right), \quad T \geqq 0.$$

By induction it is easy to prove that the function g_m has exactly $\lfloor m/2 \rfloor$ zeros in the interval $(0, \frac{1}{2})$ and all of these zeros are simple. This gives the assertion (iii) about the number of extreme points. To establish (i) we have only to observe that

$$g_m(\tfrac{1}{2})\begin{cases} >0 & \text{if } m \equiv 0 \bmod 4, \\ =0 & \text{if } m \equiv 1 \bmod 4, \\ <0 & \text{if } m \equiv 2 \bmod 4, \\ =0 & \text{if } m \equiv 3 \bmod 4, \end{cases} \qquad g'_m(\tfrac{1}{2})\begin{cases} =0 & \text{if } m \equiv 0 \bmod 4, \\ <0 & \text{if } m \equiv 1 \bmod 4, \\ =0 & \text{if } m \equiv 2 \bmod 4, \\ >0 & \text{if } m \equiv 3 \bmod 4. \end{cases}$$

This completes the proof of Lemma 1. □

Our next lemma says for how many values of T we get the same value for $y_{m,T}(0)$.

LEMMA 2. *Let $y_{m,T}$ be the function defined above. If $m \geqq 0$ and α is a sufficiently small positive number, then there exists $\lfloor m/4 \rfloor + 1$ different values T_j so that $y_{m,T_j}(0)=\frac{1}{2}-\alpha$ and $y_{m-1,T_j}(0)<\frac{1}{2}-\alpha$ if $m \geqq 1$.*

Proof. From Lemma 1 (i) and (iii) and the fact that $h_m(\infty)=0$ we see that the extreme point of h_2 must be a maximum. Then it is possible, using induction and (ii) in Lemma 1 to show that at all local maxima of $h_m(T)$ the value of the function is larger than $\frac{1}{2}$ and at all the minima smaller. Furthermore it follows from (i) and (iii) in Lemma 1 that h_m has exactly $\lfloor m/4 \rfloor$ local maxima on $\mathbf{R}^+$ (note that for some values of m there is one at 0). But then the desired conclusion follows directly from (ii) in Lemma 1 since immediately to the right of each maximum there is a point where $h_m(T)=\frac{1}{2}-\alpha$ and $h'_m(T)<0$. □

If now $\theta_{\text{low}}=\frac{1}{2}-\alpha$ and $\theta_{\text{high}}=1-\theta_{\text{low}}$, $m \geqq 0$ and T is such that $y_{m,T}(0)=\frac{1}{2}-\alpha$, then $(y_{m,T}, u_T)$ is a periodic strict solution of equation (1) if we can show that

$$\max_{s \in [0,T]} y_{m,T}(s)=y_{m,T}(T)>y_{m,T}(t), \qquad t \in [0, T). \tag{10}$$

This is of course the case if $m=0$ and we will show that it is also true if $y_{m,T}(0)>y_{m-1,T}(0)$. First we need a simple lemma.

LEMMA 3. *Let $m \geqq 0$, $T > 0$ and let $y_{m,T}$ be as defined above. Then $y_{m,T}$ can have at most one local extreme point on the interval* $(0, T)$.

Proof. We use induction to prove the stronger claim that there cannot be two extreme points in $(0, T)$ or one extreme point if the derivative is zero at 0. It obviously holds for $m = 0$. Assume that it also holds for $m = k-1$. We will show that it is true for $m = k$. Suppose that $y_{k,T}$ has at least two local extreme points in $(0, T)$. Since

$$y'_{k,T}(t) = y_{k-1,T}(t) - y_{k,T}(t), \tag{11}$$

and by symmetry

$$y'_{k-1,T}(0) = -y'_{k-1,T}(T),$$

this assumption implies that $y_{k-1,T}$ must have two extreme points in $(0, T)$ or one extreme point and vanishing derivative at zero. If $y_{k,T}$ has derivative equal to zero and one extreme point on $(0, T)$ then a similar argument shows that we again get a contradiction. □

To complete the proof of the theorem we see that if (10) is not satisfied then either $y'_{m,T}(0) \geqq 0$ which is impossible if $y_{m,T}(0) > y_{m-1,T}(0)$ by (11) or there are at least two local extreme points on the interval $(0, T)$. But this is impossible according to Lemma 3 and therefore the desired conclusion follows from Lemma 2. □

Acknowledgment. I would like to thank Dr. J. Prüss for introducing me to the problems considered in this paper.

REFERENCES

[1] K. GLASHOFF AND J. SPREKELS, *An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats*, this Journal, 12 (1981), pp. 477–486.

[2] ———, *The regulation of temperature by thermostats and set-valued integral equations*, J. Integral Equations, 4 (1982), pp. 95–112.

[3] J. PRÜSS, *Periodic solutions of the thermostat problem*, unpublished manuscript.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

EXISTENCE OF GLOBAL SOLUTIONS TO A MODEL OF A MYELINATED NERVE AXON*

CHRIS COSNER†

Abstract. The main result is a proof of the existence of solutions which are global in time for a differential equation arising in the theory of myelinated nerves. The equation differs from the usual reaction-diffusion equations occurring in nerve models in that the second derivative operator in the spatial direction is replaced at a sequence of discrete nodes by an operator defined as the jump in the first derivative in the spatial direction across the node. The nonlinear dynamics are assumed to be concentrated at the nodes. Local existence of solutions is obtained via semigroup theory; global bounds and hence global existence via energy or Lyapunov methods.

Key words. nerve axon models, global existence theory, differential/difference operators, semigroup theory, Lyapunov functions

AMS(MOS) subject classifications. 35R10, 35K57, 92A09

1. Introduction. The study of mathematical models of nerve impulse conduction has proved to be a fruitful source of interesting problems in differential equations. The purpose of this article is to prove the global existence of solutions to one such model that arises from a consideration of the effects of myelination on conduction of impulses in a nerve axon. Myelinated nerves, which include many human nerve processes, are wrapped in periodic bands of fatty myelin tissue, which acts as an insulator in damping the dynamics of the nerve axon. The dynamics of the myelinated nerve occur primarily at the gaps (known as nodes of Ranvier) between the myelin bands. The model considered here is the myelinated analogue of the simplest model for conduction in a uniform axon, the so-called reduced FitzHugh-Nagumo equation

$$\begin{aligned} &u_t = u_{xx} + f(u) \quad \text{on } (0,\infty)\times(0,\infty),\\ &u_x(0+,t) = I(t),\\ &u(x,0) = \phi(x) \end{aligned} \tag{1.1}$$

where u represents the potential across the axonal membrane, $I(t)$ a current stimulus at the end of the axon, $\phi(x)$ the initial state of the axon, and $f(u)$ describes the dynamics of the excitable membrane. The corresponding myelinated model, which we shall study in the present article, assumes nodes at integer values of x and is given by

$$\begin{aligned} &U_t = U_{xx} - gU, \quad x \in (0,\infty)\backslash\mathbb{Z}^+, \quad t>0,\\ &V_t^j = [\![U_x]\!]_j + f(V^j), \quad j\in\mathbb{Z}^+, \quad t>0,\\ &U_x(0+,t) = I(t),\\ &(U,V)|_{t=0} = (\phi_0, \psi_0) \end{aligned} \tag{1.2}$$

where $V=(V^j)_{j=1}^{\infty}$ and $[\![F]\!]_j = F(j+) - F(j-)$. In (1.2), the myelin is assumed to have the effect of limiting the dynamics of the nerve to the nodes. A discussion of the modeling leading to (1.2) is given in [2]. References on (1.2) and on models related

* Received by the editors November 4, 1985; accepted for publication April 1, 1986. This research was supported in part by National Science Foundation grant MCS-8108814 (A04) and the Institute for Advanced Study.

† Department of Mathematics and Computer Science, University of Miami, Coral Gables, Florida 33124.

to (1.2) are given in [2]. Another class of models for myelinated nerves is discussed in [6]. We shall impose the following conditions on f:

$$f(0)=0, \qquad f\in C^1(\mathbb{R}) \quad \text{with } f'(x)\leqq f_0 \quad \text{on } \mathbb{R}, \tag{1.3}$$

and assume that f can be extended to a function on $\mathbb{C}$ which, if viewed as a function of two real variables, is C^1. In the actual model, f is often taken to be a cubic $f(u)=bu(a-u)(u-1)$, $b>0$, $a\in(0,1)$.

For purposes of applications, the most interesting questions about (1.2) involve threshold phenomena and propagation of impulses; those questions for the pure initial value problem are discussed in [2] and for the initial boundary problem (1.2) in work in progress by Jonathan Bell and the author. However, it is not entirely satisfactory from a mathematical point of view to leave the question of existence of solutions open; thus the question of existence is discussed here. The main result is the existence theorem for (1.2) stated at the end of § 3. To obtain a local existence result, we cast (1.2) in terms of the operator A given by $A(u,v)=(-u_{xx}+u,(-[\![u_x]\!]_j+v^j)_{j=1}^{\infty})$, where as in (1.2) u is defined on $(0,\infty)\backslash\mathbb{Z}^+$ and $v=(v^j)_{j=1}^{\infty}$, then show that when given the appropriate domain the operator A is such that abstract results from the theory of semilinear evolution equations can be applied to (1.2). The appropriate spaces on which to consider A are slightly nonstandard; those spaces and the properties of A are studied in § 2. To obtain global existence in time, a priori bounds on solutions are needed; those bounds are obtained by a Lyapunov or energy method in § 3. The analysis is somewhat similar to those done for various models of uniform (i.e., nonmyelinated) axons in [1], [3], [8]. The analysis can be extended to models more complicated than (1.2); for example it is clear from an examination of the proof that the scalar equation $V_t^j=[\![U_x]\!]_j+f(V^j)$ could be replaced by a vector equation of the form

$$\vec{V}_t^j=\vec{P}_0[\![U_x]\!]_j+\vec{f}(\vec{V}^j,\vec{W}^j), \qquad \vec{W}_t^j=\vec{h}(\vec{V}^j,\vec{W}^j)$$

(where $\vec{P}_0$ is a constant vector with positive components) with only minor changes in the analysis under appropriate hypotheses on $\vec{f}$ and $\vec{h}$. Thus, we could replace the reduced Fitzhugh-Nagumo dynamics with Hodgkin-Huxley or other more complicated dynamics at each node. The mathematical novelty of (1.2) comes mainly from the existence of nodes at which $\partial^2/\partial x^2$ is replaced by $[\![\partial^{\cdot}/\partial x]\!]$. Thus we limit our attention to (1.2) for the sake of brevity and simplicity; our methods are adequately illustrated by that simple case. In studying the operator A, we use some rather "soft" or nonconstructive methods. The author would be very interested in seeing an explicit construction of a Green's function for the linear problem $(u,v)_t+A(u,v)=(p,q)$, $(u,v)|_{t=0}=(\phi,\psi)$ with (p,q), (ϕ,ψ) given; such a construction appears to be computationally tricky. Another approach to (1.2) might be via some sort of method of lines; however, obtaining the appropriate estimates for the convergence of line methods also seems to be a delicate problem. Thus, although the methods used here are nonconstructive, they have the advantage of simplicity.

2. Linear theory. Our analysis will use some slightly unusual spaces, all consisting of pairs (p,q) where p is a complex valued function on $(0,\infty)\backslash\mathbb{Z}^+$ and $q=(q^j)_{i=1}^{\infty}$ with $q^j\in\mathbb{C}$. Let $H=\{(p,q)\colon p\in L^2[(0,\infty)\backslash\mathbb{Z}^+], q\in l^2\}$ with inner product

$$\langle (p_1,q_1),(p_2,q_2)\rangle_H=\int_{(0,\infty)\backslash\mathbb{Z}^+} p_1\bar{p}_2+\sum_{j=1}^{\infty} q_1^j\bar{q}_2^j.$$

The space H is a complex Hilbert space and can be viewed as an L^2 space generated by the appropriate measure. Let

$$X=\{(p,q)\in H\colon p',p''\in L^2[(0,\infty)\backslash\mathbb{Z}^+],\ p'\text{ absolutely continuous on }(j-1,j)\text{ for } j\in\mathbb{Z}^+,\ (\llbracket p'\rrbracket_j)_{j=1}^{\infty}\in l^2,\ \lim_{x\to j}p(x)=q^j,\text{ and } p'(0+)=0\}.$$

It is a fairly straightforward exercise in the theory of functions of a real variable to show that X is dense in H. Finally, let

$$Y=\{(p,q)\in H\colon p'\in L^2[(0,\infty)\backslash\mathbb{Z}^+],\ p\text{ absolutely continuous on }(0,\infty)\backslash\mathbb{Z}^+,\ \lim_{x\to j}p(x)=q^j\}$$

with inner product

$$\langle (p_1,q_1),(p_2,q_2)\rangle_Y=\int_{(0,\infty)\backslash\mathbb{Z}^+}(p_1'\bar{p}_2'+p_1\bar{p}_2)+\sum_{j=1}^{\infty}q_1^j\bar{q}_2^j.$$

Clearly $X\subseteq Y\subseteq H$, with $\|(u,v)\|_H\leqq\|(u,v)\|_Y$ for $(u,v)\in Y$. Let A be the operator defined by $A(p,q)=(-p''+p,-\llbracket p'\rrbracket_j+q^j)_{j=1}^{\infty}$, with $\operatorname{dom}A=X$. We will show that A generates an analytic semigroup; to see that, we will first show that A has a bounded inverse so that A is closed (since $\operatorname{dom}A=X$ is dense in H), then calculate the numerical range to see that A is m-sectorial so that A generates a holomorphic semigroup. Inverting A is equivalent to solving

$$\begin{aligned} A(u,v)&=(p,q)\in H, \qquad (u,v)\in X\\ -u''+u&=p \quad \text{on } (0,\infty)\backslash\mathbb{Z}^+,\\ -\llbracket u'\rrbracket_j+v^j&=q^j, \qquad j\in\mathbb{Z}^+,\\ u'(0+)&=0 \end{aligned} \tag{2.1}$$

in X with $\|(u,v)\|_H\leqq K\|(p,q)\|_H$ for some constant K. To solve (2.1) we first obtain a generalized solution in Y and then show that such a solution must belong to X.

We define a generalized solution of (2.1) to be an element $(u,v)\in Y$ such that for any $(w,y)\in Y$,

$$\langle (w,y),(u,v)\rangle_Y=\langle (w,y),(p,q)\rangle_H. \tag{2.2}$$

Since $|\langle (w,y),(p,q)\rangle_H|\leqq\|(w,y)\|_H\|(p,q)\|_H\leqq\|(w,y)\|_Y\|(p,q)\|_H$, the right side of (2.2) defines a bounded linear functional on $(w,y)\in Y$. By the Riesz representation theorem, there exists a unique $(u,v)\in Y$ such that (2.2) holds for any $(w,y)\in Y$. Thus, for any given $(p,q)\in Y$ we can define $B\colon H\to H$ by taking $B(p,q)$ to be the generalized solution to (2.2) corresponding to (p,q). If $(u,v)=B(p,q)\in Y$, then by (2.2), $\|B(p,q)\|_Y^2=|\langle (u,v),(u,v)\rangle_Y|=|\langle (u,v),(p,q)\rangle_H|\leqq\|(u,v)\|_Y\|(p,q)\|_H=\|B(f,g)\|_Y\|(p,q)\|_H$ so that $\|B(p,q)\|_Y\leqq\|(p,q)\|_H$. Hence $B\colon H\to Y\hookrightarrow H$ is bounded. It remains to show that B maps H into X and that any solution (u,v) to (2.2) in X satisfies $A(u,v)=(p,q)$. Showing that B maps H to X is essentially a matter of proving the regularity of generalized solutions to (2.1). Let $H^k[\Omega]$ denote the L^2-Sobolev space of functions with k weak derivatives in $L^2[\Omega]$. If $(u,v)\in Y$ then $u|_{(j-1,j)}\in H^1[(j-1,j)]$ for all $j\in\mathbb{Z}^+$. We may follow the analysis in [4, § I.15] to conclude that $u|_{(a,b)}\in H^2[(a,b)]$ for any a,b with $j-1<a<b<j$, $j\in\mathbb{Z}^+$, provided we can show that if $(u,v)\in Y$ satisfies (2.2) for a given $(p,q)\in H$, there exists a constant C such that for any $\phi\in C_0^\infty[(j-1,j)]$ we have

$$\left|\int_{(j-1,j)}\phi'\bar{u}'\right|\leqq C\|\phi\|_{L^2[(j-1,j)]}. \tag{2.3}$$

(Inequality (2.3) is essentially inequality (15.6) in [4, § I.15]; the verification that the results in [4] carry over to our situation is routine and hence omitted.) Suppose that $(u, v) \in Y$ satisfies (2.2) and $\phi \in C_0^\infty[(j-1, j)]$ with $\operatorname{supp} \phi \subseteq [a, b] \subseteq (j-1, j)$. If we extend ϕ to $(0, \infty)\backslash\mathbb{Z}^+$ by taking $\phi \equiv 0$ outside $[a, b]$ and let $\psi = (\psi^k)_{k=1}^\infty$ with $\psi^k = 0$ for all k, then $(\phi, \psi) \in Y$ and (2.2) yields

$$\int_{(j-1,j)} \phi' \bar{u}' = \int_{(j-1,j)} \phi(\bar{p} - \bar{u}) \tag{2.4}$$

so that

$$\begin{aligned} \left| \int_{(j-1,j)} \phi' \bar{u}' \right| &= \left| \int_{(j-1,j)} \phi(\bar{p} - \bar{u}) \right| \\ &\leqq (\|(p, q)\|_H + \|B(p, q)\|_H) \|\phi\|_{L^2[(j-1,j)]}, \end{aligned} \tag{2.5}$$

which establishes (2.3) since $B: H \to H$ is bounded. Hence, $u''|_{(a,b)} \in L^2[(a, b)]$ for any $[a, b] \subseteq (j-1, j)$, so by taking complex conjugates in (2.4) we have

$$-\int_{(a,b)} u'' \bar{\phi} = \overline{\int_{(a,b)} \phi' \bar{u}'} = \int_{(a,b)} \bar{\phi}(p - u)$$

for any $\phi \in C_0^\infty[(a, b)]$. Since C_0^∞ is dense in L^2, we have for any $h \in L^2[(a, b)]$

$$\int_{(a,b)} (-u'')\bar{h} = \int_{(a,b)} (p - u)\bar{h}.$$

Thus, any bounded linear functional acting on L^2 gives the same result for $-u''$ as for $p - u$, so $-u'' = p - u$ or $-u'' + u = p$ on (a, b). Since a and b may be chosen arbitrarily as long as $j - 1 < a < b < j$, and $(p - u)|_{(j-1,j)} \in L^2[(j-1, j)]$, we have $-u'' = p - u$ on $(j-1, j)$. Since $j \in \mathbb{Z}^+$ was arbitrary, $-u'' + u = p$ on $(0, \infty)\backslash\mathbb{Z}^+$ and $u'' \in L^2[(0, \infty)\backslash\mathbb{Z}^+]$, which establishes the first equation in (2.1). Since $L^2[(j-1, j)] \subseteq L^1[(j-1, j)]$ it follows that u' is absolutely continuous on $(j-1, j)$, and that $u'(b) - u'(a) = \int_{(a,b)} u''$ for any $(a, b) \subseteq (j-1, j)$. Hence we may let $a \to j-1$ or $b \to j$ to see that $u'(j-1+)$ and $u'(j-)$ are well defined. Also, since u and u'' belong to $L^2[(0, \infty)\backslash\mathbb{Z}^+]$, so does u'. Since $(u, v) \in Y$ it remains only to verify the last two equations in (2.1). Given $j \in \mathbb{Z}^+$, define $(w_j, y_j) \in Y$ by $y_j^k = \delta_{jk}$ and $w_j' = (1/h)\chi_{(j-h,j)}$ on $(j-h, j)$, $w_j' = -(1/h)\chi_{(j,j+h)}$ on $(j, j+h)$, and $w_j' = 0$ otherwise, where $h > 0$ is arbitrary; then $w_j = (1/h)(x - j + h)$ on $(j-h, j)$, $w_j = -(1/h)(x - j - h)$ on $(j, j+h)$, and $w_j = 0$ otherwise. Clearly $(w_j, y_j) \in Y$, and (2.2) yields

$$\begin{aligned} (1/h)\int_{(j-h,j)} u' + (1/h)\int_{(j-h,j)} (x-j+h)u - (1/h)\int_{(j,j+h)} u' \\ -(1/h)\int_{(j,j+h)} (x-j-h)u + v^j \\ = (1/h)\int_{(j-h,j)} (x-j+h)p - (1/h)\int_{(j,j+h)} (x-j-h)p + q^j. \end{aligned} \tag{2.6}$$

Noting that $(1/h)\int_{(j-h,j)} u' = (1/h)[u(j) - u(j-h)] \to u'(j-)$ and $(1/h)\int_{(j,j+h)} u' \to u'(j+)$ as $h \to 0$, that $|(1/h)\int_{(j-h,j)} (x-j+h)u| \leqq \int_{(j-h,j)} |u| \to 0$ as $h \to 0$ and similarly for the remaining integral terms in (2.6), we may let $h \to 0$ in (2.6) to obtain the equation $u'(j-) - u'(j+) + v^j = q^j$, which is equivalent to the second equation in (2.1), and implies that $([\![u']\!]_j)_{j=1}^\infty = v^j - q^j \in l^2$.

The last equation in (2.1) follows as above by taking $w=(1/h)(-x+h)$ on $(0, h)$, $0<h<1$, $w=0$ otherwise, $y^k=0$ for all k, using (2.1) and letting $h\to 0$.

We have shown that $B: H\to X\subseteq H$ so that $B(p, q)$ satisfies $AB(p, q)=(p, q)$ for any $(p, q)\in H$. Also, $B: H\to H$ is bounded and by our construction $BA(u, v)=(u, v)$ for $(u, v)\in X$. Since $B=A^{-1}$ is a bounded operator defined on all of H and $X=\operatorname{dom} A$ is dense in H, it follows that A is closed. (See the discussion in [7, § III.5.2].) We have thus proved the following:

LEMMA 1. *The operator A defined by $A(u, v)=(-u''+u, (-[\![u']\!]_j+v^j)_{j=1}^{\infty})$ with* $\operatorname{dom} A=X$ *is a closed operator from X to H with $A^{-1}: H\to H$ bounded.*

To apply semigroup theory in our nonlinear problem, we will show that A satisfies

$$\|(\lambda-A)^{-1}\|\leqq C/(1+|\lambda|) \quad \text{for } \operatorname{Re}\lambda\leqq 0 \tag{2.7}$$

for some constants $C, \varepsilon>0$, where $\|\ \|$ denotes the operator norm on $\mathcal{B}(H)$. (In fact, (2.7) turns out to hold for $|\arg\lambda|>(\pi/2)-\varepsilon$ for any $\varepsilon\in(0, \pi/2)$, but we will not need that stronger result.) Let $\theta(A)\subseteq\mathbb{C}$ denote the numerical range of A; that is,

$$\theta(A)=\{\langle A(u, v), (u, v)\rangle_H: (u, v)\in\operatorname{dom} A, \|(u, v)\|_H=1\}.$$

Let Γ denote the closure of $\theta(A)$, and let $\Delta=\mathbb{C}\backslash\Gamma$. We shall use the following lemma, which is essentially a special case of Theorem V.3.2 in [7] applied to A.

LEMMA 2. *Suppose that A is closed and Δ is connected. For $\lambda\in\Delta$, $A-\lambda$ has nullity 0 and constant deficiency. If the deficiency is 0, then Δ is contained in the resolvent set of A and*

$$\|(\lambda-A)^{-1}\|\leqq 1/\operatorname{dist}(\lambda, \Gamma) \tag{2.8}$$

where $\operatorname{dist}(\lambda, \Gamma)$ *is the distance from λ to Γ in $\mathbb{C}$.*

To apply Lemma 2, we must calculate the numerical range of A. We have for $(u, v)\in X$ that

$$\begin{aligned}\langle A(u, v), (u, v)\rangle &= \int_{(0,\infty)\backslash\mathbb{Z}^+}(u''+u)\bar{u}+\sum_{j=1}^{\infty}(-[\![u']\!]_j+v^j)\bar{v}^j\\ &= \int_{(0,\infty)\backslash\mathbb{Z}^+}(|u'|^2+|u|^2)+\sum_{j=1}^{\infty}u'(j+)\bar{u}(j+)\\ &\quad -\sum_{j=1}^{\infty}u'(j-)\bar{u}(j-)-\sum_{j=1}^{\infty}[\![u']\!]_j\bar{v}^j+\sum_{j=1}^{\infty}|v^j|^2.\end{aligned}$$

Since $u(j+)=u(j-)=v^j$ for $(u, v)\in X$, the middle three terms on the right side of the second equation drop out, leaving

$$\begin{aligned}\langle A(u, v), (u, v)\rangle &= \int_{(0,\infty)\backslash\mathbb{Z}}(|u'|^2+|u|^2)+\sum_{j=1}^{\infty}|v^j|^2\\ &=\|(u, v)\|_Y^2\geqq\|(u, v)\|_H^2.\end{aligned}$$

Since $\|(u, v)\|_H=1$ in the definition of $\theta(A)$, we have $\theta(A)\subseteq[1, \infty)$ and $\Gamma\subseteq[1, \infty)$. Hence $\Delta=\mathbb{C}\backslash\Gamma$ is connected, and since $A^{-1}\in\mathcal{B}(H)$, A has deficiency zero on Δ. Thus Lemma 2 applies, and (2.7) follows immediately from (2.8) and some elementary geometry.

3. Existence. To solve (1.2) we recast it in terms of the operator A defined in § 2 and apply a result from the theory of abstract evolution equations. First, we rewrite

(1.2) as a pure initial value problem. Let $\sigma(x) \in C_0^\infty[[0,\infty)\backslash\mathbb{Z}^+]$ be such that $\sigma'(0+) = -1$ and $\operatorname{supp}\sigma \subseteq [0, 1/2)$. Let $u = U + \sigma(x)I(t)$ and $v = V$; then (u, v) satisfies

$$\begin{aligned} &u_t - u_{xx} + u = (1-g)u + (g\sigma - \sigma_{xx})I + \sigma I_t,\\ &v_t^j - [\![u_x]\!]_j + v^j = v^j + f(v^j),\\ &(u, v)|_{t=0} = (\phi, \psi) \equiv (\phi_0 + \sigma(x)I(0), \psi_0),\\ &u_x(0+) = 0. \end{aligned} \tag{3.1}$$

Letting $h(x, t) = (g\sigma - \sigma_{xx})I + \sigma I_t$, we may rewrite (3.1) as

$$\begin{aligned} &\frac{d}{dt}(u, v) + A(u, v) = F(t, (u, v)),\\ &(u, v)|_{t=0} = (\phi, \psi), \end{aligned} \tag{3.2}$$

where $F(t, (u, v)) = ((1-g)u + h, (v^j + f(v^j))_{j=1}^\infty)$. We will want $(\phi, \psi) \in \operatorname{dom} A = X$; also, we will want to differentiate (3.2) with respect to t, so we require

$$(\phi_0, \psi_0) \in X, \quad \phi_0'(0+) = I(0), \quad I(t) \in C^3[[0, \infty)]. \tag{3.3}$$

To solve (3.2) we use a form of a result due independently to Sobolevski and Tanabe (see [4, § II.16] and [5]); the present formulation is taken from [3, Thm. 1].

LEMMA 3. *Let A be a closed linear operator on a Banach space E such that* (2.7) *holds. Suppose that $F(t, p)$ is a function on $[0, T_0) \times E$ such that for some constants $\alpha, \eta \in (0, 1)$ and for any $R > 0$ there exists a constant $C(R)$ for which*

$$\|F(t_1, A^{-\alpha}p_1) - F(t_2, A^{-\alpha}p_2)\|_E \leqq C(R)[|t_1 - t_2|^\eta + \|p_1 - p_2\|_E] \tag{3.4}$$

for all $t_1, t_2 \in [0, T_0]$, $p_1, p_2 \in E$ with $\|p_1\|_E, \|p_2\|_E < R$. Then for any $p_0 \in \operatorname{dom} A$ and each $R > \|A^\sigma p_0\|_E$, there exists a $t^* = t^*(R, \|A^\alpha p_0\|_E) > 0$ such that the problem

$$\frac{dp}{dt} + Ap = F(t, p), \qquad p(0) = p_0 \tag{3.5}$$

has a unique solution in $[0, t^*]$. Furthermore, if there exists a constant $R' > 0$ such that for any solution p of (3.5) *in $[0, T_1]$, $T_1 \leqq T_0$, we have $\|Ap\|_E < R'$, then we may choose $R > R'$ and thus apply the local existence assertion of this lemma on $[0, t^*]$, $[t^*, 2t^*]$, and so on until $[0, T_0]$ is exhausted.*

Since the operator A in (3.2) was already shown to be closed and to satisfy (2.7), we need only establish (3.4) for the function F in (3.2) to conclude the local existence of solutions to (1.2). Let $(w_k, y_k) = A^{-\alpha}(u_k, v_k)$ for $k = 1, 2$. If $\|(u_k, v_k)\|_H < R$ for $k = 1, 2$, $|y_k^j| \leqq \|y_k\|_{l^2} \leqq \|(w_k, y_k)\|_H \leqq \|A^{-\alpha}\|R$, so

$$\begin{aligned} \|(f(y_1^j) - f(y_2^j))_{j=1}^\infty\|_{l^2} &\leqq \sup_{|y| < \|A^{-\alpha}\|R} |f'(y)|\,\|y_1 - y_2\|_{l^2}\\ &\leqq C_0(R)\|(w_1 - w_2, y_1 - y_2\|_H\\ &\leqq C_0(R)\|A^{-\alpha}\|\;\|(u_1, v_1) - (u_2, v_2)\|_H. \end{aligned} \tag{3.6}$$

We have

$$\begin{aligned} &\|F(t_1, A^{-\alpha}(u_1, v_1)) - F(t_2, A^{-\alpha}(u_2, v_2))\|_H\\ &\quad = \|F(t_1, (w_1, y_1)) - F(t_2, (w_2, y_2))\|_H\\ &\quad = \|((1-g)(w_1 - w_2) + h(t_1) - h(t_2), (y_1^j - y_2^j + f(y_1^j) - f(y_2^j))_{j=1}^\infty)\|_H\\ &\quad \leqq |1-g|\;\|(w_1 - w_2, 0)\|_H + \|(h(t_1) - h(t_2), 0\|_H\\ &\qquad + \|(0, (y_1^j - y_2^j + f(y_1^j) - f(y_2^j))_{j=1}^\infty)\|_H. \end{aligned} \tag{3.7}$$

Since $\|(w,0)\|_H \leqq \|(w,y)\|_H$, $\|(0,y)\|_H \leqq \|(w,y)\|_H$ and $\|(0,y)\|_H = \|(y^j)_{j=1}^\infty\|_{l^2}$, it follows from (3.6) and (3.7) that

$$\begin{aligned}\|F(t_1, A^{-\alpha}(u_1, v_1)) - F(t_2, A^{-\alpha}(u_2, v_2))\|_H &\leqq \|(h(t_1)-h(t_2), 0)\|_H \\ &\quad + (|1+g| + C_0(R))\|A^{-\alpha}\| \, \|(u_1, y_1) - (u_2, y_2)\|_H.\end{aligned} \tag{3.8}$$

Since $\sigma \in C_0^\infty([0,\infty)\mathbb{Z}^+)$ with supp $\sigma \in [0, 1/2]$ and $I(t) \in C^3([0,\infty))$, (3.4) follows from (3.8) with $\eta \in (0,1)$ arbitrary. Thus, the local existence of solutions to (1.2) is established.

To obtain global existence we must bound $\|A(u,v)\|_H$ for any solution to (3.2). Since only the situation with (ϕ, ψ) real valued is of physical interest, we shall assume that to be the case, so that (u, v) is real valued also. To bound $\|A(u,v)\|_H$ it is sufficient to bound $\|(u,v)\|_H$, $\|(h,0)\|_H$ and $\|(u_t, v_t)\|_H$, since if $\|(u,v)\|_H \leqq M$ then $|v^j| \leqq M$ so by (1.2) $|f(v^j)| \leqq \sup_{|y|<M} |f'(y)| \, |v^j|$ and thus $\|(0, (f(v^j))_{j=1}^\infty\|_H \leqq \sup_{|y|<M} |f'(y)| \, \|(0,v)\|_H$. Because we assume the initial data $(\phi, \psi) \in \operatorname{dom} A$, the proof of Lemma 3 and the smothness of F imply the existence of continuous first and second time derivatives in H for any solution (u, v) of (3.2); see the discussion following Theorem 2 in [5]. We obtain the necessary bounds via a Lyapunov or energy functional; the method is similar to those used in [1], [3], [8]. Let

$$\begin{aligned}E(t) &= \frac{1}{2}(\|(u,v)\|_H^2 + \|(u,v)_t\|_H^2) \\ &= \frac{1}{2}\int_{(0,\infty)\setminus\mathbb{Z}^+} [u^2 + u_t^2] + \frac{1}{2}\sum_{j=1}^\infty [(v_t^j)^2 + (v^j)^2].\end{aligned}$$

We have

$$\begin{aligned}E'(t) &= \int_{(0,\infty)\setminus\mathbb{Z}^+} (uu_t + u_t u_{tt}) + \sum_{j=1}^\infty v_t^j v_{tt}^j + v^j v_t^j \\ &= \int_{(0,\infty)\setminus\mathbb{Z}^+} [u(u_{xx} - gu + h) + u_t(u_{xxt} - gu_t + h_t)] \\ &\quad + \sum_{j=1}^\infty [v_t^j [\![u_{xt}]\!]_j + f'(v^j)(v_t^j)^2 + v^j [\![u_x]\!]_j + f(v^j)v^j] \\ &= -\int_{(0,\infty)\setminus\mathbb{Z}} [u_x^2 + gu^2 + u_{xt}^2 + gu_t^2] \\ &\quad + \int_{(0,\infty)\setminus\mathbb{Z}} [hu + h_t u_t] + \sum_{j=1}^\infty f'(v^j)(v_t^j)^2 + f(v^j)v^j\end{aligned} \tag{3.9}$$

where we have used the fact that (u, v), $(u, v)_t \in X$, so that $u(j+, t) = u(j-, t) = v^j(t)$ and $u_t(j+, t) = u_t(j-, t) = v_t^j(t)$ as in the computation of the numerical range of A following Lemma 2. Using Cauchy's inequality and (1.2) we have from (3.9)

$$\begin{aligned}E'(t) &\leqq -\int_{(0,\infty)\setminus\mathbb{Z}^+} (u_x^2 + u_{xt}^2) + [(1/2) - g]\int_{(0,\infty)\setminus\mathbb{Z}^+} (u^2 + u_t^2) \\ &\quad + (1/2)\int_{(0,\infty)\setminus\mathbb{Z}^+} (h^2 + h_t^2) + f_0 \sum_{j=1}^\infty (v^j)^2 + (v_t^j)^2].\end{aligned} \tag{3.10}$$

From (3.10) and the properties of h, it follows immediately that $E'(t) \leqq K_1 + K_2 E(t)$ for some constants K_1, K_2 that do not depend on (u, v), as long as (u, v) exists. We already have the existence of a local solution to (3.2); so choose $\varepsilon > 0$ such that a solution to (3.2) exists on $[0, T_2)$ for some $T_2 > \varepsilon$. We can now apply Lemma 3 again with initial data $(u(\varepsilon), v(\varepsilon)) \in X$. By our differential inequality for E, we have $E(t) \leqq K_3(t) + E(\varepsilon)\, e^{K_2 t}$ with $K_3(t)$ bounded on any fixed interval in t. By our assumptions on σ and I, $\|(h, 0)\|_H$ is bounded on any fixed interval $[0, T_1]$. Since $2E(t)$ bounds $\|(u, v)\|_H$ and $|(u, v)_t\|_H$, we have as noted above that $\|A(u, v)\|_H$ is bounded on any subinterval of $[\varepsilon, T_0]$ for which the solution exists; the bound depends only on K_2, K_3, and $E(\varepsilon)$. Hence we have the necessary bound for global existence of solutions on $[\varepsilon, T_0]$ and hence on $[0, T_0]$ by Lemma 3. Since $T_0 > 0$ was arbitrary, our solution must exist for all $t > 0$. We have thus proved the following.

THEOREM. *If* $(\phi_0, \psi_0) \in X$ *and conditions* (1.2), (3.3) *hold, then the problem* (1.2) *has a unique solution* (U, V) *for all* $t > 0$.

Acknowledgment. The author wishes to thank Jonathan Bell for much useful discussion and correspondence.

REFERENCES

[1] R. ARIMA AND Y. HASEGAWA, *On global solutions for mixed problem of a semilinear differential equation*, Proc. Japan Acad. Ser. A Math. Sci., 39 (1963), pp. 721-725.

[2] J. BELL AND C. COSNER, *Threshold conditions for a diffusive model of a myelinated axon*, J. Math. Biol., 18 (1983), pp. 39-52.

[3] ———, *Stability properties of a model of parallel nerve fibers*, J. Differential Equations, 40 (1981), pp. 303-315.

[4] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[5] ———, *Remarks on nonlinear parabolic equations*, Proc. Sympos. Appl. Math., 17 (1965), pp. 3-23.

[6] P. GRINDROD AND B. SLEEMAN, *A model of a myelinated nerve axon: threshold behaviour and propagation*, University of Dundee, Applied Analysis Technical Report AA 851 (1985).

[7] T. KATO, *Perturbation Theory for Linear Operators*, Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen v. 132, Springer-Verlag, Berlin, New York, 1966.

[8] J. RAUCH, *Global existence for the FitzHugh-Nagumo equations*, Comm. Partial Differential Equations, 1 (1976), pp. 609-621.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

010

THE BLOW-UP TIME FOR SOLUTIONS OF NONLINEAR HEAT EQUATIONS WITH SMALL DIFFUSION*

AVNER FRIEDMAN† AND ANDREW A. LACEY‡

Abstract. Consider a nonlinear heat equation $u_t - \varepsilon\Delta u = f(u)$ in a cylinder $\{x \in \Omega,\ t > 0\}$, with u vanishing on the lateral boundary and $u = \phi_\varepsilon(x)$ initially ($\phi_\varepsilon \geqq 0$). Denote by T_ε the blow-up time for the solution. Asymptotic estimates are obtained for T_ε as $\varepsilon \to 0$.

Key words. nonlinear heat equation, blow-up of solutions

AMS(MOS) subject classifications. Primary 35K55; secondary 35B25

Introduction. Consider the system

$$u_t - \varepsilon\Delta u = f(u) \qquad (x \in \Omega, t > 0), \tag{0.1}$$

$$u(x,t) = 0 \qquad (x \in \partial\Omega, t > 0), \tag{0.2}$$

$$u(x,0) = \phi(x) \qquad (x \in \Omega) \tag{0.3}$$

where $\phi(x)$ is continuous and nonnegative, $f(s)$ is positive and increasing for $s \geqq 0$, and $1/f$ is integrable. Denote by T_ε the time when the solution blows up. Denote by T_0 the time when the solution of

$$v'(t) = f(v(t)) \qquad (t > 0), \tag{0.4}$$

$$v(0) = \phi(x_0) \tag{0.5}$$

blows up, where $\phi(x_0) = \max_{x \in \bar{\Omega}} \phi(x)$. We are interested in estimating T_ε as $\varepsilon \to 0$. If $\Delta\phi(x_0) < 0$ then we prove that

$$c\varepsilon < T_\varepsilon - T_0 < C\varepsilon \tag{0.6}$$

where c, C are positive constants. More generally, if

$$\phi(x) \sim \phi(x_0) - c_0|x - x_0|^{2\alpha} \quad \text{as } x \to x_0$$

for some positive numbers c_0 and α, then

$$c\varepsilon^\alpha < T_\varepsilon - T_0 < C\varepsilon^\alpha. \tag{0.7}$$

The forms of the constants c, C in (0.6), (0.7) will be given quite explicitly.

We also consider the case where $\phi(x)$ varies with ε. Suppose

$$\phi_\varepsilon(x) = \begin{cases} \phi_0(x) + \Phi\left(\dfrac{x}{\varepsilon^\beta}\right) & \text{if } \dfrac{x}{\varepsilon^\beta} \in B_1, \\ \phi_0(x) & \text{if } \dfrac{x}{\varepsilon^\beta} \notin B_1 \end{cases}$$

where $B_1 = \{y;\ |y| < 1\}$, $\Phi(y) \geqq 0$, $\Phi(y) = 0$ if $y \in \partial B_1$, $\Phi \in C^0(\bar{B}_1)$, and let

$$\phi_0(0) = \max_{\Omega} \phi, \qquad \Phi(0) = \max_{B_1} \Phi.$$

* Received by the editors January 2, 1986; accepted for publication April 8, 1986. This work was partially supported by National Science Foundation grants DMS-8420896 and DMS-8501397.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 49707.

‡ Heriot-Watt University, Edinburgh, Scotland.

Denote by T^* and T_* the blow-up times for the solutions of (0.4), (0.5) corresponding to $\phi_0(0)$ and $\phi_0(0)+\Phi(0)$. Then if $\beta \neq \frac{1}{2}$, $T_\varepsilon \to T_0$ as $\varepsilon \to 0$, where

$$T_0 = T_* \quad \text{if } \beta < \tfrac{1}{2}, \qquad T_0 = T^* \quad \text{if } \beta > \tfrac{1}{2}; \tag{0.8a}$$

also

$$T_* < \varliminf_{\varepsilon \to 0} T_\varepsilon \leqq \varlimsup_{\varepsilon \to 0} T_\varepsilon < T^* \quad \text{if } \beta = \tfrac{1}{2}. \tag{0.8b}$$

Notice that if $\phi_0(0) < \max_\Omega \phi_0$ then the asymptotic behavior of T_ε does not depend on Φ.

The assertion (0.6) is proved in § 1, (0.7) is proved in § 2 and (0.8) is proved in § 3.

We finally mention that recent literature on blow-up of solutions of nonlinear heat equations can be found in [1]-[5] and in the references given there.

1. The estimate (0.6). Let Ω be a bounded domain in $\mathbb{R}^n$ whose boundary $\partial\Omega$ is locally a Lipschitz graph. For any $s > 0$ set

$$\Omega_s = \Omega \times \{0 < t < s\}, \qquad \partial\Omega_s = \partial\Omega \times \{0 < t < s\}$$

and consider the parabolic problem

$$L_\varepsilon u_\varepsilon \equiv \frac{\partial u_\varepsilon}{\partial t} - \varepsilon \Delta u_\varepsilon - f(u_\varepsilon) = 0 \quad \text{in } \Omega_\infty, \tag{1.1}$$

$$u_\varepsilon = 0 \quad \text{on } \partial\Omega_\infty, \tag{1.2}$$

$$u_\varepsilon(x, 0) = \phi(x) \quad \text{for } x \in \Omega \tag{1.3}$$

where

$$\begin{gathered} \phi \in C^2(\Omega) \cap C^0(\bar{\Omega}), \qquad \phi(x) \geqq 0, \\ \phi = 0 \quad \text{on } \partial\Omega \end{gathered} \tag{1.4}$$

and $f(s)$ satisfies

$$\begin{gathered} f(s) > 0, \quad f'(s) > 0, \quad f''(s) \geqq 0 \quad \text{if } s \geqq 0, \\ \int^\infty \frac{ds}{f(s)} < \infty; \end{gathered} \tag{1.5}$$

at the end of § 3 we shall extend all the results to the case where $f(0) = 0$.

It is easily seen that there exists a unique function $\zeta(s)$ and a positive number s_0 such that

$$\begin{gathered} \frac{d\zeta}{ds} = -f(\zeta) \quad \text{if } 0 < s < s_0, \\ \zeta(s) \to \infty \quad \text{if } s \to 0, \qquad \zeta(s_0 - 0) = 0. \end{gathered} \tag{1.6}$$

Clearly

$$\zeta' < 0, \qquad \zeta'' < 0. \tag{1.7}$$

The inverse function ζ^{-1} satisfies

$$(\zeta^{-1})'(s) = -\frac{1}{f(s)}. \tag{1.8}$$

The most interesting examples for f are

$$f(s)=e^s, \text{ and then } \zeta(s)=-\log s,\ \zeta^{-1}(s)=e^{-s}, \tag{1.9}$$

$$\begin{gathered} f(s)=(s+\lambda)^p(\lambda>0, p>1), \text{ and then} \\ \zeta(s)=\alpha s^{-1/(p-1)}-\lambda,\ \zeta^{-1}(s)=\frac{1}{p-1}(s+\lambda)^{1-p} \qquad (\alpha=(p-1)^{-1/(p-1)}); \end{gathered} \tag{1.10}$$

the case $\lambda=0$ is considered at the end of § 3.

For any $\bar{x}\in\Omega$ consider

$$\frac{dv}{dt}=f(v) \quad \text{for } t>0, \qquad v(0)=\phi(\bar{x})$$

and denote the solution by $u_{\bar{x}}(t)$. The solution must blow up in finite time $T_{\bar{x}}$ and

$$u_{\bar{x}}(t)=\zeta(T_{\bar{x}}-t). \tag{1.11}$$

Taking $t=0$ we get $\phi(\bar{x})=\zeta(T_{\bar{x}})$, or

$$T_{\bar{x}}=\zeta^{-1}(\phi(\bar{x})). \tag{1.12}$$

Notation. We denote $T_{\bar{x}}$ by $T(\phi(\bar{x}))$, and we denote the blow-up time for the solution u_ε of (1.1)-(1.3) by T_ε.

THEOREM 1.1. *Let $\phi(x_0)=\max_{x\in\Omega}\phi(x)$, $x_0\in\Omega$. Then*

$$T_\varepsilon \geqq T(\phi(x_0)). \tag{1.13}$$

Proof. The function $v(t)=u_{x_0}(t)$ satisfies

$$v_t-\Delta v=f(v) \quad \text{in } \Omega_{T_{x_0}},$$

and $v\geqq u_\varepsilon$ on the parabolic boundary. Hence, by comparison, $v>u_\varepsilon$ in $\Omega_{T_{x_0}}$. Since $u_\varepsilon\geqq 0$ it follows that u_ε cannot blow up in time smaller than T_{x_0}.

In the sequel we shall obtain a much more refined estimate than (1.13).

THEOREM 1.2. *Let x_0 be any point in Ω such that $\phi(x_0)=\max_{x\in\Omega}\phi(x)$. Then*

$$T_\varepsilon \leqq T(\phi(x_0))+\varepsilon T(\phi(x_0))\frac{|\Delta\phi(x_0)|}{f(\phi(x_0))}+o(\varepsilon) \tag{1.14}$$

as $\varepsilon\to 0$.

Proof. We shall construct a subsolution of the form

$$w(x,t)=\zeta(V(x,t)-t).$$

Using (1.6) and the fact that $\zeta''>0$, we have

$$\begin{aligned} w_t-\varepsilon\Delta w-f(w) &= \zeta'(V-t)(V_t-1-\varepsilon\Delta V)-\varepsilon\zeta''(V-t)|\nabla V|^2-f(\zeta(V-t)) \\ &\leqq -f(\zeta(V-t))(V_t-\varepsilon\Delta V). \end{aligned}$$

Taking

$$V(x,t)=U(x,\varepsilon t)$$

where $U(x,\tau)$ satisfies

$$U_\tau=\Delta U,$$

we conclude that w is a subsolution.

We choose the initial and boundary conditions for U such that

$$(1.15)\qquad \begin{aligned} &U(x,0)=\zeta^{-1}(\phi(x)) \quad \text{if } x\in\Omega,\\ &U(x,\tau)=\frac{\tau}{\varepsilon}+s_0 \quad \text{if } x\in\partial\Omega, \qquad 0<\tau<C\varepsilon \end{aligned}$$

where C is a positive constant. Then

$$\zeta(U(x,0))=\phi(x)=u_\varepsilon(x,0) \quad \text{if } x\in\Omega,$$

$$\zeta\left(U(x,\tau)-\frac{\tau}{\varepsilon}\right)=0 \quad \text{if } x\in\partial\Omega, \quad 0<\tau<C\varepsilon,$$

so that $w=u_\varepsilon$ on the parabolic boundary of Ω_C. The maximum principle then gives

$$(1.16)\qquad u_\varepsilon(x,t)\geqq\zeta(U(x,\varepsilon t)-t) \quad \text{if } t<C.$$

Notice that the boundary values of U are bounded, i.e.,

$$0\leqq U(x,\tau)\leqq C+s_0 \quad \text{if } x\in\partial\Omega, \quad \tau\leqq C\varepsilon.$$

Hence we can write

$$(1.17)\qquad U(x_0,\varepsilon t)=U(x_0,0)+\varepsilon tU_\tau(x_0,0)+o(\varepsilon)$$

where $o(\varepsilon)/\varepsilon\to 0$ as $\varepsilon\to 0$, uniformly in t if $0<t<C$.

The function w blows up at time $\leqq\tilde{t}$ where $\tilde{t}$ satisfies

$$U(x_0,\varepsilon\tilde{t})-\tilde{t}=0.$$

Choosing $C>U(x_0,0)+1$ we see, upon recalling (1.17), that $\tilde{t}$ is determined by

$$\tilde{t}(1-\varepsilon U_\tau(x_0,0))=U(x_0,0)+o(\varepsilon)$$

or, since $U_\tau=\Delta U$,

$$\tilde{t}=U(x_0,0)+\varepsilon U(x_0,0)\Delta U(x_0,0)+o(\varepsilon).$$

In view of (1.16) we also have $T_\varepsilon\leqq\tilde{t}$, and recalling (1.15), (1.12) and the fact that

$$\Delta\zeta^{-1}(\phi(x))=-\frac{\Delta\phi(x)}{f(\phi(x))} \quad \text{at } x=x_0$$

(by (1.18) and $\nabla\phi(x_0)=0$), the assertion (1.14) follows.

THEOREM 1.3. *Let ϕ achieve its maximum at the unique point $x_0\in\Omega$ and suppose that $\Delta\phi(x_0)<0$. Then*

$$(1.18)\qquad T_\varepsilon\geqq T(\phi(x))+\varepsilon\gamma(\phi(x_0))|\Delta\phi(x_0)|+o(\varepsilon)$$

as $\varepsilon\to 0$, where $\gamma(\phi(x_0))$ is defined by

$$(1.19)\qquad \gamma(\phi(x_0))=\max_{0<\lambda<\zeta^{-1}(\phi(x_0))}\lambda/f(\zeta(\zeta^{-1}(\phi(x_0))-\lambda)).$$

Remark 1.1. In case (1.9), the combined estimates (1.14) and (1.18) become

$$(1.20)\qquad \tfrac{1}{4}\varepsilon\, e^{-2\phi(x_0)}|\Delta\phi(x_0)|+o(\varepsilon)\leqq T_\varepsilon\leqq\varepsilon\, e^{-2\phi(x_0)}|\Delta\phi(x_0)|+o(\varepsilon).$$

Proof. Taking $x_0=0$ we first construct a supersolution of the form

$$w=\zeta(\zeta^{-1}(\phi(0))-t)+W(x,t), \qquad 0<t<C$$

where $C \in (0, T(\phi(0)))$ is to be chosen,

$$W(x,t) = Z(x, \varepsilon t)$$

and $Z(x, \tau)$ satisfies

$$\begin{aligned} &Z_\tau = \Delta Z \quad \text{if } x \in \Omega, \quad 0 < \tau < C\varepsilon, \\ &Z(x,0) = \phi(x) - \phi(0) \quad \text{if } x \in \Omega, \\ &Z(x,\tau) = \phi(x) - \phi(0) \quad \text{if } x \in \partial\Omega, \quad 0 < \tau < C\varepsilon. \end{aligned}$$

Clearly $Z \leqq 0$, whence

$$\begin{aligned} L_\varepsilon w &= w_t - \varepsilon \Delta w - f(w) \\ &= f(\zeta(\zeta^{-1}(\phi(0)) - t)) + W_t - \varepsilon \Delta W - f(w) \\ &= f(\zeta(\zeta^{-1}(\phi(0)) - t)) - f(\zeta(\zeta^{-1}(\phi(0)) - t) + W) \geqq 0. \end{aligned}$$

Also

$$w(x,0) = \phi(0) + (\phi(x) - \phi(0)) = u_\varepsilon(x) \quad \text{if } x \in \Omega$$

and

$$w(x,t) = \zeta(\zeta^{-1}(\phi(0)) - t) + \phi(x) = \phi(0) \geqq 0 \quad \text{if } x \in \partial\Omega,$$

since $\zeta(\zeta^{-1}(\phi(0)) - t) \geqq \phi(0)$ and $\phi(x) = 0$ for $x \in \partial\Omega$.

It follows that

$$u_\varepsilon \leqq w \quad \text{if } x \in \Omega, \quad 0 < t < C. \tag{1.21}$$

We now choose μ arbitrarily small and η small enough depending on μ so that

$$-\Delta\phi(x) > -\Delta\phi(0) - \mu \quad \text{if } |x| < \eta.$$

Then

$$Z \leqq (\Delta\phi(0) + \mu)\tau + o(\tau) \quad \text{as } \tau \to 0 \quad \text{for } |x| < \eta,$$

while $Z \leqq 0$ if $|x| \geqq \eta$. It follows that, for $\varepsilon \to 0$,

$$\begin{aligned} u_\varepsilon(x,c) \leqq w(x,C) &= \zeta(\zeta^{-1}(\phi(0)) - C) + W(x,C) \\ &\leqq \zeta(\zeta^{-1}(\phi(0)) - C) + (\Delta\phi(0) + \mu)C + o(\varepsilon). \end{aligned}$$

Denote by $\tilde{u}(x,t)$ the solution of (1.1), (1.2) for $t > C$ with $\tilde{u}(x, C) = w(x, C)$ and denote by $\tilde{T}$ the blow-up time for $\tilde{u}$. From a comparison of the form (1.13),

$$\begin{aligned} T_\varepsilon &\geqq \tilde{T} \geqq C + \zeta^{-1}\{\max_{\Omega} u_\varepsilon(\cdot, C)\} \\ &\geqq C + \zeta^{-1}\{\zeta(\zeta^{-1}(\phi(0)) - C + (\Delta\phi(0) + \mu)\varepsilon C + o(\varepsilon)\} \\ &= C + (\zeta^{-1}(\phi(0)) - C) - \varepsilon C(\Delta\phi(0) + \mu)/f(\zeta(\zeta^{-1}(\phi(0)) - C)) + o(\varepsilon) \quad \text{(by (1.18))} \\ &= T(\phi(0)) - \varepsilon C(\Delta\phi(0) + \mu)/f(\zeta(\zeta^{-1}(\phi(0)) - C)) + o(\varepsilon). \end{aligned}$$

But as this holds for μ arbitrarily small and $\Delta\phi(0) < 0$,

$$T_\varepsilon \geqq T(\phi(0)) + \varepsilon|\Delta\phi(0)|C/f(\zeta(\zeta^{-1}(\phi(0)) - C)) + o(\varepsilon).$$

Choosing $C = C_M \in (0, \zeta^{-1}(\phi(0)))$, which maximizes the coefficient of ε, assertion (1.18) follows with $\gamma(\phi(x_0))$ defined by (1.19).

Remark 1.2. Theorem 1.1 does not require the assumption that $\phi \in C^2(\Omega)$; Theorems 1.2 and 1.3 require only that $\phi \in C^2$ in a neighborhood of x_0.

Remark 1.3. If $\phi(x)$ achieves its maximum at several points x_i, then in assertion (1.18) one should replace $|\Delta\phi(x_0)|$ by $\min_i |\Delta\phi(x_i)|$; the proof is the same as before.

2. The cases $\Delta\phi_0(x_0)=0$ or $-\infty$. Consider the case where at a point x_0 of maximum of ϕ

$$\phi(x) \geqq \phi(x_0) - c_0|x-x_0|^{2\alpha} \quad \text{if } x \to x_0 \tag{2.1}$$

where c_0 is a positive constant and $\alpha > 0$, $\alpha \neq 1$; the assumption that $\phi \in C^2(\Omega)$ made in § 1 is dropped (cf. Remark 1.2). The inequality (2.1) is taken in the usual sense that

$$\liminf_{x \to x_0} \frac{\phi(x)-\phi(x_0)}{|x-x_0|^{2\alpha}} \geqq -c_0.$$

THEOREM 2.1. *If* (2.1) *holds then*

$$T_\varepsilon \leqq T(\phi(x_0)) + \kappa_\alpha c_0 \varepsilon^\alpha \{\zeta^{-1}(\phi(x_0)\}^\alpha / f(\phi(x_0)) + o(\varepsilon^\alpha) \tag{2.2}$$

as $\varepsilon \to 0$, *where* κ_α *is a universal positive constant, defined by* (2.3) *below.*

Proof. We proceed as in the proof of Theorem 1.2. The only difference occurs when we analyze the behavior of the function $U(x_0, \tau)$. Using the relation

$$\int_\Omega \frac{e^{-(|x-\xi|^2/4\tau)}}{(4\pi\tau)^{n/2}} |\xi - x_0|^{2\alpha}\, d\xi = \kappa_\alpha \tau^\alpha + o(\tau^\alpha) \quad \text{as } \tau \to 0 \tag{2.3}$$

in the representation of $U(x_0, \tau)$ by means of Green's function, we get

$$U(x_0, \tau) \leqq \zeta^{-1}(\phi(x_0)) - \kappa_\alpha c_0 \tau^\alpha + o(\tau^\alpha) \quad \text{as } \tau \to 0.$$

The blow-up time for the subsolution w is $\leqq \tilde{t}$ where

$$\tilde{t} = U(x_0, \varepsilon\tilde{t}) \leqq T(\phi(x_0)) + \kappa_\alpha c_0 \zeta^{-1}(\phi(x_0)) \varepsilon^\alpha \tilde{t}^\alpha + o(\tilde{t}^\alpha),$$

from which the assertion (2.2) follows.

We next obtain a lower bound on T_ε, assuming that

$$\phi(x) \leqq \phi(x_0) - c_0|x-x_0|^{2\alpha} \quad \text{as } x \to x_0. \tag{2.4}$$

As before, the assumption $\phi \in C^2(\Omega)$ is dropped.

THEOREM 2.2. *Assume that* ϕ *achieves its maximum at the unique point* x_0 *and that* (2.4) *holds for some* $c_0 > 0$ *and* $\alpha > 0$. *Then*

$$T_\varepsilon \geqq T(\phi(x_0) + \kappa_\alpha c_0 \varepsilon^\alpha \gamma_\alpha(\phi(x_0)) + o(\varepsilon^\alpha) \tag{2.5}$$

as $\varepsilon \to 0$, *where* $\gamma_\alpha(\phi(x_0))$ *is defined by*

$$\gamma_\alpha(\phi(x_0)) = \max_{0<\lambda<\zeta^{-1}(\phi(x_0))} \lambda^\alpha / f(\zeta(\zeta^{-1}(\phi(x_0)) - \lambda)). \tag{2.6}$$

Proof. We proceed as in the proof of Theorem 1.3, modified as in the proof of Theorem 2.1, to obtain the local behavior of Z:

$$Z(x, \tau) = -\kappa_\alpha c_0 \tau^\alpha + o(\tau^\alpha) \quad \text{as } \tau \to 0.$$

The blow-up time for the supersolution $\tilde{u}$, applicable for $T > C$, is

$$\tilde{T} = T(\phi(x_0)) + \kappa_\alpha c_0 \varepsilon^\alpha / f(\zeta(\zeta^{-1}(\phi(x_0)) - C)) + o(\varepsilon^\alpha);$$

the assertions (2.5) and (2.6) now readily follow.

Remark 2.1. If

$$\phi(x) \sim \phi(x_0) - c_0|x-x_0|^\alpha \quad \text{as } x \to x_0$$

then from Theorems 2.1 and 2.2 we obtain

$$T(\phi(x_0)) + \kappa_\alpha c_0 \varepsilon^\alpha \gamma_\alpha(\phi(x_0)) + o(\varepsilon^\alpha) \leqq T_\varepsilon$$
$$\leqq T(\phi(x_0)) + \kappa_\alpha c_0 \varepsilon^\alpha \{\zeta^{-1}(\phi(x_0))\}^\alpha / f(\phi(x_0)) + o(\varepsilon^\alpha).$$

In particular, for the special case $f(s) = e^s$,

$$e^{-\phi(x_0)} + \frac{\alpha^\alpha}{(1+\alpha)^{1+\alpha}} \kappa_\alpha c_0 \varepsilon^\alpha \, e^{-(1+\alpha)\phi(x_0)} + o(\varepsilon^\alpha) \leqq T_\varepsilon$$
$$\leqq e^{-\phi(x_0)} + \kappa_\alpha c_0 \varepsilon^\alpha \, e^{-(1+\alpha)\phi(x_0)} + o(\varepsilon^\alpha).$$

3. Proof of (0.8). In this section we consider the case where ϕ depends on the parameter ε, say $\phi = \phi_\varepsilon$. The function ϕ_ε is a sum of two functions:

$$\phi_\varepsilon(x) = \phi_0(x) + \Phi_\varepsilon(x) \tag{3.1}$$

where

$$\Phi_\varepsilon(x) = \begin{cases} \Phi\left(\dfrac{x}{\varepsilon^\beta}\right) & \text{if } |x| < \varepsilon^\beta, \\ 0 & \text{if } |x| > \varepsilon^\beta, \qquad \beta > 0, \end{cases} \tag{3.2}$$

and

$$\phi_0(0) = \max_{x \in \Omega} \phi(x), \qquad 0 \in \Omega. \tag{3.3}$$

We assume that

$$\begin{gathered} \Phi \in C^0(\bar{B}_1), \quad \Phi \geqq 0, \quad \Phi(0) = \max_{x \in \bar{B}_1} \Phi(x), \\ \Phi = 0 \quad \text{on } \partial B_1, \\ \Phi \in C^2(B_\rho) \quad \text{for some } \rho > 0 \end{gathered} \tag{3.4}$$

where $B_\rho = \{x; |x| < \rho\}$, and set

$$\phi_0 = \phi(0), \qquad \Phi_0 = \Phi(0).$$

To avoid a trivial case we assume that $\Phi_0 > 0$.

Introduce the blow-up times

$$T^* = T(\phi_0), \qquad T_* = T(\phi_0 + \Phi_0); \tag{3.5}$$

clearly $T_* < T^*$.

THEOREM 3.1. (i) *If* $\beta < \frac{1}{2}$ *then* $T_\varepsilon \to T_*$ *as* $\varepsilon \to 0$; (ii) *if* $\beta > \frac{1}{2}$ *then* $T_\varepsilon \to T^*$ *as* $\varepsilon \to 0$; (iii) *if* $\beta = \frac{1}{2}$ *then* $T_0^- \equiv \lim_{\varepsilon \to 0} T_\varepsilon$ *and* $T_0^+ = \overline{\lim}_{\varepsilon \to 0} T_\varepsilon$ *satisfy* $T_* < T_0^- \leqq T_0^+ < T^*$.

Proof. For simplicity consider first the case where $\phi_0 \equiv 0$. Let $\tilde{\Phi}(x)$ be a function satisfying: $\tilde{\Phi}(0) = \Phi(0)$, $\tilde{\Phi} \geqq \Phi_\varepsilon$, $\tilde{\Phi} = 0$ on $\partial\Omega$. By comparison

$$u_{0,\varepsilon} \leqq u_\varepsilon \leqq u_{1,\varepsilon}$$

where $u_{0,\varepsilon}$, $u_{1,\varepsilon}$ are the solutions of (1.1)-(1.3) corresponding to $\phi \equiv 0$ and $\phi \equiv \tilde{\Phi}$ respectively. Since, by Theorems 1.1 and 1.2, the blow-up times $T_{i,\varepsilon}$ corresponding to u_i satisfy

$$T_{0,\varepsilon} \to T^*, \qquad T_{1,\varepsilon} \to T_* \quad \text{as } \varepsilon \to 0,$$

it follows that

$$T_* \leqq \varliminf_{\varepsilon \to 0} T_\varepsilon \leqq \varlimsup_{\varepsilon \to 0} T_\varepsilon \leqq T^*. \tag{3.6}$$

In order to prove (i) it suffices to show that

$$\varlimsup_{\varepsilon \to 0} T_\varepsilon \leqq T_*. \tag{3.7}$$

To do this we construct a subsolution w of the form

$$w(x, t) = \zeta(V(x, t) - t)$$

where

$$V(x, t) = U\left(\frac{x}{\varepsilon^\beta}, \frac{\varepsilon t}{\varepsilon^{2\beta}}\right)$$

and $U(u, \tau)$ satisfies

$$\begin{aligned} &U_\tau = \Delta U \quad \text{if } y \in \mathbb{R}^n, \quad \tau > 0,\\ &\zeta(U(y, 0)) = \Phi(y) \quad \text{if } |y| < 1,\\ &\zeta(U(y, 0)) = 0 \quad \text{if } |y| > 1,\\ &\zeta(U(y, \tau)) = 0 \quad \text{if } |y| = 1. \end{aligned}$$

As in the proof of Theorem 1.2, w is a subsolution and $u_\varepsilon \geqq w$. Hence $T_\varepsilon \leqq \tilde{T}_\varepsilon$ where $\tilde{T}_\varepsilon$ is the blow-up time for w. Further, $\tilde{T}_\varepsilon \leqq \tilde{t}$ where $\tilde{t}$ satisfies

$$\tilde{t} \sim V\left(0, \frac{\varepsilon \tilde{t}}{\varepsilon^{2\beta}}\right) = \zeta^{-1}(\Phi_0) + O(\varepsilon^{1-2\beta})$$

since V_t is continuous about (0, 0). The estimate (3.7) now immediately follows.

To prove (ii) it suffices, in view of (3.6), to show that

$$\varliminf_{\varepsilon \to 0} T_\varepsilon \geqq T^*. \tag{3.8}$$

We proceed similarly to the proof of Theorem 1.3, to construct a supersolution w of the form

$$w = W(x, t) + ta$$

where a is a positive constant,

$$W(x, t) = U\left(\frac{x}{\varepsilon^\beta}, \frac{\varepsilon t}{\varepsilon^{2\beta}}\right)$$

and $U(y, \tau)$ satisfies

$$\begin{aligned} &U_\tau = \Delta U \quad \text{if } y \in \mathbb{R}^n, \quad \tau > 0,\\ &U(y, 0) = \Phi(y) \quad \text{if } |y| < 1,\\ &U(y, 0) = 0 \quad \text{if } |y| > 1. \end{aligned}$$

By the maximum principle $0 \leqq W \leqq \Phi_0$. Hence

$$w_t - \varepsilon \Delta w - f(w) = a - f(W + ta) > 0$$

if $a > 1 + f(\Phi_0)$, $t \leqq b$, provided b is some small positive number. It follows that

$$u_\varepsilon \leqq w \quad \text{in } \Omega_b.$$

Since $U(y, \tau) \leqq C\tau^{-n/2}$, we have

$$w \leqq at + C(t\varepsilon^{1-2\beta})^{-n/2}.$$

Noting that

$$t_0 \equiv \varepsilon^{(2\beta-1)n/(n+2)} < b$$

if ε is small enough, we get

$$u_\varepsilon(x, t_0) \leqq w(x, t_0) < C\varepsilon^{(2\beta-1)n/(n+2)}. \tag{3.9}$$

Using comparison for $t > t_0$, we conclude from (3.9) (by (1.8), (1.9)) that

$$u(x, t) \leqq \zeta(\zeta^{-1}(C\varepsilon^{(2\beta-1)n/(n+2)}) - t + t_0).$$

The right-hand side blows up at time $\sim \tilde{t}$ where

$$\begin{aligned} \tilde{t} &= t_0 + \zeta^{-1}(C\varepsilon^{(2\beta-1)n/(n+2)}) \\ &= t_0 + T^* + (\zeta^{-1})'(0)C\varepsilon^{(2\beta-1)n/(n+2)}. \end{aligned}$$

Consequently

$$T_\varepsilon > T^* + O(\varepsilon^{(2\beta-1)n/(n+2)}),$$

from which (3.8) follows.

To prove (iii) let $v_\varepsilon(x, t) = u(x\sqrt{\varepsilon}, t)$. Then

$$\begin{aligned} &\frac{\partial v_\varepsilon}{\partial t} - \Delta v_\varepsilon = f(v_\varepsilon) \quad \text{if } x \in \Omega_\varepsilon \equiv \frac{1}{\sqrt{\varepsilon}}\Omega, \quad 0 < t < T_\varepsilon, \\ &v_\varepsilon(x, 0) = \Phi(x) \quad \text{if } |x| < 1, \\ &v_\varepsilon = 0 \text{ on the remaining part of the parabolic boundary of } \Omega_\varepsilon \times (0, T_\varepsilon). \end{aligned} \tag{3.10}$$

It is clear that $v_\varepsilon \uparrow W$ where

$$\begin{aligned} &W_t - \Delta W = f(W) \quad \text{in } \mathbb{R}^n x(0, \tilde{T}), \\ &W(x, 0) = \Phi(x) \quad \text{if } |x| < 1, \\ &W(x, 0) = 0 \quad \text{if } |x| > 1 \end{aligned} \tag{3.11}$$

where $\tilde{T}$ is the blow up time of W, and $T_\varepsilon \downarrow T_0$ where $T_0 \geqq \tilde{T}$. We shall first prove that

$$\tilde{T} > T_* \tag{3.12}$$

and, consequently, $T_0 > T_*$.

Let $\Psi(x) \equiv \Psi_0(|x|) \geqq \Phi(x)$ be such that $\Psi_0(1) = 0$, $\Psi_0'(r) \leqq 0$ and $\Psi_0(0) = \Phi_0$. Denote by $\tilde{u}$ the corresponding solution of (3.11) which is obtained by limit of problems analogous to (3.10). Then $\tilde{u}$ is symmetric about the origin, i.e., $\tilde{u} = U(r, t)$, and it is decreasing in r, by the maximum principle applied to $\partial\tilde{u}/\partial r$. Further, by the maximum principle applied to $\partial\tilde{u}/\partial x_i$ we get that $\partial\tilde{u}/\partial x_i < 0$ in $\{x_i > 0\} \times \{t > 0\}$, and

$$\frac{\partial^2 \tilde{u}}{\partial x_i^2} < 0 \quad \text{on } \{x_i = 0\} \times \{t > 0\};$$

in particular,

$$\Delta \tilde{u}(0, t) < 0.$$

It follows that $\tilde{U}(t) \equiv U(0, t)$ satisfies

$$\tilde{U}' < f(\tilde{U}).$$

Hence the blow-up time $\tilde{t}$ of U satisfies

$$\tilde{t} > \zeta^{-1}(\Psi(0)) = \zeta^{-1}(\Phi_0) = T_*.$$

Since $W \leqq \tilde{u}$ (by comparison), it follows that $\tilde{T} \geqq \tilde{t} > T_*$, and (3.12) is thereby proved.

In order to complete the proof of (iii) it remains to show that $T_0 < T_*$. We proceed as in Theorem 1.2 with $\varepsilon = 1$, taking

$$w = \zeta(V(x, t) - t)$$

where

$$V_t = \Delta V \quad \text{if } |x| < 1, \quad t > 0,$$
$$V(x, 0) = \zeta^{-1}(\Phi(x)) \quad \text{if } |x| < 1,$$
$$V(x, t) = 0 \quad \text{if } |x| = 1, \quad t > 0.$$

Then w is a subsolution and $u \geqq w$ in $\{|x| < 1, t > 0\}$. It follows that $T_\varepsilon \leqq \tilde{t}$ where

$$V(0, \tilde{t}) = \tilde{t}. \tag{3.13}$$

Since $\Phi \geqq 0$ we have

$$\zeta^{-1}(\Phi) \leqq \zeta^{-1}(0) = T^*.$$

Also, $\zeta^{-1}(\Phi(0)) < T^*$ since $\Phi(0) > 0$. Applying the strong maximum principle we conclude that

$$V(x, t) < T^* \quad \text{if } |x| < 1, \quad t > 0$$

and, in particular,

$$V(0, \tilde{t}) < T^*. \tag{3.14}$$

Combining (3.13), (3.14) we deduce that $\tilde{t} < T^*$ and, consequently,

$$\overline{\lim_{\varepsilon \to 0}}\, T_\varepsilon \leqq t < T^*.$$

So far we have proved Theorem 3.1 in the case $\phi_0 \equiv 0$. The same proof extends with minor changes to the case $\phi_0 \not\equiv 0$. The only difference is that now we cannot assert that the corresponding function $v_\varepsilon(x, t) = u(x\sqrt{\varepsilon}, t)$ is monotonically increasing to W and that consequently $\lim T_\varepsilon$ exists (if $\phi_\varepsilon(x)$ is decreasing in every radial direction then $v_\varepsilon(x, t)$ is still increasing in ε).

Remark 3.1. If condition (3.3) is not satisfied, then the blow-up times corresponding to ϕ_0 and ϕ_ε coincide, i.e., the "spike" Φ_ε does not affect the blow-up time for small ε.

Remark 3.2. Theorem 3.1 clearly extends to the case of several spikes of the type Φ_ε with different β's.

Remark 3.3. *The case* $f(0)=0$. So far we have assumed that $f(0)>0$ and consequently $s_0<\infty$. This was needed in the proofs for upper bounds on T_ε (e.g. Theorems 1.2 and 2.1). Consider now the case $f(0)=0$ and assume that for all ε small enough

$$f(\phi)+\varepsilon\Delta\phi \geqq 0 \quad \text{in } \Omega. \tag{3.15}$$

Then, by the maximum principle $\partial u_\varepsilon/\partial t \geqq 0$. It follows that for some small ball $B_\rho(x_0)$ with center x_0 and radius ρ we have

$$u_\varepsilon \geqq \tilde{c} > 0 \text{ on the parabolic boundary of } B_\rho(x_0)\times(0, T_\varepsilon). \tag{3.16}$$

We can now modify the proofs of Theorems 1.2 and 2.1 by constructing subsolutions w only in $B_\rho(x_0)\times(0, t)$; since $\zeta^{-1}(\tilde{c})$ is well defined, the proofs can be extended with just trivial changes.

REFERENCES

[1] A. Friedman and M. McLeod, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425-447.

[2] Y. Giga and R. V. Kohn, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297-319.

[3] A. A. Lacey, *Mathematical analysis of thermal runaway for spatially inhomogeneous reactions*, SIAM J. Appl. Math., 43 (1983), pp. 1350-1366.

[4] ———, *The form of blow-up for nonlinear parabolic equations*,. Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 183-202.

[5] F. B. Weissler, *Single point blow-up of semilinear initial value problems*, J. Differential Equations, 55 (1984), pp. 202-224.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

011

A CRITERION FOR BLOW-UP OF SOLUTIONS TO SEMILINEAR HEAT EQUATIONS*

HAMID BELLOUT†

Abstract. We prove that for the parabolic initial value problem $u_t = \Delta u + \delta f(u)$ there is a finite time blow-up of the solution, provided δ is greater than the upper bound to the spectrum of the steady state problem and (f/f') is concave. An upper bound of the blow-up time is given. The proof is based on a comparison with a subsolution to the parabolic initial value problem.

Key words. blow-up, parabolic, comparison, supersolution

AMS(MOS) subject classification. 35K55

1. The main result. We are interested in the finite time blow-up of the solution to the ·problem:

$$u_t = \Delta u + \delta f(u) \quad \text{in } Q_T, \tag{1.1}$$

$$u(x, 0) = u_0(x) \quad \text{in } D, \tag{1.2}$$

$$\alpha u_\nu + \beta u = 0 \quad \text{on } \Gamma \tag{1.3}$$

where D is a bounded domain in R^n with $C^{2,\mu}$ boundary and

$$Q_T = D \times (0, T), \qquad \Gamma = \partial D \times (0, T),$$

u_0 is a positive continuous function on $\bar{D}$, ν is the outward normal to D, δ is a positive number (which is called the scaled Damköhler number in the combustion theory [3, p. 3]) and α, β are real numbers satisfying

$$\alpha \geqq 0, \quad \beta \geqq 0, \quad \alpha + \beta > 0.$$

Let $R_+ = \{0 < s < \infty\}$. We shall need the following conditions:

$$f \in C^3([0, \infty)), \quad f(0) > 0, \quad f' > 0 \quad \text{on } R_+, \tag{1.4}$$

$$\left(\frac{f}{f'}\right)'' \leqq 0 \quad \text{and} \quad M \equiv \int_0^\infty \frac{ds}{f(s)} < \infty, \tag{1.5}$$

$$\left(\frac{sf'(s)}{f(s)}\right)' \geqq 0 \quad \text{on } R_+. \tag{1.6}$$

By standard results (see, for instance, [7]) there exists a unique positive solution $u(x, t)$ of (1.1)-(1.3) for $0 \leqq t < T$ with either T finite or $+\infty$. If $T < \infty$ then

$$\|u(\cdot, t)\|_{C^0(\bar{D})} \to \infty \quad \text{as } t \to T;$$

in this case we say that there is a finite time blow-up.

Under assumption (1.4), it is well known (Amann [2]) that there is a critical value δ^* such that, for any δ less than δ^*, there is a positive classical solution to the steady state problem:

$$\Delta u + \delta f(u) = 0 \quad \text{in } D, \tag{1.7}$$

$$\alpha u_\nu + \beta u = 0 \quad \text{on } \partial D, \tag{1.8}$$

* Received by the editors June 17, 1985; accepted for publication (in revised form) May 13, 1986.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

while, for δ greater than δ^*, such a solution does not exist. We will need the following definition.

DEFINITION. The spectrum of (1.7), (1.8) is the set of numbers $\delta > 0$ such that a positive classical solution of (1.7), (1.8) exists.

THEOREM 1.[1] *Assume that f satisfies conditions* (1.4), (1.5), *and that if $\alpha \neq 0$ and $\beta \neq 0$ then f also satisfies* (1.6). *If $\delta > \delta^*$ then the solution of* (1.1)-(1.3) *blows up in finite time T, and*

$$(1.9) \qquad T \leqq \frac{4M(2\delta^* + \delta')}{\delta'(4\delta^* + \delta')} \cdot \left[1 - \left[\frac{4\delta^* + \delta'}{2(2\delta^* + \delta')}\right]^{1/2}\right]^{-1} \quad \textit{where } \delta' = \delta - \delta^*.$$

The proof is given in § 2 and is based on comparison of u with a subsolution to the problem (1.1)-(1.3).

The conditions on f are satisfied by a large class of functions such as e^u and $(A+u)^p$ $(p > 1, A > 0)$, as well as functions with growth like $u \cdot (Lnu)^2$.

A. A. Lacey proved that

(i) If $\delta < \delta^*$ then there is blow-up infinite time if u_0 is not "too small" in the average norm;

(ii) If $\delta > \delta^*$ and δ^* is in the spectrum, then the solution of problem (1.1)-(1.3) blows up in finite time.

Although he does not require that $(f/f')'' \leqq 0$, as we did, his condition that δ^* be in the spectrum has been proved only for "small" dimension n, and then only for the Dirichlet problem (see [4]). Thus, in the case of $(A+u)^p$, he requires that $n \leqq 4$ and, for e^u, that $n \leqq 9$.

2. Auxiliary lemmas. Let

$$g(t) = a^2 t^2,$$

$$a = \frac{\delta'}{M} \frac{4\delta^* + \delta'}{4(2\delta^* + \delta')} \left[1 - \left[\frac{4\delta^* + \delta'}{2(2\delta^* + \delta')}\right]^{1/2}\right],$$

$$T_1 = \frac{1}{a},$$

and consider the problem

$$(2.1) \qquad v_t = \Delta v + \delta g(t) f(v) \quad \text{in } Q_T,$$

$$(2.2) \qquad \alpha v_\nu + \beta v = 0 \quad \text{on } \Gamma,$$

$$(2.3) \qquad v(x, 0) = 0 \quad \text{in } D.$$

The local existence, uniqueness and regularity of a solution v to the problem (2.1)-(2.3) follow from Theorem 6.1 of [7, p. 452]. Furthermore, v ceases to exist only by becoming infinite. We want to prove that v blows up at finite time $\leqq T_1$. Without loss of generality, we can assume that v is finite in $D \times [0, T_1)$; then v is $C^{6,3}_{x,t}(\bar{D} \times [0, T_1))$, and $v_t/f(v)$ is in $C^{2,1}_{x,t}(\bar{D} \times [0, T_1))$.

LEMMA 1. *There holds*

$$\left(\frac{v_t}{f(v)}\right)_t \geqq 0 \quad \forall t \leqq T_1.$$

In the special case $f(s) = s^\gamma$ the lemma is due to F. B. Weissler [8]. However, our method of proof is different.

[1] The author expresses his thanks to Professor A. Friedman for suggesting this theorem.

Proof. Dividing (2.1) by $f(v)$ we get

$$\frac{v_t}{f(v)}=\frac{\Delta v}{f(v)}+\delta g.$$

Taking $w=\int_0^v ds/f(s)$, we have that w satisfies

$$w_t=\Delta w+(\nabla w)^2 f'(v)+\delta g(t). \tag{2.4}$$

Since

$$\begin{aligned}((\nabla w)^2 f'(v))_{tt} &= (2\nabla w_t\cdot\nabla w f'+(\nabla w)^2 w_t f''f)_t\\ &=2\nabla w_{tt}\cdot\nabla w f'+2(\nabla w_t)^2 f'+4\nabla w_t\cdot\nabla w w_t f'f\\ &\quad+(\nabla w)^2 w_{tt} f''f+(\nabla w)^2(w_t)^2 f'''f+(\nabla w)^2(w_t)^2 f''f'f,\end{aligned}$$

if we differentiate (2.4) twice with respect to t, we find that the function

$$z\equiv\left(\frac{v_t}{f(v)}\right)_t=w_{tt}$$

satisfies

$$\begin{aligned}L(z)&\equiv z_t-\Delta z-2f'\cdot\nabla z\cdot\nabla w-f''f(\nabla w)^2 z\\ &=\delta g''+(\nabla w)^2(w_t)^2(f'''f^2+f''f'f)+2(\nabla w_t)^2 f'+4\nabla w_t\cdot\nabla w w_t f''f.\end{aligned}$$

Using

$$|\nabla w_t\cdot\nabla w\cdot w_t|\leqq\frac{1}{2}\left[\varepsilon(\nabla w)^2(w_t)^2+\frac{1}{\varepsilon}(\nabla w_t)^2\right]$$

we obtain

$$L(z)\geqq\delta g''+(\nabla w)^2(w_t)^2(f'''f^2+f''f'f-2\varepsilon f''f)+2(\nabla w_t)^2\left(f'-\frac{1}{\varepsilon}f''f\right).$$

Taking $\varepsilon=(f''f/f')$ we get

$$\begin{aligned}L(z)&\geqq\delta g''+(\nabla w)^2(w_t)^2\left[f'''f^2+f''f'-2\frac{f''^2f^2}{f'}\right]\\ &=\delta g''-(\nabla w)^2 f'^2 f\left(\frac{f}{f'}\right)'',\end{aligned}$$

and by assumption (1.5),

$$L(z)\geqq\delta g''>0. \tag{2.5}$$

From (2.3) and the definition of z we have that

$$z(x,0)=0. \tag{2.6}$$

If $\alpha=0$ (respectively, $\beta=0$) then $z(x,t)=0$ (respectively, $z_\nu(x,t)=0$) on Γ.

Since the coefficients of L remain bounded as long as v is bounded, we conclude by the maximum principle that

$$z(x,t)\geqq 0\quad\forall t<T_1.$$

Consider now the case $\alpha \neq 0$, $\beta \neq 0$ and let $b = \beta/\alpha$. Then from (2.2) we have

$$v_\nu = -bv.$$

Differentiating with respect to t and dividing by f we obtain

$$\frac{v_{\nu t}}{f} = -b\frac{v_t}{f}. \tag{2.7}$$

On the other hand,

$$\left(\frac{v_t}{f}\right)_\nu = \frac{v_{\nu t}}{f} - \frac{v_t v_\nu f'}{f^2} = \frac{v_{\nu t}}{f} + bv \cdot \frac{v_t}{f}\frac{f'}{f},$$

so that

$$\frac{v_{\nu t}}{f} = \left(\frac{v_t}{f}\right)_\nu - bv\frac{v_t}{f}\frac{f'}{f}.$$

Substituting into (2.7) we obtain

$$\left(\frac{v_t}{f}\right)_\nu^e = b\frac{v_t}{f}\left(v\frac{f'}{f} - 1\right).$$

Since $w_t = v_t/f$, we then have

$$w_{\nu t} = bw_t\left(v\frac{f'}{f} - 1\right),$$

and differentiating with respect to t we get

$$\begin{aligned} w_{tt\nu} &= bw_{tt}\left(v\frac{f'}{f} - 1\right) + bw_t\left[w_t f' + vw_t f'' - vw_t\frac{f'^2}{f}\right] \\ &= bw_{tt}\left(v\frac{f'}{f} - 1\right) + b(w_t)^2 f'\left[1 + v\left(\frac{f''}{f'} - \frac{f'}{f}\right)\right]. \end{aligned}$$

Since $z = w_{tt}$, this yields

$$z_\nu = bz\left(v\frac{f'}{f} - 1\right) + b(w_t)^2 f' v \cdot \left(Ln\left(v\frac{f'}{f}\right)\right) \geqq bz\left(v\frac{f'}{f} - 1\right) \tag{2.8}$$

where condition (1.6) was used.

Since $v(x, 0) = 0$, there exists an $\varepsilon > 0$ such that the coefficient of z in (2.8) is negative in $\bar{D} \times (0, \varepsilon)$. From (2.5), (2.6), (2.8) we deduce by the strong maximum principle that $z > 0$ in $\bar{D} \times (0, \varepsilon)$.

Suppose z takes negative values in $D \times (0, T_1)$. Then let t^* be the smallest positive time where z has a zero. By (2.5), $z(x, t^*) > 0$ in the interior of D. Thus $z(x, t^*)$ has a zero at a point $x^* \in \partial D$, and by the maximum principle

$$z_\nu(x^*, t^*) < 0.$$

On the other hand, from (2.8), $z_\nu(x^*, t^*) \geqq 0$, a contradiction. This completes the proof that

$$z \geqq 0 \quad \text{in } D \times (0, T_1).$$

LEMMA 2. *There exists a point x_0 in D such that*

$$v(x_0, t) \to \infty \quad \textit{as } t \to T_1.$$

Proof. Let

$$t_0 = \frac{1}{a}\left[\frac{4\delta^* + \delta'}{2(2\delta^* + \delta')}\right]^{1/2}.$$

Then

$$g(t_0)\left(\delta^* + \frac{\delta'}{2}\right) = \delta^* + \frac{\delta'}{4}$$

and, at $t = t_0$, we have

$$v_t = \Delta v + \left(\delta^* + \frac{\delta'}{4}\right) f(v) + \frac{\delta'}{2} g(t_0) f(v). \tag{2.9}$$

Since $\delta' > 0$, the steady state problem (1.7), (1.8) has no solution for $\delta = \delta^* + (\delta'/4)$. By Theorem 1 of [1] the problem (1.7), (1.8) has no positive subsolution; thus there exists x_0 such that at the point (x_0, t_0),

$$\Delta v + \left(\delta^* + \frac{\delta'}{4}\right) f(v) > 0.$$

From (2.9) we therefore have that

$$\frac{v_t}{f(v)} \geqq \frac{\delta'}{2} g(t_0)$$

at the point (x_0, t_0). Also, by Lemma 1

$$\frac{v_t}{f(v)} \geqq \frac{\delta'}{2} g(t_0) \tag{2.10}$$

at (x_0, t) for any $t \geqq t_0$.

Integrating (2.10) with respect to t, $t \in (t_0, T_1)$, we get

$$\int_0^{v(x_0, T_1)} \frac{ds}{f(s)} \geqq \frac{\delta'}{2} g(t_0)(T_1 - t_0).$$

By the choice of T_1, we have that

$$\frac{\delta'}{2} g(t_0)(T_1 - t_0) = M$$

and thus $v(x_0, t) \to \infty$ as $t \to T_1$.

2.1. Proof of Theorem 1. Since $g(t) \leqq 1$ for $t \leqq T_1$, v satisfies

$$v_t \leqq \Delta v + \delta f(v) \quad \text{in } Q_{T_1}.$$

The function $w = u - v$ satisfies

$$w_t \geqq \Delta w + \gamma w, \quad w(x, 0) = 0, \quad \alpha w_\nu + \beta w = 0,$$

where $\gamma = f'(\theta u + (1 - \theta) v)$, and γ is a bounded function as long as u and v are both bounded. By the maximum principle, we conclude that $w \geqq 0$ and $u \geqq v$. Consequently, u blows up at some time $\leqq T_1$.

Remark. Theorem 1 holds for more general equations such as

$$u_t = \Delta u + \sum_{i=0}^{n} a_i(x) i_{x_i} + \delta f(u) + \gamma |\nabla u|^2 \qquad (\gamma \in R^1);$$

the result of [1], [2] used in the proof of Theorem 1 has been generalized to the quasilinear case in [5].

REFERENCES

[1] H. AMANN, *Supersolutions, monotinic iteration and stability*, J. Differential Equations, 21 (1976), pp. 363-377.

[2] ———, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620-709.

[3] J. D. BUCKMASTER AND G. S. S. LUDFORD, *Lectures on Mathematical Combustion*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[4] M. G. CRANDALL AND P. H. RABINOWITZ, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Rational Mech. Anal., 58 (1975), pp. 207-218.

[5] J. DEUEL AND P. HESS, *A criterion for the existence of solutions of nonlinear elliptic boundary value problems*, Proc. Roy. Soc. Edinburgh Sect. A, 74 (1975-75), pp. 49-54.

[6] A. A. LACEY, *Mathematical analysis of thermal runaway for spatially inhomogeneous reactions*, SIAM J. Appl. Math., 43 (1983), pp. 1350-1366.

[7] O. A. LADYZENSKAYA, V. A. SOLONIKOV AND N. N. URAL'TZEVA, *Linear and quasilinear equations of parabolic type*, AMS Translation Monograph 23, American Mathematical Society, Providence, RI, 1968.

[8] F. B. WEISSLER, *Single point blow-up for a semilinear initial value problem*, J. Differential Equations, 55 (1984), pp. 204-224.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

012

LIMITING PROFILES IN CONTAMINANT TRANSPORT THROUGH POROUS MEDIA*

C. J. VAN DUYN† AND J. M. DE GRAAF‡

Abstract. In this paper the following degenerate nonlinear diffusion problem is investigated:

$$\beta(u)_t + u_x = u_{xx}, \qquad t>0, \quad -\infty<x<\infty,$$

$$u(x,0) = u_0(x), \qquad -\infty<x<\infty.$$

For a special choice of β ($\beta(u) = u + u^p$, $p>0$) the equation describes the one-dimensional transport of contaminant in a fluid flow through a homogeneous saturated porous medium.

Here the large time behaviour of the solution of the above problem is studied for more general β. It turns out that, depending upon the shape of β (convex or concave) and the values $u_0(-\infty)$ and $u_0(+\infty)$, the solution converges to a travelling wave g of the form $g(x-at)$ or to a function ω^* of the form $\omega^*(x/(t+1))$.

Key words. degenerate diffusion, large time behaviour, porous media

AMS(MOS) subject classifications. Primary 35B40, 35B50

1. Introduction. Consider the flow of an incompressible fluid through a homogeneous saturated porous medium. It is assumed that the flow is steady-state, one-dimensional and directed in what is chosen to be the positive x-axis. Suppose the fluid is contaminated by a solute, whose concentration C is defined as the mass of solute per unit volume of solution.

If no chemical reactions occur between the solute and the surrounding solid part of the porous medium, then the transport of the contaminant is determined by convection, molecular diffusion and mechanical dispersion; see Bear [1] or Freeze and Cherry [5].

But when chemical reactions do occur, this has to be taken into account when describing the transport process. In this reactive case, let S denote the mass of chemical constituent adsorbed on the solid matrix per unit mass of solid. If the boundary conditions and flow conditions are such that both S and C can be assumed constant in planes perpendicular to the x-axis, which implies $C = C(x, t)$ and $S = S(x, t)$, then mass-conservation for the contaminant leads to the equation (see [1])

$$C_t + \frac{\rho_b}{n} S_t + vC_x = DC_{xx} \tag{1.1}$$

where t and x denote, respectively, a time and a space coordinate, the subscripts t and x denote differentiation with respect to these variables, ρ_b is the bulk mass density of the porous medium, n is the porosity, v is the average fluid velocity and D is the coefficient which incorporates molecular diffusion and mechanical dispersion.

In view of our assumptions, all coefficients in (1.1) can be viewed as being constant and positive. The term $(\rho_b/n)S_t$ in (1.1) represents the change in concentration on the porous matrix caused by adsorption or desorption.

The adsorption reactions for contaminant in groundwater are normally considered as being very rapid with respect to the flow velocity. In view of this, the amount of

* Received by the editors November 5, 1984; accepted for publication (in revised form) April 8, 1986.

† Department of Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands.
‡ Mathematical Institute, University of Leiden, Leiden, The Netherlands.

contaminant that is adsorbed by the solid is commonly assumed to be a function of the concentration in the fluid:

$$S = f(C). \tag{1.2}$$

For a particular contaminant and porous material, (1.2) is determined by laboratory experiments. Since these experiments are normally carried out on soil-samples at constant temperature, relations such as (1.2) are known as *isotherms*. An important example is the Freundlich isotherm, when (1.2) is given by

$$S = K_d C^p, \tag{1.3}$$

where K_d and p are positive constants.

Substitution of the isotherm (1.2) into (1.1) leads to the partial differential equation

$$\left(C + \frac{\rho_b}{n} f(C)\right)_t + vC_x = DC_{xx}. \tag{1.4}$$

This contaminant transport equation is frequently encountered in the groundwater literature (e.g. Bolt [2], further references given there). In particular many numerical models, which are often designed for prediction capabilities, are based on (1.4) or on higher dimensional and more complicated versions of it. When the first author was at the Delft Soil Mechanics Laboratory, he did contaminant transport calculations with a model which is based on a numerical approximation of (1.4) with f given by (1.3). He made the following observations.

Let $p \in (0, 1)$ and let C be kept at some positive value far upstream and at zero far downstream. Then the numerical results indicate that when t increases the contaminant profile C converges towards a travelling front, which moves at a constant velocity. However, reversing the boundary conditions, thus keeping C at zero far upstream and at a positive value far downstream, the numerical results show a flattening profile when t increases.

Note that the first choice of boundary conditions describes the situation where polluted groundwater is contaminating the (originally clean) soil. The second choice corresponds to the situation where polluted soil is washed by uncontaminated groundwater.

For values of $p > 1$, the situation seems to be the reverse of the case $p \in (0, 1)$. Then the numerical calculations indicate a travelling front, which increases from zero (upstream) to some constant positive value (downstream) and a flattening profile when the boundary conditions are reversed.

Inspired by these observations, we study in this paper the large time behaviour of solutions of the initial value problem and of a boundary value problem related to (1.4). We do this for more general isotherms f.

Before we proceed, notice that when f is given by (1.3), (1.4) can be rescaled to an equation with only one parameter. For general f this cannot be done. However, the coefficients in (1.4) are constant and positive and remain fixed throughout this paper. Therefore, we set all coefficients in (1.4) equal to one. Furthermore, let C be denoted by u and define

$$\beta(s) = s + f(s) \quad \text{for } s \geqq 0. \tag{1.5}$$

We consider two problems.

The Cauchy problem.

$$\text{I}\begin{cases} (\beta(u))_t + u_x = u_{xx} & \text{in } S_T, \quad (1.6) \\ u(\cdot, 0) = u_0(\cdot) & \text{on } \mathbb{R}, \quad (1.7) \end{cases}$$

where $S_T=\{(x,t):x\in\mathbb{R}, 0<t\leqq T\}$ in which T is some fixed positive number which eventually will tend to infinity.

The Cauchy-Dirichlet problem.

$$\begin{array}{ll} (1.8) & \\ (1.9) & \\ (1.10) & \end{array}\quad \text{II}\begin{cases}(\beta(u))_t+u_x=u_{xx} & \text{in } H_T,\\ u(0,t)=u^0 & 0<t\leqq T,\\ u(\cdot,0)=u_0(\cdot) & \text{on } \mathbb{R}^+,\end{cases}$$

where $H_T=\{(x,t):x\in\mathbb{R}^+, 0<t\leqq T\}$.

With respect to the function β we assume the following hypotheses.

H1. $\beta:[0,M]\to[0,\infty)$ for some $M>0$ satisfies

(i) $\beta\in C([0,M])\cap C^1((0,M])$; $\beta(0)=0$ and $\beta(s)>0$ on $(0,M]$;

(ii) $\beta'(s)>0$ on $(0,M]$ and $\lim_{s\downarrow 0}(1/(\beta'(s)))$ exists;

(iii) β'' exists and is locally Hölder continuous on $(0,M]$ with exponent $\mu\in(0,1)$;

(iv) $(\beta\cdot\beta'')/(\beta')^2\in L^1(0,M]$.

The initial function u_0 from Problem I is chosen such that H2 is satisfied.

H2. $u_0:\mathbb{R}\to[0,M]$ is uniformly Lipschitz continuous on $\mathbb{R}$.

With respect to the initial function u_0 from Problem II we assume the following.

H2$^+$. $u_0:[0,\infty)\to[0,M]$ satisfies

(i) $u_0\in C([0,\infty))$;

(ii) u_0 is uniformly Lipschitz continuous on $\mathbb{R}^+$;

(iii) $u_0(0)=u^0$, where $u^0\in[0,M]$ is constant.

Equation (1.6) is a nonlinear second order equation of parabolic type. Since $\beta'(s)$ may tend to infinity when s tends to zero, this equation can degenerate at points where its solution vanishes. Therefore we cannot expect Problem I and Problem II to have classical solutions for all admissible choices of β and thus we introduce the notion of weak solutions.

Definition I. $u:\bar{S}_T\to\mathbb{R}$ is a weak solution of Problem I if

(i) $u\in C(\bar{S}_T)$ and u is uniformly bounded and nonnegative;

(ii) u has a bounded generalized derivative with respect to x;

(1.11) (iii) $\iint_{S_T}\{\phi_x(u-u_x)+\phi_t\beta(u)\}\,dx\,dt+\int_{\mathbb{R}}\phi(x,0)\beta(u_0(x))\,dx=0$ for all $\phi\in C^1(S_T)$ which vanish for large $|x|$ and $t=T$.

Definition II. $u:\bar{H}_T\to\mathbb{R}$ is a weak solution of Problem II if

(i) $u\in C(\bar{H}_T)$ and u is uniformly bounded and nonnegative;

(ii) $u(0,t)=u^0$ for all $t\in[0,T]$;

(iii) u has a bounded generalized derivative with respect to x;

(1.12) (iv) $\iint_{H_T}\{\phi_x(u-u_x)+\phi_t\beta(u)\}\,dx\,dt+\int_{\mathbb{R}^+}\phi(x,0)\beta(u_0(x))\,dx=0$ for all $\phi\in C^1(\bar{H}_T)$ which vanish for $x=0$, for large x and for $t=T$.

The hypotheses on β and u_0 assure the existence of a unique weak solution of Problem I and Problem II in $\bar{S}_T$ and $\bar{H}_T$, respectively, for every $T>0$. This solution is a classical solution of (1.6) in a neighbourhood of a point in S_T or H_T where $u>0$, see Gilding [7] and Gilding and Peletier [8].

To study the asymptotic behaviour for $t\to\infty$ of solutions of Problem I and Problem II we set $M=1$ in H1, H2 and H2$^+$, for convenience. With respect to the initial function $u_0:\mathbb{R}\to[0,1]$ from the Cauchy problem we assume the following.

H3. The limits $\lim_{x\to-\infty}u_0(x)=u^-$ and $\lim_{x\to+\infty}u_0(x)=u^+$ exist.

In particular we consider the following two cases:

$$\text{BC1: } u^-=0 \quad\text{and}\quad u^+=1,$$

$$\text{BC2: } u^-=1 \quad\text{and}\quad u^+=0.$$

Similarly, the initial function $u_0:[0,\infty)\to[0,1]$ from the Cauchy-Dirichlet problem satisfies:

H3$^+$. The limit $\lim_{x\to+\infty} u_0(x)=u^+$ exists.

As for the Cauchy problem, we consider the following two cases:

$$\text{BC1}^+\text{: } u^0=0 \quad\text{and}\quad u^+=1,$$

$$\text{BC2}^+\text{: } u^0=1 \quad\text{and}\quad u^+=0.$$

In § 2 we study the existence of travelling wave solutions of (1.6). In particular we look for solutions g of the form $g(x-at)$ $(a>0)$ with $g(-\infty)=u^-$ and $g(+\infty)=u^+$. It will turn out that for given u^- and u^+ the existence of such travelling waves depends critically upon the shape of β. In fact we shall show that (1.6) has a travelling wave solution g with $u^-=0$ and $u^+=1$ (respectively $u^-=1$ and $u^+=0$) if and only if $\beta(s)<\beta(1)\cdot s$ for all $s\in(0,1)$ (respectively $\beta(s)>\beta(1)\cdot s$ for all $s\in(0,1)$). This travelling wave solution is unique modulo translations.

The convergence of solutions of Problem I towards these travelling waves was studied by Osher and Ralston in [10]. In fact, making the additional hypothesis

$$(1.13)\qquad \int_{-\infty}^{0} |\beta(u_0(x))-\beta(u^-)|\,dx<\infty \quad\text{and}\quad \int_{0}^{\infty} |\beta(u_0(x))-\beta(u^+)|\,dx<\infty,$$

Osher and Ralston proved the next theorem with semigroup methods.

THEOREM 1. *Let assumptions* H1, H2, H3 *and* (1.13) *hold and suppose* u_0 *and* β *are such that there exists a travelling wave* $g(x-at)$. *Then the solution* u *of Problem* I *converges to* g *in the following sense*:

$$\lim_{t\to\infty}\int_{-\infty}^{\infty} |\beta(u(x,t))-\beta(g(x-y-at))|\,dx=0,$$

where $y\in\mathbb{R}$ *is such that* $\int_{-\infty}^{\infty}\{\beta(u_0(x))-\beta(g(x-y))\}\,dx=0$.

The convergence of solutions of Problem II, with u_0 satisfying BC2$^+$ and with β such that $0<\beta'<\infty$, was studied by Khusnytdinova in [9]. She obtained exponential convergence towards a travelling wave. However, her proof depends critically upon the upper bound of β' on $[0,1]$.

In §§ 3 and 4 we study Problem I when u_0 and β are such that no travelling wave exists. To be definite we shall assume that u_0 satisfies BC1 and that $\beta''<0$. To investigate the behaviour of the solution of this problem as $t\to\infty$ it will be convenient to change to the new independent variables

$$\eta:=\frac{x}{t+1},\qquad \tau:=\log(t+1).$$

Then $u(x,t)$ is a weak solution of Problem I if and only if $w(\eta,\tau):=u(x,t)$ is a weak solution of the transformed problem;

$$P\begin{cases}(\beta(w))_\tau+(1-\eta\beta'(w))w_\eta=e^{-\tau}w_{\eta\eta} & \text{in } S_{T'}, \qquad (1.14)\\ w(\cdot,0)=u_0(\cdot) & \text{on } \mathbb{R} \qquad (1.15)\end{cases}$$

where $S_{T'}=\{(\eta,\tau)\colon \eta\in\mathbb{R}, 0<\tau\leqq T'\}$ and $T'=\log(T+1)$.

Here a weak solution is a function that satisfies the following.

DEFINITION. $w:\bar{S}_{T'}\to\mathbb{R}$ is a weak solution of Problem P if

(i) $w\in C(\bar{S}_{T'})$ and w is uniformly bounded and nonnegative;

(ii) w has a bounded generalized derivative with respect to η;

(1.16) (iii) $\iint_{S_{T'}} \{\beta(w)\phi_\tau + [w - \eta\beta(w) - w_\eta e^{-\tau}]\phi_\eta - \beta(w)\phi\}\, d\eta\, d\tau$
$+\int_{\mathbb{R}} \beta(u_0(\eta))\phi(\eta, 0)\, d\eta = 0$
for all $\phi \in C^1(\bar{S}_{T'})$ which vanish for large $|\eta|$ and for $\tau = T'$.

To prove our results we need the following additional hypotheses on the functions β and u_0.

H4. (i) $\beta \in C^\alpha([0,1])$ for some $\alpha \in (0,1)$;
(ii) $-\beta''/(\beta')^2 \geqq \nu > 0$ on $(0,1)$ for some $\nu > 0$;
(iii) β''' exists and $(\beta''' \cdot \beta')/(\beta'')^2 < 2$ on $(0,1)$.

H5. There exist constants $k_1 > 1/\alpha$, $k_2 > 1$ such that
(1.17) (i) $u_0(x) = O((-x)^{-k_1})$ when $x \to -\infty$,
(1.18) (ii) $u_0(x) - 1 = O(x^{-k_2})$ when $x \to +\infty$.
Here α is the Hölder exponent defined in H4.

We shall show that $w(\eta, \tau)$ converges, as $\tau \to \infty$, to a function w^* which is a solution of the reduced problem:

$$\bar{P}_\infty \begin{cases} (1 - \eta\beta'(w))\dfrac{dw}{d\eta} = 0 & \text{on } \mathbb{R}, \\ w(-\infty) = 0, & w(+\infty) = 1. \end{cases}$$

Using the notation $u^*(x,t) = w^*(\eta)$ we prove the following theorem.

THEOREM 2. *Suppose β and u_0 satisfy* H1 *and* H4, *respectively* H2 *and* H5. *Let $u(x,t)$ be the weak solution of Problem* I. *Then the following estimate holds*:

$$\sup_{x \in \mathbb{R}} |\beta(u(x,t)) - \beta(u^*(x,t))| \leqq (t+1)^{-\alpha/(\alpha+1)}(A_1 + A_2 \log(t+1))^{\alpha/(\alpha+1)}$$

for all $t \geqq 0$, where A_1 and A_2 are positive constants.

Remark. In the case of the Freundlich isotherm, with $K_d = 1$ in (1.3) for convenience, $\beta(s) = s + s^p$. Then when $p \in (0,1)$ Theorem 2 implies that

$$\sup_{x \in \mathbb{R}} |\beta(u(x,t)) - \beta(u^*(x,t))| \leqq (t+1)^{-p/(p+1)}(A_1 + A_2 \log(t+1))^{p/(p+1)}$$

for all $t \geqq 0$.

For this choice of β the function w^* is explicitly known and is given by (3.8) in § 3. In terms of the variables x and t, u^* behaves as in Fig. 1. Indeed this gives a limiting profile, which becomes flatter when t increases.

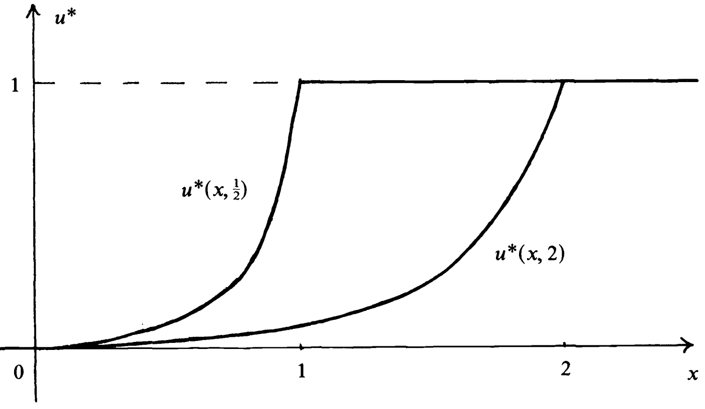


FIG. 1. *The function u^* for the case $\beta(s) = s + s^{1/2}$.*

For the Cauchy-Dirichlet problem we have a similar result. We make the following assumption on the initial function u_0.

H5$^+$. There exists a constant $k_0 > 1/\alpha$ such that

$$u_0(x) - u^+ = O(x^{-k_0}) \quad \text{when } x \to +\infty. \tag{1.19}$$

Then the following theorem can be proved.

THEOREM 3. *Suppose* β *satisfies* H1, H4 *and* u_0 *satisfies* H2$^+$, H5$^+$. *Let* $u(x, t)$ *be the weak solution of Problem* II. *Then the following estimate holds*:

$$\sup_{x\in\mathbb{R}} |\beta(u(x,t)) - \beta(u^*(x,t))| \leqq (t+1)^{-\alpha/(\alpha+1)}(A_1^+ + A_2^+ \log(t+1))^{\alpha/(\alpha+1)}$$

for all $t \geqq 0$, *where* A_1^+ *and* A_2^+ *are positive constants and* $u^*(x, t) = w^*(\eta)$ *with* w^* *the solution of* P_∞.

Finally note that Theorems 2 and 3 only partially answer the questions raised by the numerical computations. In particular, when u_0 satisfies the boundary conditions BC2 or BC2$^+$ and when $\beta(s) = s + s^p$ with $p > 1$, our results do not apply. With respect to the convergence towards travelling waves, it is of interest to generalize the results of Khusnytdinova [9] for the Cauchy-Dirichlet problem and to obtain convergence estimates for the Cauchy problem studied by Osher and Ralston [10]. We leave these questions for future study.

2. Travelling waves. In this section we study the existence and uniqueness of travelling wave solutions of Problem I.

DEFINITION. The pair (a, g) with $a \in \mathbb{R}$ and $g: \mathbb{R} \to [0, 1]$ is called a travelling wave solution of Problem I if

(i) $\beta(g)$ and g' are absolutely continuous on $\mathbb{R}$;

(ii) $a(\beta(g))' - g' = -g''$ a.e. on $\mathbb{R}$;

(iii) $g(-\infty) = 0$; $g(+\infty) = 1$ (in case BC1);
$g(-\infty) = 1$; $g(+\infty) = 0$ (in case BC2).

THEOREM 4. *Let* β *satisfy assumption* H1(i). *Then*

(i) *Problem* I *has a unique travelling wave solution* (a, g) *satisfying* BC1 *if and only if* $\beta(s) < \beta(1) \cdot s$ *for all* $s \in (0, 1)$.

(ii) *Problem* I *has a unique travelling wave solution* (a, g) *satisfying* BC2 *if and only if* $\beta(s) > \beta(1) \cdot s$ *for all* $s \in (0, 1)$.

Remarks. (1) The uniqueness in the theorem must be understood modulo translations. (2) The unique speed a of the travelling wave in the theorem is given by $a = 1/(\beta(1))$.

Proof. We shall only prove part (ii) of the theorem because part (i) can be handled in exactly the same way.

Define $\eta := x - at$. Then if (a, g) is a travelling wave solution of Problem I, a and g satisfy

$$a(\beta(g))' - g' = -g'' \quad \text{a.e. on } \mathbb{R}, \tag{2.1}$$

$$g(-\infty) = 1, \qquad g(+\infty) = 0, \tag{2.2}$$

where primes denote differentiation. Integration of (2.1) yields

$$a\beta(g(\eta)) - g(\eta) + A = -g'(\eta) \quad \text{for } \eta \in \mathbb{R}, \tag{2.3}$$

where A is a constant. Because g' is continuous we have

$$g(\eta + h) - g(\eta) = hg'(\eta + \theta h) = h(-a\beta(g(\eta + \theta h)) + g(\eta + \theta h) - A)$$
$$(0 < \theta(\eta) < 1, h > 0).$$

Taking the limit $\eta \to +\infty$ respectively $\eta \to -\infty$ in this expression we find

$$0 = a\beta(0) - 0 + A \quad \text{and} \quad 0 = a\beta(1) - 1 + A.$$

Since $\beta(0) = 0$, $A = 0$ and hence for the speed a we have

$$a = \frac{1}{\beta(1)}. \tag{2.4}$$

Substituting a and A in (2.3) we derive the following equation for g

$$g' = g - \frac{1}{\beta(1)} \cdot \beta(g) \quad \text{on } \mathbb{R}. \tag{2.5}$$

Since $\beta \in C^1((0, 1])$ it follows from (2.5) that when $g(\eta_0) > 0$ for some $\eta_0 \in \mathbb{R}$, there exists a neighbourhood of the point η_0 where g is a classical solution.

Suppose that

$$\beta(s) > \beta(1) \cdot s \quad \text{on } (0, 1). \tag{$*$}$$

Then (2.5) has a unique (modulo translations) solution given by

$$\int_{1/2}^{g(\eta)} \frac{ds}{s - (\beta(s)/\beta(1))} = \eta,$$

where we have chosen $g(0) = 1/2$.

Note that the solution vanishes for finite $\eta > 0$ if and only if $(\beta(s) - s\beta(1))^{-1} \in L^1(0, \frac{1}{2})$. Furthermore, since $(\beta(s) - s\beta(1))^{-1} \notin L^1(\frac{1}{2}, 1)$, $g(\eta) < 1$ on $\mathbb{R}$ and $g(\eta)$ tends to 1 as $\eta \to -\infty$.

Next suppose that there exist points where ($*$) does not hold. It is clear from (2.5) that if $\beta(\hat{g}) < \beta(1) \cdot \hat{g}$ at some point $\hat{g} \in (0, 1)$, then a travelling wave which satisfies (2.2) cannot exist. If $\beta(\hat{g}) = \beta(1) \cdot \hat{g}$ at some point $\hat{g} \in (0, 1)$, then it follows from the local uniqueness of solutions of (2.5) that the only solution for which $g(\eta_0) = \hat{g}$ at some point $\eta_0 \in \mathbb{R}$ is the constant solution $g = \hat{g}$ and hence (2.2) is not satisfied.

Finally note that if (a, g) is a travelling wave solution, then g is the weak solution of Problem I (for an appropriate choice of u_0). The proof follows by direct calculation. □

Remark. Suppose there exists a $s_0 \in (0, 1)$ where β' does not exist and where $\beta(s_0) = \beta(1) \cdot s_0$. Then if β is Lipschitz continuous on $[0, 1]$, the standard uniqueness result for ordinary differential equations gives that a travelling wave does not exist. If on the other hand $\beta(s) - \beta(1)s = O(|s - s_0|^\sigma)$ with $\sigma \in (0, 1)$, then a travelling wave can still be constructed. However for this case there is no uniqueness.

In the case of the Freundlich isotherm $f(s) = s^p$, we have the following.

Example. Choose $\beta(s) = s + s^p$ with $0 < p < 1$. Then β satisfies H1 and $\beta'' < 0$. Therefore we are in case (ii) of Theorem 4, so we know that a travelling wave solution (a, g) exists with $g(-\infty) = 1$, $g(+\infty) = 0$ and $a = 1/2$.

For this choice of β we can explicitly calculate the travelling wave

$$g(\eta) = \begin{cases} [1 - e^{\lambda(\eta - \eta_0)}]^{1/(1-p)} & \text{if } \eta < \eta_0 \\ 0 & \text{if } \eta \geqq \eta_0 \end{cases} \quad \text{for some } \eta_0,$$

where $\eta = x - t/2$ and $\lambda = (1 - p)/2$.

A trivial consequence of Theorem 4 is the following.

COROLLARY 1. *Suppose β satisfies* H1(i) *and β'' exists and is of one sign. Then*

(i) *Problem* I *has a unique travelling wave solution (a, g) satisfying* BC1 *if and only if $\beta''(s) > 0$ for all $s \in (0, 1)$.*

(ii) *Problem* I *has a unique travelling wave solution (a, g) satisfying* BC2 *if and only if $\beta''(s) < 0$ for all $s \in (0, 1)$.*

3. The function w^*. In this section we study the problem:

$$P_\infty \begin{cases} (1-\eta\beta'(w))\dfrac{dw}{d\eta}=0 & \text{on } \mathbb{R}, \\ w(-\infty)=0, & w(+\infty)=1, \end{cases}$$

where the prime denotes differentiation with respect to the argument of β. The solution of this problem arises in the study of the large time behaviour of solutions of Problem P (or Problem I) in the case where u_0 satisfies BC1 and $\beta''<0$. This will be treated in § 4.

DEFINITION. $w^*:\mathbb{R}\to\mathbb{R}$ is a weak solution of Problem P_∞ if it satisfies

(i) w^* is absolutely continuous on $\mathbb{R}$ and $0 \leqq w^* \leqq 1$;

(ii) $w^*(-\infty)=0$; $w^*(+\infty)=1$;

(iii) $\int_{\mathbb{R}} \{(w^*-\eta\beta(w^*))\phi_\eta - \beta(w^*)\phi\}\, d\eta = 0$
for all $\phi \in C^1(\mathbb{R})$ which vanish for large $|\eta|$.

THEOREM 5. *If β satisfies* H1(i, ii) *and β' is monotone decreasing, then Problem P_∞ has one and only one weak solution.*

Proof. Let w^* be a weak solution of Problem P_∞. Then it follows from (iii) and the continuity of β that

$$\frac{d}{d\eta}(w^*-\eta\beta(w^*)) = -\beta(w^*) \quad \text{on } \mathbb{R} \tag{3.1}$$

and $(w^*-\eta\beta(w^*)) \in C^1(\mathbb{R})$. Integrating (3.1) with respect to η from η_1 to η_2 gives

$$[w^*(\eta_2)-\eta_2\beta(w^*(\eta_2))]-[w^*(\eta_1)-\eta_1\beta(w^*(\eta_1))] = -\int_{\eta_1}^{\eta_2} \beta(w^*(s))\, ds. \tag{3.2}$$

Suppose that $w^*(\eta_1)=w^*(\eta_2)$ and $w^*(\eta)>w^*(\eta_1)$ on (η_1, η_2) (respectively $w^*(\eta) < w^*(\eta_1)$ on (η_1, η_2)). Then (3.2) yields

$$(\eta_2-\eta_1)\beta(w^*(\eta_1)) = \int_{\eta_1}^{\eta_2} \beta(w^*(s))\, ds,$$

which gives a contradiction. This means that w^* is nondecreasing, i.e.,

$$\frac{dw^*}{d\eta} \geqq 0 \quad \text{a.e. on } \mathbb{R}. \tag{3.3}$$

Next we assume that $w^*(\eta_0)>0$ for some $\eta_0<0$. Then $w^*>0$ in a neighbourhood N of η_0 with $N \subset \mathbb{R}^-$. Using the smoothness of β in (3.1) it follows that

$$(1-\eta\beta'(w^*))\frac{dw^*}{d\eta}=0 \quad \text{a.e. in } N.$$

But $1-\eta\beta'(w^*)>1$ in N. Therefore $(dw^*)/d\eta=0$ a.e. in N, which implies that $w^* \equiv 0$ on $\overline{\mathbb{R}^-}$.

Define

$$\eta^* := \sup\{\eta \in \mathbb{R}\colon w^*(\eta) = 0\}. \tag{3.4}$$

Then $\eta^* \geqq 0$ and w^* satisfies

$$w^*(\eta)\begin{cases} =0 & \text{if } \eta \leqq \eta^*, \\ >0 & \text{if } \eta > \eta^*. \end{cases} \tag{3.5}$$

Furthermore

$$(1 - \eta\beta'(w^*))\frac{dw^*}{d\eta} = 0 \quad \text{a.e. on } (\eta^*, \infty). \tag{3.6}$$

Finally we show that w^* is strictly increasing when $0 < w^* < 1$. Suppose not; then there exist points $\eta^* < \eta_1 < \eta_2 < \eta_3 < \eta_4$ such that $(dw^*)/d\eta > 0$ a.e. on $(\eta_1, \eta_2) \cup (\eta_3, \eta_4)$, and $w^*(\eta) \equiv \alpha_0 > 0$ on $[\eta_2, \eta_3]$. Then from (3.6) we have

$$\frac{1}{\eta_2} = \beta'(w^*(\eta_2)) = \beta'(\alpha_0) = \beta'(w^*(\eta_3)) = \frac{1}{\eta_3},$$

which proves the assertion.

Thus if $\tilde{\eta}$ is such that $1 - \tilde{\eta}\beta'(1) = 0$, we may conclude that a weak solution of Problem P_∞ is of the form

$$w^*(\eta)\begin{cases} =0 & \text{on } (-\infty, \eta^*], \\ \text{satisfies } 1 - \eta\beta'(w^*) = 0 & \text{on } (\eta^*, \tilde{\eta}), \\ =1 & \text{on } [\tilde{\eta}, \infty). \end{cases} \tag{3.7}$$ □

Remark. The number η^* in (3.4) is given by

$$\eta^* = \lim_{s \downarrow 0} \frac{1}{\beta'(s)}.$$

Example. In the case $\beta(s) = s + s^p$ with $0 < p < 1$ we can calculate w^* explicitly and find

$$w^*(\eta) = \begin{cases} 0 & \text{if } \eta \leqq 0, \\ \left(\dfrac{p\eta}{1-\eta}\right)^{1/(1-p)} & \text{if } 0 < \eta < \dfrac{1}{1+p}, \\ 1 & \text{if } \eta \geqq \dfrac{1}{1+p}. \end{cases} \tag{3.8}$$

4. Convergence to the function $u^*(x, t)$ for the Cauchy problem. In this section we study the asymptotic behaviour as $t \to \infty$ of the solution $u(x, t)$ of Problem I when u_0 satisfies BC1 and $\beta'' < 0$. From the corollary in § 2 we know that Problem I does not have a travelling wave solution in this case.

Throughout this section we assume that β satisfies H1 and H4 and u_0 satisfies H2 and H5.

Set

$$\eta = \frac{x}{t+1} \quad \text{and} \quad \tau = \log(t+1) \tag{4.1}$$

and consider the transformed problem

$$P\begin{cases}(\beta(w))_\tau+(1-\eta\beta'(w))w_\eta=e^{-\tau}w_{\eta\eta} & \text{in } S_T,\\ w(\cdot,0)=u_0(\cdot) & \text{on } \mathbb{R}.\end{cases}$$

The main result of this section is the following.

PROPOSITION 1. *Let $w(\eta,\tau)$ be the weak solution of Problem P with β satisfying* H1, H4 *and u_0 satisfying* H2, H5. *Let $w^*(\eta)$ be the weak solution of Problem P_∞. Then the following integral estimate holds*:

$$\int_{-\infty}^{\infty}|\beta(w(\eta,\tau))-\beta(w^*(\eta))|\,d\eta\leqq(C_1+C_2\tau)\,e^{-\tau} \tag{4.2}$$

for all $\tau\in[0,T']$, where C_1 and C_2 are positive constants.

Proof. The proof consists of two parts. In the first part we approximate the initial function u_0 by a sequence of smooth positive functions $\{u_{0n}\}$. Then we consider the family of problems $\{P_n\}$ obtained from Problem P by using u_{0n} as the initial function. Next we show that the classical solution w_n of the corresponding Problem P_n converges towards a function w_n^* as $\tau\to\infty$, where w_n^* is an approximation of w^*.

In the second part we let n tend to infinity and obtain the desired integral estimate (4.2).

Without loss of generality we may assume that u_0 is nondecreasing. In the case where u_0 is nonmonotone, two nondecreasing functions ψ_1 and ψ_2 can always be found such that $\psi_1\leqq u_0\leqq\psi_2$ on $\mathbb{R}$. Then if Proposition 1 holds for the monotone initial functions ψ_1 and ψ_2, it follows by a maximum principle argument (see [7]) that the integral estimate (4.2) also holds in the case of arbitrary initial functions u_0.

Part 1. Because the initial function u_0 is uniformly Lipschitz continuous and nondecreasing on $\mathbb{R}$ we can construct a sequence of smooth functions $\{u_{0n}\}_{n=1}^{\infty}$ such that $u_{0n}\to u_0$ as $n\to\infty$, uniformly on $\mathbb{R}$, and where every function u_{0n} satisfies

(i) $u_{0n}\in C^\infty(\mathbb{R})$ and $0\leqq u_{0n}'\leqq M^*$ (for some constant $M^*>0$, independent of n),

(ii) $u_{0n}(-\infty)=1/n$; $u_{0n}(+\infty)=1$.

Now consider for each $n\in\mathbb{N}$ the approximate problem:

$$P_n\begin{cases}(\beta(w))_\tau+(1-\eta\beta'(w))w_\eta=e^{-\tau}w_{\eta\eta} & \text{in } S_T,\\ w(\cdot,0)=u_{0n}(\cdot) & \text{on } \mathbb{R}.\end{cases} \tag{4.3}$$

This problem has a unique solution $w_n\in C^{3+\mu}(\bar{S}_{T'})$, with $\mu\in(0,1)$ as given in H1(iii), and $1/n\leqq w_n\leqq 1$. This follows directly from the observation that w_n is the transformed of u_n, where u_n is the unique classical solution of Problem I with initial function u_{0n} (see e.g. van Duyn and Ye [4]).

We want to show that w_n tends to an approximation of w^* as $\tau\to\infty$, where this approximation is defined by

$$w_n^*=\max\left\{w^*,\frac{1}{n}\right\},\qquad n\in\mathbb{N}. \tag{4.4}$$

Note that the function w_n^* satisfies

$$\begin{aligned}&(1-\eta\beta'(w_n^*))\frac{dw_n^*}{d\eta}=0\quad\text{a.e. on }\mathbb{R},\\ &w_n^*(-\infty)=\frac{1}{n},\qquad w_n^*(+\infty)=1.\end{aligned} \tag{4.5}$$

In order to prove the convergence of w_n to w_n^* we need the following lemma.

LEMMA 1. *There exists a constant $M_0 > 0$ such that for all $n \in \mathbb{N}$*

$$0 \leqq w_{n\eta} \leqq M_0 \quad \text{in } S_{T'}.$$

Proof. Differentiate (4.3) with respect to η. Then $v := w_{n\eta}$ satisfies the following problem;

$$P'_n \begin{cases} \beta'(w_n)v_\tau + (1-\eta\beta'(w_n))v_\eta + \left\{ -\dfrac{\beta''(w_n)}{\beta'(w_n)} v - \beta'(w_n) \right\} v \\ \qquad + \dfrac{\beta''(w_n)}{\beta'(w_n)} e^{-\tau} v v_\eta = e^{-\tau} v_{\eta\eta} \quad \text{in } S_{T'}, \\ v(\cdot, 0) = u'_{0n}(\cdot) \qquad\qquad \text{on } \mathbb{R}. \end{cases}$$

It is known that $0 \leqq u'_{0n} \leqq M^*$. Then obviously $\underline{v} \equiv 0$ is a subsolution for Problem P'_n. Furthermore the function $\bar{v} \equiv M_0$, with $M_0 = \max\{M^*, 1/\nu\}$, is a supersolution for Problem P'_n. This follows from H4(ii). Then Cosner [3, Lemma 1] implies that $0 \leqq w_{n\eta} \leqq M_0$, where the constant M_0 does not depend on n and τ.

To investigate the asymptotic behaviour of w_n for large $|\eta|$ we construct a time-independent sub- and supersolution.

LEMMA 2. (i) *There exist constants $\gamma > 0$, $1 < k < k_2$, $\eta_0 > 1$ such that the function $\underline{s}_n(\eta)$, for each $n \in \mathbb{N}$ defined by*

$$\underline{s}_n(\eta) := \begin{cases} \dfrac{1}{n}, & \eta \leqq \eta_0 \\ \max\left\{\dfrac{1}{n}, 1 - \gamma(\eta - \eta_0)^{-k}\right\}, & \eta > \eta_0, \end{cases} \tag{4.6}$$

satisfies: $\underline{s}_n$ *is continuous on* $\mathbb{R}$ *and* $w_n(\cdot, \tau) \geqq \underline{s}_n(\cdot)$ *on* $\mathbb{R}$ *for all* $0 \leqq \tau \leqq T'$ *and for all* $n \in \mathbb{N}$.

(ii) *There exist constants $\tilde{\gamma} > 0$, $1/\alpha < \tilde{k} < k_1$, $\eta_1 < -1$ such that the function $\bar{s}_n(\eta)$, for each $n \in \mathbb{N}$ defined by*

$$\bar{s}_n(\eta) := \begin{cases} 1, & \eta \geqq \eta_1, \\ \min\left\{1, \tilde{\gamma}(-\eta + \eta_1)^{-\tilde{k}} + \dfrac{1}{n}\right\}, & \eta < \eta_1, \end{cases} \tag{4.7}$$

satisfies: $\bar{s}_n$ *is continuous on* $\mathbb{R}$ *and* $w_n(\cdot, \tau) \leqq \bar{s}_n(\cdot)$ *on* $\mathbb{R}$ *for all* $0 \leqq \tau \leqq T'$ *and for all* $n \in \mathbb{N}$.

Proof. For each $n \in \mathbb{N}$ the approximating initial functions u_{0n}, used in Problem P_n, are given by

$$u_{0n}(x) = \int_{\mathbb{R}} \rho_n(x-y) u_0^+(y)\, dy.$$

Here $u_0^+ = \max\{u_0, 1/n\}$ and $\rho_n(x) = n\rho(nx)$, where $\rho \in C^\infty(\mathbb{R})$ is such that $\operatorname{Supp} \rho = [-1, 1]$, $\rho(x) \geqq 0$ on $\mathbb{R}$ and $\|\rho\|_{L^1(\mathbb{R})} = 1$.

Below we use the following property of the functions u_{0n}: for any increasing function f defined on $\mathbb{R}$, let $f^+ = \max\{f, 1/n\}$. Then for each $x_0 \geqq 1$ we have:

(a) If $u_0 \geqq f$ on $\mathbb{R}$, then $u_{0n}(x) \geqq f^+(x - x_0)$ for all $n \in \mathbb{N}$ and for all $x \in \mathbb{R}$;

(b) If $u_0 \leqq f$ on $\mathbb{R}$, then $u_{0n}(x) \leqq f^+(x + x_0)$ for all $n \in \mathbb{N}$ and for all $x \in \mathbb{R}$.

We shall only prove part (i) of Lemma 2. The proof of part (ii) is almost identical and is therefore omitted.

Choose constants $\gamma > 0$, $k > 1$, $\hat{\eta} \geqq (\beta'(1))^{-1}(1 + (k+1)\gamma^{-1/k})$ such that $\mu_0(\eta) \geqq 1 - \gamma(\eta - \hat{\eta})^{-k}$ for $\eta > \hat{\eta}$. Because μ_0 satisfies H5 this is possible if we choose $k < k_2$.

Next let $\eta^* = \hat{\eta} + \gamma^{1/k}$ and define

$$\underline{s}(\eta) := \begin{cases} 0, & \eta \leqq \eta^*, \\ 1 - \gamma(\eta - \hat{\eta})^{-k}, & \eta > \eta^*. \end{cases}$$

From property (a) we know that there exists a constant $\xi_0 > 1$ such that $u_{0n}(\eta) \geqq \underline{s}(\eta - \xi_0)$ on $\mathbb{R}$. Thus if we choose in (4.6) $\eta_0 = \hat{\eta} + \xi_0$, then $u_{0n}(\eta) \geqq \underline{s}_n(\eta)$ on $\mathbb{R}$. Furthermore if

$$\mathscr{L}u := (\beta(u))_\tau + (1 - \eta\beta'(u))u_\eta - e^{-\tau}u_{\eta\eta},$$

then an elementary computation shows that $\mathscr{L}\underline{s}_n \leqq 0$ for $\eta > \bar{\eta}$. Applying [3, Lemma 1] to the difference $w_n - \underline{s}_n$, part (i) of the lemma follows immediately. □

Thus Lemma 2 gives

$$\underline{s}_n(\cdot) \leqq w_n(\cdot, \tau) \leqq \bar{s}_n(\cdot) \quad \text{on } \mathbb{R} \tag{4.8}$$

for all $\tau \in [0, T']$ and for all $n \in \mathbb{N}$.

From (4.8) and H1(i) we may now conclude that $(\beta(w_n(\cdot, \tau)) - \beta(w_n^*(\cdot))) \in L^1(\mathbb{R})$ for all $\tau \in [0, T']$.

To estimate

$$\|\beta(w_n(\tau)) - \beta(w_n^*)\|_{L^1(\mathbb{R})} := \int_{-\infty}^{\infty} |\beta(w_n(\eta, \tau)) - \beta(w_n^*(\eta))| \, d\eta$$

it will be convenient to consider the following auxiliary problem:

$$\tilde{P}_n \begin{cases} (\beta(w))_\tau + (1 - \eta\beta'(w))w_\eta = e^{-\tau}w_{\eta\eta} & \text{in } H^{\tilde{\eta}}_{T'}, \\ w(\tilde{\eta}, \tau) = 1, \qquad 0 < \tau \leqq T', \\ w(\cdot, 0) = \tilde{u}_{0n}(\cdot) \quad \text{on } (-\infty, \tilde{\eta}), \end{cases}$$

where $H^{\tilde{\eta}}_{T'} = \{(\eta, \tau)\colon \eta \in (-\infty, \tilde{\eta}),\ 0 < \tau \leqq T'\}$. In this problem $\tilde{\eta}$ is chosen such that $w^*(\tilde{\eta}) = 1$ and $w^*(\eta) < 1$ if $\eta < \tilde{\eta}$, thus $\tilde{\eta} = 1/\beta'(1)$ (see § 3). The function $\tilde{u}_{0n}$ is given by $\tilde{u}_{0n} = \max\{u_{0n}, w_n^*\}$.

We want to transform Problem $\tilde{P}_n$ into one in which the differential equation has bounded coefficients and in which the boundary condition is essentially unchanged. To this end we set

$$y = (\eta - \tilde{\eta})\, e^\tau \quad \text{and} \quad t = e^\tau - 1.$$

Then $w(\eta, \tau)$ is a classical solution of Problem $\tilde{P}_n$ if and only if $h(y, t) := w(\eta, \tau)$ is a classical solution of the problem

$$\begin{aligned} &(\beta(h))_t + (h - \tilde{\eta}\beta(h))_y = h_{yy} \quad \text{in } H_T^- = (-\infty, 0) \times (0, T], \\ &h(0, t) = 1, \qquad 0 < t \leqq T, \\ &h(y, 0) = \tilde{u}_{0n}(y + \tilde{\eta}) \quad \text{on } (-\infty, 0). \end{aligned}$$

Now this problem has a unique classical solution $h_n \in C^{2,1}(H_T^-) \cap C(\overline{H_T^-})$ (see e.g. [7]). Then $v_n(\eta, \tau) := h_n(y, t)$ is the unique solution of Problem $\tilde{P}_n$ such that $v_n \in C^{2,1}(H^{\tilde{\eta}}_{T'}) \cap (\overline{H^{\tilde{\eta}}_{T'}})$ and $1/n \leqq v_n \leqq 1$.

If we define the function $\tilde{w}_n$ as

$$\tilde{w}_n(\eta, \tau) := \begin{cases} v_n(\eta, \tau) & \eta < \tilde{\eta}, \\ 1 & \eta \geqq \tilde{\eta}, \end{cases}$$

then by a maximum principle argument we have

$$\tilde{w}_n \geqq w_n \quad \text{in } \overline{S_{T'}}. \tag{4.9}$$

Furthermore it can be proved by direct calculation that w_n^* is a subsolution for Problem $\tilde{P}_n$, where we use the fact that β satisfies H4(iii). Also it can be shown that $\bar{s}_n(\eta)$ is a supersolution for Problem $\tilde{P}_n$. Therefore

$$w_n^*(\cdot) \leqq \tilde{w}_n(\cdot, \tau) \leqq \bar{s}_n(\cdot) \quad \text{on } \mathbb{R} \tag{4.10}$$

for all $\tau \in [0, T']$ and for all $n \in \mathbb{N}$.

Now it follows from (4.8), (4.10) and the definition of $\underline{s}_n$, $\bar{s}_n$ that $(\beta(\tilde{w}_n(\cdot, \tau)) - \beta(w_n(\cdot, \tau))) \in L^1(\mathbb{R})$ and $(\beta(\tilde{w}_n(\cdot, \tau)) - \beta(w_n^*(\cdot))) \in L^1(\mathbb{R})$ for all $\tau \in [0, T']$.

Since

$$\begin{aligned} \|\beta(w_n(\tau)) - \beta(w_n^*)\|_{L^1(\mathbb{R})} \leqq \|\beta(w_n(\tau)) - \beta(\tilde{w}_n(\tau))\|_{L^1(\mathbb{R})} \\ + \|\beta(\tilde{w}_n(\tau)) - \beta(w_n^*)\|_{L^1(\mathbb{R})}, \end{aligned} \tag{4.11}$$

it is sufficient to find estimates for the norms on the right-hand side of (4.11).

LEMMA 3. (i) $\|\beta(w_n(\tau)) - \beta(\tilde{w}_n(\tau))\|_{L^1(\mathbb{R})} \leqq (B_0 + B_1\tau)\, e^{-\tau}$ *for all* $\tau \in [0, T']$, *where* B_0 *and* B_1 *are constants independent of* n. (ii) $\|\beta(\tilde{w}_n(\tau)) - \beta(w_n^*)\|_{L^1(\mathbb{R})} \leqq (B_2 + B_3\tau)\, e^{-\tau}$ *for all* $\tau \in [0, T']$, *where* B_2 *and* B_3 *are constants independent of* n.

Proof. (i) The function $\tilde{w}_n$ satisfies

$$(\beta(\tilde{w}_n))_\tau + (1 - \eta\beta'(\tilde{w}_n))\tilde{w}_{n\eta} = e^{-\tau}\tilde{w}_{n\eta\eta} \quad \text{on } \mathbb{R}|\{\tilde{\eta}\} \times (0, T'].$$

The function w_n satisfies

$$(\beta(w_n))_\tau + (1 - \eta\beta'(w_n))w_{n\eta} = e^{-\tau}w_{n\eta\eta} \quad \text{on } \mathbb{R} \times (0, T'].$$

Subtraction and integration with respect to η from $-\infty$ to ∞ yields

$$\begin{aligned} &\frac{d}{d\tau}\|\beta(\tilde{w}_n) - \beta(w_n)\|_{L^1(\mathbb{R})} + (\tilde{w}_n - w_n)|_{\eta=-\infty}^{\infty} \\ &\qquad - \eta\{\beta(\tilde{w}_n) - \beta(w_n)\}|_{\eta=-\infty}^{\infty} + \|\beta(\tilde{w}_n) - \beta(w_n)\|_{L^1(\mathbb{R})} \\ &\quad = e^{-\tau}\{(\tilde{w}_n - w_n)_\eta|_{\eta=-\infty}^{\tilde{\eta}-} + (\tilde{w}_n - w_n)_\eta|_{\eta=\tilde{\eta}+}^{\infty}\}. \end{aligned} \tag{4.12}$$

Because (4.8) and (4.10) hold and because k, $\tilde{k} > 1$ in (4.6), (4.7) respectively, the third term on the left-hand side of (4.12) vanishes. Then, using Lemma 1 and Lemma 2, we find

$$\begin{aligned} &\frac{d}{d\tau}\|\beta(\tilde{w}_n) - \beta(w_n)\|_{L^1(\mathbb{R})} + \|\beta(\tilde{w}_n) - \beta(w_n)\|_{L^1(\mathbb{R})} \\ &\quad \leqq e^{-\tau}\{\tilde{w}_{n\eta}(\tilde{\eta}^-, \tau) - \tilde{w}_{n\eta}(-\infty, \tau) + C_0\} \end{aligned} \tag{4.13}$$

for all $\tau \in (0, T']$, where the constant C_0 does not depend on n. For each $n \in \mathbb{N}$ the solution v_n of Problem $\tilde{P}_n$ satisfies $v_{n\eta}(\tilde{\eta}, \tau) \geqq 0$ for $\tau \in [0, T']$ and $v_{n\eta}(\eta, 0) \geqq 0$ for $\eta \in (-\infty, \tilde{\eta}]$, except possibly at the points where $u_{0n} = w_n^*$. It then follows from a maximum principle argument that $v_{n\eta} = \tilde{w}_{n\eta} \geqq 0$ in $H_{T'}^{\tilde{\eta}}$. Moreover, using (4.10), we see that

$$\tilde{w}_{n\eta}(\tilde{\eta}^-, \tau) \leqq \frac{dw_n^*(\tilde{\eta}^-)}{d\eta}$$

for all $\tau \in (0, T']$ and for all $n \in \mathbb{N}$.

Therefore the right-hand side of (4.13) can be estimated uniformly with respect to n, which implies that there exists a constant C such that

$$\frac{d}{d\tau}\|\beta(\tilde{w}_n)-\beta(w_n)\|_{L^1(\mathbb{R})}+\|\beta(\tilde{w}_n)-\beta(w_n)\|_{L^1(\mathbb{R})}\leqq Ce^{-\tau} \tag{4.14}$$

for all $\tau\in(0, T']$ and for all $n\in\mathbb{N}$.

From this, inequality (i) follows immediately.

(ii) The proof of this inequality is almost identical and will be omitted. □

Next use (4.11) and Lemma 3 to find

$$\|\beta(w_n(\tau))-\beta(w_n^*)\|_{L^1(\mathbb{R})}\leqq(C_1+C_2\tau)\,e^{-\tau} \tag{4.15}$$

for all $\tau\in[0, T']$, where C_1 and C_2 are constants independent of n.

Part 2. The convergence of $\{w_n\}$ to the weak solution w of Problem P as $n\to\infty$.

From Lemma 1 it follows that

$$|w_n(\eta_1, \tau)-w_n(\eta_2, \tau)|\leqq M_0|\eta_1-\eta_2| \tag{4.16}$$

for all (η_1, τ), $(\eta_2, \tau)\in\bar{S}_{T'}$ and for all $n\in\mathbb{N}$.

Next consider the sequence of rectangles $D_m:=(-m, m)\times(0, T')$ with $m\in\mathbb{N}$. Then by a result of Gilding [6] about the Hölder continuity of solutions of parabolic equations

$$|w_n(\eta, \tau_1)-w_n(\eta, \tau_2)|\leqq C(m)|\tau_1-\tau_2|^{1/2} \quad \text{in } D_m \tag{4.17}$$

for all $n\in\mathbb{N}$ and for all (η, τ_1) and $(\eta, \tau_2)\in D_m$ with $|\tau_1-\tau_2|\leqq 1$. Thus, for each $m\geqq 1$, the set of functions $\{w_n(\eta, \tau)\}$ is bounded ($0\leqq w_n(\eta, \tau)\leqq 1$ in $\bar{S}_{T'}$) and equicontinuous on D_m. Hence the Arzela–Ascoli Theorem implies the existence of a subsequence $\{w_{n_k}\}$ and a function $w_m\in C(\bar{D}_m)$ such that $w_{n_k}\to w_m$ as $n_k\to\infty$, uniformly on $\bar{D}_m$. Then, by a diagonal process, it follows that there exists a function w, defined on $\bar{S}_{T'}$, and a convergent subsequence $\{w_{n_j}(\eta, \tau)\}$ such that $w_{n_j}\to w$ as $j\to\infty$, pointwise on $\bar{S}_{T'}$. Since this convergence is uniform on every bounded subset of $\bar{S}_{T'}$, the limit function w is continuous on $\bar{S}_{T'}$.

To show that the function $w(\eta, \tau)$ is a weak solution of Problem P we must check properties (i), (ii) and (iii) in the definition of weak solution. Because the proof is standard (see e.g. [4]), we shall omit it here.

Finally we derive the integral estimate (4.2).

Because $|\beta(w_n)-\beta(w_n^*)|\leqq(\beta(\tilde{w}_n)-\beta(w_n))+(\beta(\tilde{w}_n)-\beta(w_n^*))$ and (4.8), (4.10) hold, we have

$$|\beta(w_n)-\beta(w_n^*)|\leqq\begin{cases}\beta(1)-\beta(\underline{s}_n), & \eta\geqq\tilde{\eta},\\ 2\left\{\beta(\bar{s}_n)-\beta\left(\dfrac{1}{n}\right)\right\}, & \eta<\tilde{\eta}.\end{cases} \tag{4.18}$$

Using the assumptions H1(i), H4(i) on β and the definition of $\underline{s}_n$, $\bar{s}_n$ (given in (4.6) and (4.7)) it is clear that the right-hand side of (4.18) can be estimated, uniformly in n, by an L^1-function which is $O(\eta^{-k})$ for $\eta\to\infty$ and $O((-\eta)^{-\bar{k}\alpha})$ for $\eta\to-\infty$. Then if we take the limit $n\to\infty$ in the estimates in Lemma 3 and use the Dominated Convergence Theorem, we find

$$\|\beta(w(\tau))-\beta(w^*)\|_{L^1(\mathbb{R})}\leqq(C_1+C_2\tau)\,e^{-\tau}$$

for all $\tau\in[0, T']$, where C_1 and C_2 are positive constants independent of n.

This proves Proposition 1. □

To derive a pointwise estimate from the integral estimate (4.2) we need the following lemma.

LEMMA 4. *Let ϕ be a function defined on $\mathbb{R}$, satisfying $\phi(x) \geqq 0$ on $\mathbb{R}$ and $|\phi(x) - \phi(x_0)| \leqq A|x - x_0|^\alpha$ on $\mathbb{R}$, where A and α are positive constants. Furthermore, suppose*

$$\int_{-\infty}^{\infty} \phi(x)\, dx \leqq l.$$

Then

$$\sup_{x \in \mathbb{R}} \phi(x) \leqq l^{\alpha/(\alpha+1)} A^{1/(\alpha+1)} \left(\frac{\alpha+1}{\alpha}\right)^{\alpha/(\alpha+1)}.$$

Proof. The proof is similar to the proof of Lemma 3 in Gilding and Peletier [8]. □

PROPOSITION 2. *Let β and u_0 satisfy* H1, H4, *respectively,* H2, H5. *Then if $w(\eta, \tau)$ is the weak solution of Problem P and $w^*(\eta)$ is the weak solution of Problem P_∞, the following estimate holds*:

$$\sup_{\eta \in \mathbb{R}} |\beta(w(\eta, \tau)) - \beta(w^*(\eta))| \leqq [e^{-\tau}(A_1 + A_2 \tau)]^{\alpha/(\alpha+1)}$$

for all $\tau \in [0, T']$, where A_1 and A_2 are positive constants.

Proof. Because w is uniformly Lipschitz continuous with respect to η, which is a direct consequence of (4.16), and β satisfies H4(i), we have

$$|\beta(w(\eta, \tau)) - \beta(w(\eta_0, \tau))| \leqq C|\eta - \eta_0|^\alpha$$

for η, $\eta_0 \in \mathbb{R}$ and for all $\tau \in [0, T']$, where C is a positive constant. The same is true for the function w^*

$$|\beta(w^*(\eta)) - \beta(w^*(\eta_0))| \leqq C^*|\eta - \eta_0|^\alpha$$

for η, $\eta_0 \in \mathbb{R}$ and for some positive constant C^*.

Now apply Lemma 4 with $\phi(\eta) = |\beta(w(\eta, \tau)) - \beta(w^*(\eta))|$ and Proposition 2 is proved. □

Finally, let $u^*(x, t) = w^*(\eta)$, where w^* is the weak solution of Problem P_∞.

Since the result of Proposition 2 does not depend on T', Theorem 2 follows immediately.

5. Convergence to the function $u^*(x, t)$ for the Cauchy–Dirichlet problem.

In this section we give results on the asymptotic behaviour of the solution of the Cauchy-Dirichlet problem II when u_0 satisfies the boundary condition $\mathrm{BC1}^+$ and $\beta'' < 0$.

Furthermore we assume that β satisfies H1, H4 and u_0 satisfies $\mathrm{H2}^+$, $\mathrm{H5}^+$.

As in § 4 we use transformation (4.1) and consider the transformed problem:

$$P^+ \begin{cases} (\beta(w))_\tau + (1 - \eta\beta'(w))w_\eta = e^{-\tau} w_{\eta\eta} & \text{in } H_{T'}, \\ w(0, \tau) = 0, \quad 0 < \tau \leqq T', \\ w(\cdot, 0) = u_0(\cdot) & \text{on } \mathbb{R}^+, \end{cases}$$

where $H_{T'} = \{(\eta, \tau) : \eta \in \mathbb{R}^+, 0 < \tau \leqq T'\}$, $T' = \log(T + 1)$ and $w(\eta, \tau) = u(x, t)$.

The reduced problem

$$P_\infty^+ \begin{cases} (1 - \eta\beta'(w)) \dfrac{dw}{d\eta} = 0 \quad \text{on } \mathbb{R}^+, \\ w(0) = 0, \quad w(+\infty) = 1 \end{cases}$$

has a unique weak solution w^* (as in § 3).

The following proposition holds.

PROPOSITION 3. *Let w be the weak solution of Problem P^+ and let w^* be the weak solution of Problem P^+_∞. Suppose β satisfies* H1, H4 *and u_0 satisfies* H2$^+$, H5$^+$. *Then*

$$\sup_{\eta\in\mathbb{R}^+} |\beta(w(\eta,\tau))-\beta(w^*(\eta))| \leqq [e^{-\tau}(A1^+ + A2^+\tau)]^{\alpha/(\alpha+1)}$$

for all $\tau\in[0, T']$, where $A1^+$ and $A2^+$ are positive constants.

Proof. The proof is essentially the same as the proof of Proposition 2. The main difference is the following. When introducing the approximate Problem P_n in $H_{T'}$, with solution w_n, one needs a uniform bound with respect to n on $w_{n\eta}$ at $\eta=0$. This bound is obtained by observing that

$$\frac{1}{n} \leqq w_n(\eta,\tau) \leqq \min\left\{\frac{1}{n} + M\eta, 1\right\}$$

for all $(\eta, \tau)\in H_{T'}$ and for all $n\in\mathbb{N}$. Here the constant M is chosen such that $M=\max\{M^*, (1/\tilde{\eta})\}$, where $\tilde{\eta}=1/(\beta'(1))$ as defined in § 3. The rest of the proof will be omitted. □

Again define $u^*(x,t)=w^*(\eta)$; then an immediate consequence of Proposition 3 is Theorem 3.

Acknowledgment. The authors are greatly indebted to Professor L. A. Peletier for his assistance with this work.

REFERENCES

[1] J. BEAR, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.

[2] G. H. BOLT, ed., *Soil Chemistry, B. Physico-Chemical Models*, Developments in Soil Science 5 B, Elsevier, Amsterdam, New York, 1982.

[3] C. COSNER, *Asymptotic behaviour of solutions of second order parabolic partial differential equations with unbounded coefficients*, J. Differential Equations, 35 (1980), pp. 407-428.

[4] C. J. VAN DUYN AND Q. X. YE, *The flow of two immiscible fluids through a porous medium*, Nonlinear Anal. Theory Meth. Appl., 8 (1984), pp. 353-367.

[5] R. A. FREEZE AND J. A. CHERRY, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[6] B. H. GILDING, *Hölder continuity of solutions of parabolic equations*, J. London Math. Soc. (2), 13 (1976), pp. 103-106.

[7] ———, *A nonlinear degenerate parabolic equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (1977), pp. 393-432.

[8] B. H. GILDING AND L. A. PELETIER, *The Cauchy problem for an equation in the theory of infiltration*, Arch. Rational Mech. Anal., 61 (1976), pp. 127-140.

[9] N. V. KHUSNYTDINOVA, *The limiting moisture profile during infiltration into a homogeneous soil*, Prikl. Mat. Mekh., 31 (1967), pp. 770-776; J. Appl. Math. Mech., 31 (1967), pp. 783-789.

[10] S. OSHER AND J. RALSTON, *L^1-stability of travelling waves with applications to convective porous media flow*, Comm. Pure Appl. Math., 35 (1982), pp. 737-749.

[11] L. A. PELETIER, *Asymptotic behaviour of solutions of the porous media equation*, SIAM J. Appl. Math., 21 (1971), pp. 542-551.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

013

GLOBAL EXISTENCE AND BOUNDEDNESS IN REACTION-DIFFUSION SYSTEMS*

SELWYN L. HOLLIS†, ROBERT H. MARTIN, JR.† AND MICHEL PIERRE‡

Abstract. In many applications, systems of reaction-diffusion equations arise in which the nature of the nonlinearity in the reaction terms renders ineffective the standard techniques (such as invariant sets and differential inequalities) for establishing global existence, boundedness, and asymptotic behavior of solutions. In this paper we prove global existence and uniform boundedness for a class of reaction-diffusion systems involving two unknowns in which an a priori bound is available for one component as long as solutions exist. Among this class of systems is the so-called Brusselator, a model from the study of instabilities in chemical processes.

Key words. reaction-diffusion systems, global existence, stability

AMS (MOS) subject classifications. 35K45, 35B35

Introduction. It is well known that the addition of diffusion in a reaction system can cause the loss of stability properties of equilibrium solutions. This follows from the fact that the subtraction of a positive diagonal matrix from a square matrix can increase the maximum real part of the eigenvalues. In general, it does not seem to be known whether adding diffusion in a reaction system (or changing the diffusion coefficients in a reaction-diffusion system) can affect the global existence of solutions. In some situations general results on invariant sets and differential inequalities may be applied to establish global existence as well as estimate the growth of solutions. However, the class of equations under consideration here cannot be effectively analyzed by these general methods.

A typical type of system that can be studied with the methods of this paper is a reaction-diffusion equation of the form

$$u_t = d_1\Delta u - uv^\beta, \qquad v_t = d_2\Delta v + uv^\beta \tag{1}$$

with various homogeneous boundary conditions and nonnegative initial data, where Ω is a bounded region in $\mathbb{R}^n$, Δ is the Laplacian, $\beta \geqq 1$, and d_1, $d_2 > 0$. N. Alikakos [1] established global existence and $\mathscr{L}^\infty$-bounds of solutions when $1 \leqq \beta < (n+2)/n$. K. Masuda [8] shows that solutions to this system exist globally for every $\beta \geqq 1$ and, in addition, shows that the solutions converge as $t \to \infty$. One should note that the global existence assertion is immediate if $\beta = 1$ and also if $d_1 = d_2$ (for if $d_1 = d_2$ and $w \equiv u + v$, then $w_t = d_1\Delta w$, and it easily follows that u and v remain uniformly bounded since $u, v \geqq 0$). On the other hand, if $d_1 \neq d_2$, the global boundedness is not obvious; one can even show that there is no global $\mathscr{L}^\infty$ estimate in terms of $\|u_0\|_\infty$ and $\|v_0\|_\infty$ for t large if $d_2 = 0$ (see § 5).

A second illustration is the Brusselator, a model of a certain chemical morphogenetic process due to Turing [12], which has the form

$$\begin{aligned} &u_t = d_1\Delta u - uv^2 + Bv, && \\ &v_t = d_2\Delta v + uv^2 - (B+1)v + A, && x \in \Omega, \quad t > 0, \\ &u(x,t) = B/A,\ v(x,t) = A, && t > 0, \quad x \in \partial\Omega \end{aligned} \tag{2}$$

* Received by the editors May 6, 1985; accepted for publication (in revised form) April 8, 1986.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

‡ UER Sciences Mathematiques, Université de Nancy I, Nancy, France.

with nonnegative initial data where A, B, d_1 and d_2 are positive constants. This system is discussed for $\Omega=(0,1)$ in Prigogine and Nicolis [10]. Except for inhomogenous boundary conditions, (2) is very similar to (1) with $\beta=2$. In [6], H. Lange proved that solutions to (2) with $\Omega=(0,1)$ exist globally, *provided* that the diffusion coefficients satisfy

$$(d_1-d_2)^2/(4d_1d_2)<[1+B^2/(4A^2)]^{-1}.$$

Observe that this result may be regarded as a perturbation of the routinely deduced fact that nonnegative solutions to (2) exist globally when $d_1=d_2$ (for if $d_1=d_2$ and $w\equiv u+v$, then $w_t=d_1w_{xx}+A-v\leqq d_1w_{xx}+A$). Auchmuty and Nicolis [2] proved the global existence of solutions to (2) for spatial dimension 1 and arbitrary d_1, $d_2>0$. A proof of global existence and boundedness for solutions of (2) with Neumann boundary conditions; d_1, $d_2>0$; and spatial dimension $n\leqq 3$ may be found in Rothe [11]. Our results show that nonnegative solutions to (2) exist globally and remain uniformly bounded for the stated boundary conditions, arbitrary spatial dimensions, and d_1, $d_2>0$.

The present paper is organized as follows: 1, The main results; 2, Semigroup formulation and local existence; 3, Preliminary estimates; 4, Proof of the theorems; 5, An example and a counterexample.

1. The main results. Throughout this paper it is assumed that Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$. Also, Δ is the Laplacian operator on Ω and $\partial/\partial n$ denotes the outward normal derivative on $\partial\Omega$. We consider the reaction-diffusion system

$$(1.1)\qquad \begin{array}{lll} \text{(a)} & \begin{array}{l} u_t=d_1\Delta u+f(t,u,v),\\[4pt] v_t=d_2\Delta v+g(t,u,v),\end{array} & x\in\Omega,\quad t>0,\\[12pt] \text{(b)} & \begin{array}{l} \alpha_1u+(1-\alpha_1)\dfrac{\partial u}{\partial n}=\beta_1,\\[8pt] \alpha_2v+(1-\alpha_2)\dfrac{\partial v}{\partial n}=\beta_2,\end{array} & x\in\partial\Omega,\quad t>0,\\[12pt] \text{(c)} & u=u_0,\ v=v_0, & x\in\Omega,\quad t=0\end{array}$$

where the following basic hypotheses are assumed to hold:

(H1) d_1, d_2, α_1, α_2, β_1 and β_2 are constants with $d_1, d_2>0$; β_1, $\beta_2\geqq 0$; and either $0<\alpha_1$, $\alpha_2<1$, $\alpha_1=\alpha_2=1$, or $\alpha_1=\alpha_2=0$. Also, $\beta_1=\beta_2=0$ if $\alpha_1=\alpha_2=0$;

(H2) f and g are continuously differentiable functions from $[0,\infty)^3$ into $\mathbb{R}$ with $f(t,0,\eta)\geqq 0$ and $g(t,\xi,0)\geqq 0$ for all t, ξ, $\eta\geqq 0$;

(H3) Both u_0 and v_0 are measurable on Ω and there is an $M_0>0$ such that $0\leqq u_0(x)$, $v_0(x)\leqq M_0$ for all $x\in\Omega$.

PROPOSITION 1. *Suppose that* (H1)-(H3) *are satisfied. Then* (1.1) *has a unique, noncontinuable* (*classical*) *solution* (u,v) *on* $\Omega\times[0,T^*)$, *and there are continuous functions* N_1, $N_2:[0,T^*)\to[0,\infty)$ *such that*

$$(1.2)\qquad 0\leqq u(x,t)\leqq N_1(t),\qquad 0\leqq v(x,t)\leqq N_2(t)\quad \text{for } (x,t)\in\Omega\times[0,T^*).$$

Moreover, if $T^*<\infty$ *then*

$$(1.3)\qquad \limsup_{t\uparrow T^*}\ \sup_{x\in\Omega}\{|u(x,t)|+|v(x,t)|\}=\infty.$$

This local existence result follows from the basic existence theory for abstract semilinear differential equations (see, e.g., Henry [4] or Pazy [9]). Since some modifications of these results are needed, the ideas of the proof are included in § 2. In general, (H1)-(H3) are not sufficient to insure that solutions to (1.1) exist for all $t \geqq 0$. The purpose of this paper is to give sufficient conditions guaranteeing that $T^* = \infty$ in Proposition 1. In addition, we establish a result on the uniform boundedness of solutions on $\Omega \times [0, \infty)$.

THEOREM 1. *In addition to* (H1)-(H3) *suppose that*

(H4) *The function* N_1 *in* (1.2) *is bounded if* $T^* < \infty$;

(H5) *There is a* $\gamma \geqq 1$ *and a continuous* $L_0: [0, \infty)^2 \to [0, \infty)$ *such that* $|g(t, \xi, \eta)| \leqq L_0(t, r)(1+\eta)^\gamma$ *for all* $t, \xi, \eta \geqq 0$ *with* $\xi \leqq r$;

(H6) *There is a continuous* $\mu_0: [0, \infty)^2 \to [0, \infty)$ *such that* $f(t, \xi, \eta) + g(t, \xi, \eta) \leqq \mu_0(t, r)$ *for all* $t, \xi, \eta \geqq 0$ *with* $\xi \leqq r$.

Then the solution (u, v) *exists on* $\Omega \times [0, \infty)$ *and* (1.2) *holds with* $T^* = \infty$.

Remark 1. Condition (H4) says that in each bounded t interval, $u(\cdot, t)$ remains uniformly bounded so long as the solution to (1.1) exists. This is the case, for example, if $f(t, \xi, \eta) \leqq \mu_1(t)\xi + \mu_2(t)$ for all $t, \xi, \eta \geqq 0$, where $\mu_1, \mu_2: [0, \infty) \to [0, \infty)$ are continuous.

THEOREM 2. *In addition to the suppositions in Theorem* 1, *suppose that there is an* $\bar{N}_1 > 0$ *and for each* $r > 0$ *there are numbers* $L_\infty(r)$ *and* $\mu_\infty(r)$ *such that*

$$N_1(t) \leqq \bar{N}_1, \quad L_0(t, r) \leqq L_\infty(r), \quad \mu_0(t, r) \leqq \mu_\infty(r) \quad \textit{for all } t \geqq 0.$$

If $\alpha_2 = 0$, *assume also that* $\mu_\infty(r) \equiv 0$. *Then, there is an* $\bar{N}_2$ *such that* $N_2(t) \leqq \bar{N}_2$ *for all* $t \geqq 0$, *and so the solution to* (1.1) *is uniformly bounded on* $\Omega \times [0, \infty)$.

2. Semigroup formulation and local existence. For each $p \in (1, \infty)$ and $j \in \{1, 2\}$ define $A_{j,p}$ on $\mathscr{L}^p(\Omega)$ by

$$\begin{aligned} &A_{j,p}w = d_j \Delta w \quad \text{for } w \in \mathscr{D}(A_{j,p}), \quad \text{and} \\ &\mathscr{D}(A_{j,p}) = \left\{ w \in W^{2,p}(\Omega) \colon \alpha_j w + (1-\alpha_j)\frac{\partial w}{\partial n} = 0 \text{ on } \partial\Omega \right\} \end{aligned} \tag{2.1}$$

where $W^{2,p}(\Omega)$ is the usual Sobolev space and d_j, α_j are as in (H1). It is well known that $A_{j,p}$ generates a compact, analytic semigroup $\mathscr{T}_{j,p} = \{\mathscr{T}_{j,p}(t) \colon t \geqq 0\}$ of bounded linear operators on $\mathscr{L}^p(\Omega)$, and that

$$|\mathscr{T}_{j,p}(t)w|_p \leqq e^{\omega t}|w|_p \quad \text{for } t \geqq 0, \quad w \in \mathscr{L}^p(\Omega) \tag{2.2}$$

where $\omega < 0$ if $\alpha_j > 0$, $\omega = 0$ if $\alpha_j = 0$, and

$$|w|_p \equiv \left[\int_\Omega |w(x)|^p \, dx\right]^{1/p}.$$

For each $\alpha > 0$ and $\lambda > \omega$ the fractional powers $(\lambda I - A_{j,p})^{-\alpha}$ exist and are injective, bounded linear operators on $\mathscr{L}^p(\Omega)$ (see Pazy [9]). For each $\alpha > 0$ define $B_{j,p}^{-\alpha} = (-A_{j,p})^{-\alpha}$ if $\alpha_j > 0$, $B_{j,p}^{-\alpha} = (I - A_{j,p})^{-\alpha}$ if $\alpha_j = 0$, and $B_{j,p}^{\alpha} = (B_{j,p}^{-\alpha})^{-1}$. Recall that $\mathscr{D}(B_{j,p}^{\alpha})$ is a Banach space with the graph norm $|w|_\alpha = |B_{j,p}^{\alpha} w|_p$, and that, for $\alpha > \beta \geqq 0$, $\mathscr{D}(B_{j,p}^{\alpha})$ is a dense subspace of $\mathscr{D}(B_{j,p}^{\beta})$ with the inclusion $\mathscr{D}(B_{j,p}^{\alpha}) \subset \mathscr{D}(B_{j,p}^{\beta})$ compact (we use the convention $\mathscr{D}(B_{j,p}^{0}) = \mathscr{L}^p(\Omega)$). Also (see Henry [4, p. 40]),

$$\begin{aligned} &\text{if } \alpha > n/(2p) \text{ then } \mathscr{D}(B_{j,p}^{\alpha}) \subset \mathscr{L}^\infty(\Omega), \text{ and} \\ &|w|_\infty \leqq M_{\alpha,p}|B_{j,p}^{\alpha} w|_p \text{ for all } w \in \mathscr{D}(B_{j,p}^{\alpha}). \end{aligned} \tag{2.3}$$

In addition, the following properties are satisfied by $\mathcal{T}_{j,p}$ and $B_{j,p}^{\alpha}$; a proof can be found in Pazy [9, p. 74].

LEMMA 1. *Suppose that $\mathcal{T}_{j,p}$ and $B_{j,p}^{\alpha}$ are as above. Then*

(i) *$\mathcal{T}_{j,p}(t)$: $\mathcal{L}^p(\Omega)\to\mathcal{D}(B_{j,p}^{\alpha})$ for all $t>0$,*

(ii) *$|B_{j,p}^{\alpha}\mathcal{T}_{j,p}(t)w|_p\leqq C_{\alpha,p}t^{-\alpha}e^{\omega t}|w|_p$ for $t>0$ and $w\in\mathcal{L}^p(\Omega)$, and*

(iii) *$\mathcal{T}_{j,p}(t)B_{j,p}^{\alpha}w=B_{j,p}^{\alpha}\mathcal{T}_{j,p}(t)w$ for $t>0$, $w\in\mathcal{D}(B_{j,p}^{\alpha})$.*

In order to use a modification of a local existence result in Henry [4], we write (1.1) as an integral equation system via variation of constants. So define F and G on $[0,\infty)\times\mathcal{L}^{\infty}(\Omega)^2$ by

$$(2.4)\qquad \begin{aligned}&[F(t,w_1,w_2)](x)=f(t,w_1(x),w_2(x))\\&[G(t,w_1,w_2)](x)=g(t,w_1(x),w_2(x))\end{aligned}\quad \text{for } x\in\Omega,\quad t>0,\quad w_1,w_2\in\mathcal{L}^{\infty}(\Omega),$$

and let z_j satisfy

$$(2.5)\qquad \Delta z_j=0\quad\text{on }\Omega,\qquad \alpha_j z_j+(1-\alpha_j)\frac{\partial z_j}{\partial n}=\beta_j\quad\text{on }\partial\Omega$$

where $z_j=0$ if $\alpha_j=0$ (and hence $\beta_j=0$—see (H1)). By variation of constants (see Pazy [9]) it follows that if (u,v) is a solution in $\mathcal{L}^p(\Omega)\times\mathcal{L}^p(\Omega)$ to the system

$$(2.6)\qquad \begin{aligned}u(t)&=\mathcal{T}_{1,p}(t)(u_0-z_1)+z_1+\int_0^t\mathcal{T}_{1,p}(t-s)F(s,u(s),v(s))\,ds,\\ v(t)&=\mathcal{T}_{2,p}(t)(v_0-z_2)+z_2+\int_0^t\mathcal{T}_{2,p}(t-s)G(s,u(s),v(s))\,ds\end{aligned}$$

then $u(x,t)\equiv[u(t)](x)$, $v(x,t)\equiv[v(t)](x)$ is the solution to (1.1). Of course, z_1 and z_2 reflect the possibility of the inhomogeneous boundary conditions in (1.1b), and clearly $z_j=0$ whenever $\beta_j=0$.

Proof of Proposition 1. Select $\alpha\in(0,1)$ and $p>1$ so that (2.3) holds and use the techniques in Henry [4, § 3.3], modified to take into account the inhomogeneous terms, to show the existence of a unique noncontinuable soltuion (u,v) to (2.6) for

$$(u_0-z_1,v_0-z_2)\in\mathcal{D}(B_{1,p}^{\alpha})\times\mathcal{D}(B_{2,p}^{\alpha}).$$

Also, $u(t)$ and $v(t)$ have nonnegative values on Ω since $f(t,0,\eta)\geqq 0$ and $g(t,\xi,0)\geqq 0$ for all $t,\xi,\eta\geqq 0$ (see, e.g., Lightbourne and Martin [7]). Application of the results of Ball [3, Thm. 3.1] shows that (u,v) is defined on an interval of the form $[0,T^*)$, and that

$$(2.7)\qquad \begin{aligned}&\text{if } T^*<\infty,\text{ then } |B_{1,p}^{\alpha}(u(t)-z_1)|_p+|B_{2,p}^{\alpha}(v(t)-z_2)|_p\to\infty \text{ as } t\uparrow T^*,\text{ and}\\ &\limsup_{t\uparrow T^*}\textstyle\int_0^t(t-s)^{-\alpha}[|F(s,u(s),v(s))|_p+|G(s,u(s),v(s))|_p]\,ds=\infty.\end{aligned}$$

For each $T,R>0$ there is an $M(T,R)>0$ such that

$$|f(t,\xi,\eta)|,|g(t,\xi,\eta)|\leqq M(T,R)\quad\text{for } t\in[0,T],\quad \xi,\eta\in[0,R].$$

Hence,

$$(2.8)\qquad \begin{aligned}&|F(t,w_1,w_2)|_{\infty},|G(t,w_1,w_2)|_{\infty}\leqq M(T,R)\\ &\text{whenever } t\in[0,T]\text{ and } w_1,\ w_2\in\mathcal{L}^{\infty}(\Omega)\text{ with } |w_1|_{\infty},|w_2|_{\infty}\leqq R.\end{aligned}$$

Since the function $\psi(t)\equiv \mathcal{T}_{j,p}(t)(w_j-z_j)+z_j$ satisfies

$$\psi_t=d_j\Delta\psi \quad \text{on } \Omega\times(0,\infty),$$

$$\alpha_j\psi+(1-\alpha_j)\frac{\partial\psi}{\partial n}=\beta_j \quad \text{on } \partial\Omega\times(0,\infty),$$

$$\psi(0)=w_j \quad \text{on } \Omega,$$

application of the maximum principle yields

$$\text{(2.9)} \qquad |\mathcal{T}_{j,p}(t)(w_j-z_j)+z_j|_\infty \leqq R \quad \text{for all } t\geqq 0 \text{ whenever } |z_j|_\infty,\ |w_j|_\infty\leqq R.$$

If $0\leqq t_1<T^*$, $|z_j|_\infty\leqq R_1<R_2$ and $|u(t_1)|_\infty$, $|v(t_1)|_\infty\leqq R_1$, then by (2.8) and (2.9)

$$\begin{aligned}|u(t)|_\infty&=|\mathcal{T}_{1,p}(t-t_1)(u(t_1)-z_1)+z_1+\int_{t_1}^{t}\mathcal{T}_{1,p}(t-s)F(s,u(s),v(s))\,ds|_\infty\\ &\leqq|\mathcal{T}_{1,p}(t-t_1)(u(t_1)-z_1)+z_1|_\infty+\int_{t_1}^{t}|\mathcal{T}_{1,p}(t-s)F(s,u(s),v(s))|\,ds\\ &\leqq R_1+(t-t_1)M(T^*,R_2)\end{aligned}$$

so long as $|u(s)|_\infty$, $|v(s)|_\infty\leqq R_2$ for $s\in[t_1,t]$. A similar estimate holds for $|v(t)|_\infty$, and, taking $R_2=R_1+1$, it follows that

(2.10) if $0\leqq t_1<T^*$ and $|u(t_1)|_\infty$, $|v(t_1)|_\infty\leqq R_1$ where $R_1\geqq\max\{|z_1|_\infty,|z_2|_\infty\}$, then $T^*-t_1>M(T^*,R_1+1)^{-1}$, and $|u(t)|_\infty$, $|v(t)|_\infty\leqq R_1+1$ for $t\in[t_1,t_1+M(T^*,R_1+1)^{-1}]$.

Now suppose that $T^*<\infty$. Then (2.7) implies that

$$\begin{aligned}\infty&=\limsup_{t\uparrow T^*}|F(t,u(t),v(t))|_p+|G(t,u(t),v(t))|_p\\ &\leqq C\limsup_{t\uparrow T^*}|F(t,u(t),v(t))|_\infty+|G(t,u(t),v(t))|_\infty.\end{aligned}$$

Hence by (2.8)

$$\limsup_{t\uparrow T^*}|u(t)|_\infty+|v(t)|_\infty=\infty.$$

This, combined with (2.10), shows that

$$|u(t)|_\infty+|v(t)|_\infty\to\infty \quad \text{as } t\uparrow T^* \text{ if } T^*<\infty.$$

Each of the assertions in Proposition 1 follows whenever

$$(u_0-z_1,v_0-z_2)\in\mathcal{D}(B_{1,p}^{\alpha})\times\mathcal{D}(B_{2,p}^{\alpha}).$$

Suppose now that $(u_0,v_0)\in\mathcal{L}^\infty(\Omega)^2$ and let $\{(u_0^k,v_0^k)\}_1$ be a sequence in $\mathcal{D}(B_{1,p}^{\alpha})\times\mathcal{D}(B_{2,p}^{\alpha})$ such that u_0^k, $v_0^k\geqq 0$ and

$$|u_0^k-u_0|_p,\ |v_0^k-v_0|_p\to 0 \quad \text{as } k\to\infty.$$

By the first equation in (2.6) and the properties of $B_{1,p}$ stated in Lemma 1, it follows that

$$|B_{1,p}^{\beta}(u^k(t)-z_1)|_p\leqq C_\beta t^{-\beta}|u_0^k-z_1|_p+\int_0^t C_\beta(t-s)^{-\beta}|F(s,u^k(s),v^k(s))|_p\,ds$$

for all $t \in [0, T_k^*)$. If $R_1 \geqq |z_1|_\infty, |z_2|_\infty$ is chosen so that $R_1 \geqq |u_0^k|_\infty, |v_0^k|_\infty$ for all $k \geqq 1$, then it follows from (2.10) that if $\delta = M(R_1, R_1+1)^{-1}$ then $T_k^* > \delta$ for all $k \geqq 1$. From these estimates one can deduce the existence of a $\bar{C}_\beta$ such that

$$|B_{1,p}^\beta(u^k(t) - z_1)|_p \leqq \bar{C}_\beta t^{-\beta} \quad \text{for all } t \in [0, \delta], k \geqq 1 \text{ and } \alpha \leqq \beta < 1$$

and, similarly,

$$|B_{2,p}^\beta(v^k(t) - z_2)|_p \leqq \bar{C}_\beta t^{-\beta} \quad \text{for all } t \in [0, \delta], k \geqq 1 \text{ and } \alpha \leqq \beta < 1.$$

Because of the compact embedding of $\mathscr{D}(B_{j,p}^\beta)$ in $\mathscr{D}(B_{j,p}^\alpha)$ for $\alpha < \beta < 1$ and the compactness of $T_{j,p}$, it follows, by selecting a subsequence and relabeling if necessary, that

$$\lim_{k\to\infty} |u^k(t) - u(t)|_p = \lim_{k\to\infty} |v^k(t) - v(t)|_p = 0$$

for $t \in [0, \delta]$ and that (u, v) is a solution to (2.6) on $[0, \delta]$ with $u(t) - z_1 \in \mathscr{D}(B_{1,p}^\alpha)$ and $v(t) - z_2 \in \mathscr{D}(B_{2,p}^\alpha)$ for $0 < t \leqq \delta$ (see, e.g., Lemmas 6 and 7 in Lightbourne and Martin [7]). By replacing $[0, T^*)$ with $[\delta, T^*)$ and (u_0, v_0) with $(u(\delta), v(\delta))$ and using the results already established when $u_0 - z_1 \in \mathscr{D}(B_{1,p}^\alpha)$ and $v_0 - z_2 \in \mathscr{D}(B_{2,p}^\alpha)$, Proposition 1 follows.

3. Preliminary estimates. For $0 \leqq \tau < T < \infty$ let $Q_{\tau,T} = \Omega \times (\tau, T)$ and for $q \in [1, \infty)$ let $\mathscr{L}^q(Q_{\tau,T})$ be the space of measurable $\phi: Q_{\tau,T} \to \mathbb{R}$ with

$$\|\phi\|_{q,\tau,T} \equiv \left[\int_{Q_{\tau,T}} |\phi(x, t)|^q \, dx \, dt\right]^{1/q} < \infty.$$

Observe that if $\phi \in \mathscr{L}^q(Q_{\tau,T})$ then $\phi(\cdot, t) \in \mathscr{L}^q(\Omega)$ for almost all $t \in (\tau, T)$, $t \mapsto \phi(\cdot, t)$ is measurable, and that

$$\|\phi\|_{q,\tau,T} = \left[\int_\tau^T |\phi(\cdot, t)|_q^q \, dt\right]^{1/q}.$$

Throughout this section we assume that $\theta \in \mathscr{L}^q(Q_{\tau,T})$ and that $\phi: Q_{\tau,T} \to \mathbb{R}$ is the solution to

$$\text{(3.1)} \qquad \begin{aligned} &\phi_t = -d_2 \Delta \phi - \theta, && x \in \Omega, \quad t \in (\tau, T), \\ &\alpha_2 \phi + (1 - \alpha_2) \frac{\partial \phi}{\partial n} = 0, && x \in \partial\Omega, \quad t \in (\tau, T), \\ &\phi = 0, && x \in \Omega, \quad t = T \end{aligned}$$

where d_2 and α_2 are as in (H1). If $\bar{\phi}(x, t) \equiv \phi(x, T - t)$ and $\bar{\theta}(x, t) \equiv \theta(x, T - t)$ for $(x, t) \in \Omega \times (0, T - \tau)$, then

$$\text{(3.2)} \qquad \begin{aligned} &\bar{\phi}_t = d_2 \Delta \bar{\phi} + \bar{\theta}, && x \in \Omega, \quad t \in (0, T - \tau), \\ &\alpha_2 \bar{\phi} + (1 - \alpha_2) \frac{\partial \bar{\phi}}{\partial n} = 0, && x \in \partial\Omega, \quad t \in (0, T - \tau), \\ &\bar{\phi} = 0, && x \in \Omega, \quad t = 0. \end{aligned}$$

The maximum principle implies that $\bar{\phi} \geqq 0$ (and hence $\phi \geqq 0$) whenever $\theta \geqq 0$. Thus it follows immediately that $\partial\phi/\partial n \leqq 0$ on $\partial\Omega \times (\tau, T)$ whenever $\theta \geqq 0$.

LEMMA 2. *Suppose that $q \in (1, \infty)$, $\theta \in \mathscr{L}^q(Q_{\tau,T})$ and ϕ is the solution to* (3.1). *Then there is a constant $C(q, T-\tau)$ (independent of θ) such that*

$$\|\phi\|_{q,\tau,T}, \|\Delta\phi\|_{q,\tau,T} \leqq C(q, T-\tau)\|\theta\|_{q,\tau,T}. \tag{3.3}$$

Also, $C(q, T-\tau)$ can be chosen so that

$$|\phi(\cdot, \tau)|_q \leqq C(q, T-\tau)\|\theta\|_{q,\tau,T}. \tag{3.4}$$

Proof. The estimates in (3.3) follow directly from Theorem 9.1 of Ladyzenskaja et al. [5, p. 341]. Applying variation of constants to (3.2) and using (2.2) and Hölder's inequality we have, taking p conjugate to q,

$$\begin{aligned} |\phi(\cdot, \tau)|_q = |\bar{\phi}(\cdot, T-\tau)|_q &= \left| \int_0^{T-\tau} \mathscr{T}_{2,q}(T-\tau-s)\bar{\theta}(\cdot, s)\, ds \right|_q \\ &\leqq \int_0^{T-\tau} |\bar{\theta}(\cdot, s)|_q\, ds \\ &\leqq (T-\tau)^{1/p}\|\theta\|_{q,\tau,T} \end{aligned}$$

and (3.4) follows by taking $C(q, T-\tau) \geqq (T-\tau)^{1/p}$.

In order to obtain uniform boundedness of solutions to (1.1), we need to control the dependence on $T-\tau$ of the constants in Lemma 2.

LEMMA 3. *Under the assumptions of Lemma 2, there is a $\bar{C}(q)$ such that*

$$|P\phi(\cdot,\tau)|_q, \|P\phi\|_{q,\tau,T}, \|\Delta\phi\|_{q,\tau,T} \leqq \bar{C}(q)\|\theta\|_{q,\tau,T}$$

where we set

$$P\phi(\cdot, t) = \begin{cases} \phi(\cdot, t) & \text{if } \alpha_2 \neq 0, \\ \phi(\cdot, t) - |\Omega|^{-1} \int_\Omega \phi(x, t)\, dx & \text{if } \alpha_2 = 0 \end{cases}$$

for $t \in (\tau, T)$. Thus, if $\alpha_2 \neq 0$, one can assume that $C(q, T-\tau) \leqq \bar{C}(q)$ in Lemma 2.

Proof. Note first that C can be chosen so that

$$C(q, T-\tau) \leqq C(q, T'-\tau) \quad \text{if } T' \geqq T > \tau \geqq 0. \tag{3.5}$$

Suppose that $\theta_1 \in \mathscr{L}^q(Q_{\tau,T})$ and define $\theta_2 \in \mathscr{L}^q(Q_{\tau,T})$ by $\theta_2 = \theta_1$ on $Q_{\tau,T}$ and $\theta_2 = 0$ on $Q_{T,T'}$. Then $\|\theta_1\|_{q,\tau,T} = \|\theta_2\|_{q,\tau,T'}$ and if ϕ_1 and ϕ_2 are solutions to (3.1) with $\theta = \theta_1$ and $\theta = \theta_2$, respectively, then $\phi_2 = \phi_1$ on $Q_{\tau,T}$. Thus, by (3.3),

$$\|\phi_1\|_{q,\tau,T} \leqq \|\phi_2\|_{q,\tau,T'} \leqq C(q, T'-\tau)\|\theta_2\|_{q,\tau,T'} = C(q, T'-\tau)\|\theta_1\|_{q,\tau,T}.$$

Since the same estimate holds for $\|\Delta\phi_1\|_{q,\tau,T}$ and $|\phi_1(\cdot, \tau)|_q$, we see that (3.5) is true. Let us first assume that $\alpha_2 \neq 0$. Then (2.2) holds with $\omega = -\delta < 0$. By variation of constants

$$\begin{aligned} |\phi(\cdot, \tau)|_q = |\bar{\phi}(\cdot, T-\tau)|_q &= \left| \int_0^{T-\tau} \mathscr{T}_{2,q}(T-\tau-s)\bar{\theta}(\cdot, s)\, ds \right| \\ &\leqq \int_0^{T-\tau} e^{-\delta(T-\tau-s)}|\bar{\theta}(\cdot, s)|_q\, ds \\ &\leqq \left[\int_0^{T-\tau} e^{-p\delta(T-\tau-s)}\, ds \right]^{1/p} \left[\int_0^{T-\tau} |\bar{\theta}(\cdot, s)|_q^q\, ds \right]^{1/q} \\ &\leqq (p\delta)^{-1/p}\|\theta\|_{q,\tau,T} \end{aligned}$$

where $p^{-1}+q^{-1}=1$. Therefore,

$$|\phi(\cdot,\tau)|_q \leqq k_q\|\theta\|_{q,\tau,T}. \tag{3.6}$$

This shows that the constant in (3.4) may be chosen independent of $T-\tau$. Again by variation of constants

$$\begin{aligned}\|\phi\|_{q,\tau,T}^q &= \int_0^{T-\tau} |\bar\phi(\cdot,t)|_q^q\,dt\\ &= \int_0^{T-\tau}\left|\int_0^t \mathcal{T}_{2,q}(t-s)\bar\theta(\cdot,s)\,ds\right|_q^q dt\\ &\leqq \int_0^{T-\tau}\left\{\int_0^t e^{-\delta(t-s)}|\bar\theta(\cdot,s)|_q\,ds\right\}^q dt.\end{aligned}$$

Let us set $y(s)=|\bar\theta(\cdot,s)|_q$, By Hölder's inequality

$$\int_0^t e^{-\delta(t-s)}y(s)\,ds \leqq \left[\int_0^t e^{-\delta(t-s)}y(s)^q\,ds\right]^{1/q}\left[\int_0^t e^{-\delta(t-s)}\,ds\right]^{1/p},$$

and the last integral is bounded above by δ^{-1}. Hence

$$\begin{aligned}\|\phi\|_{q,\tau,T}^q &\leqq \delta^{-q/p}\int_0^{T-\tau}\int_0^t e^{-\delta(t-s)}y(s)^q\,ds\,dt\\ &= \delta^{-q/p}\int_0^{T-\tau} y(s)^q\int_s^{T-\tau} e^{-\delta(t-s)}\,dt\,ds\end{aligned}$$

where the last integral is bounded above by δ^{-1}. This implies that

$$\|\phi\|_{q,\tau,T}^q \leqq \delta^{-q}\|\theta\|_{q,\tau,T}^q \tag{3.7}$$

and shows that the constant in the estimate for $\|\phi\|_{q,\tau,T}$ in (3.3) may be chosen independent of $T-\tau$.

To show the same for $\|\Delta\phi\|_{q,\tau,T}$, we argue as follows. Let A be an even regular function on $\mathbb{R}$ such that

$$0\leqq A\leqq 1, \qquad A\equiv 1 \quad \text{on } (0,1), \qquad A\equiv 0 \quad \text{on } [2,\infty).$$

For $t_0\geqq 2$, set $\psi(t,x)=A(t-t_0)\bar\phi(t,x)$ where $\bar\phi$ is defined for $t\geqq 0$ by (3.2). Then

$$\begin{aligned}&\psi_t = d_2\Delta\psi + A(t-t_0)\bar\theta + A'(t-t_0)\bar\phi, && x\in\Omega, \quad t>t_0-2,\\ &\alpha_2\psi+(1-\alpha_2)\frac{\partial\psi}{\partial n}=0, && x\in\partial\Omega, \quad t>t_0-2,\\ &\psi(t_0-2)=0.\end{aligned}$$

Set $C(q)=C(q,4)$. By Lemma 2,

$$\begin{aligned}\|\Delta\psi\|_{q,t_0-2,t_0+2}^q &\leqq C(q)^q(\|\bar\theta\|_{q,t_0-2,t_0+2}+\|A'\|_\infty\|\bar\phi\|_{q,t_0-2,t_0+2})^q\\ &\leqq 2^qC(q)^q(\|\bar\theta\|_{q,t_0-2,t_0+2}^q+\|A'\|_\infty^q\|\bar\phi\|_{q,t_0-2,t_0+2}^q).\end{aligned}$$

By the choice of A,

$$\|\Delta\psi\|_{q,t_0-2,t_0+2}^q \geqq \|\Delta\psi\|_{q,t_0-1,t_0+1}^q \geqq \|\Delta\bar\phi\|_{q,t_0-1,t_0+1}^q.$$

Applying these inequalities with $t_0=2,4,\cdots,2k,\cdots$ and summing over k, we obtain

$$\|\Delta\bar\phi\|_{q,1,\infty}^q \leqq 2^qC(q)^q(\|\bar\theta\|_{q,0,\infty}^q+\|A'\|_\infty^q\|\bar\phi\|_{q,0,\infty}^q).$$

Using again Lemma 2 on (0, 1), we deduce that

$$\|\Delta\bar{\phi}\|^q_{q,0,\infty} \leqq (2^q+1)C(q)^q(\|\bar{\theta}\|^q_{q,0,\infty} + \|A'\|^q_\infty \|\bar{\phi}\|^q_{q,0,\infty}).$$

Combining this with (3.7) yields

$$\|\Delta\bar{\phi}\|_{q,0,\infty} \leqq \bar{C}(q)\|\bar{\theta}\|_{q,0,\infty}$$

where $\bar{C}(q) = (2^q+1)^{1/q}C(q)(q + \|A'\|^q_\infty \delta^{-q})^{1/q}$. Applying this with $\bar{\theta} = 0$ for $t \geqq T-\tau$ yields the required estimate.

Assume now that $\alpha_2 = 0$. For $\psi \in \mathscr{L}^q(\Omega)$ set

$$\hat{\psi} = |\Omega|^{-1}\int_\Omega \psi \quad \text{and} \quad P\psi = \psi - \hat{\psi}.$$

If ϕ and θ are as in (3.1), then $\hat{\phi}_t = -\hat{\theta}$ since $\partial\phi/\partial n = 0$ on $\partial\Omega$, and

$$\begin{aligned} &(P\phi)_t = -d_2\Delta(P\phi) - P\theta, && x \in \Omega, \quad t \in (0, T-\tau), \\ &\frac{\partial}{\partial n}(P\phi) = 0, && x \in \partial\Omega, \quad t \in (0, T-\tau), \\ &P\phi(T) = 0. && \end{aligned}$$

Moreover, there exists $\delta > 0$ such that

$$|P\mathscr{T}_{2,q}(t)w|_q \leqq e^{-\delta t}|w|_q \quad \text{for } w \in \mathscr{L}^q(\Omega), \quad t \geqq 0.$$

Arguing as for $\alpha_2 \neq 0$ but with ϕ replaced by $P\phi$ completes the proof of Lemma 3.

Our final estimate relates ϕ and θ to the solution (u, v) of (1.1):

LEMMA 4. *Let the supposition of Proposition* 1 *be satisfied, let* $0 \leqq \tau < T < T^*$, *and let* q, θ *and* ϕ *be as in Lemma* 2 *with* $\theta \geqq 0$. *Then there are constants* $E_1(\tau, T)$, $E_2(\tau, T)$ *and* $P_0(q)$ *such that*

$$\text{(i)} \quad \int_{Q_{\tau,T}} u\theta \leqq \int_{Q_{\tau,T}} \phi f(t, u, v) + \int_\Omega \phi(\cdot, \tau)u(\cdot, \tau) + E_1(\tau, T),$$

$$\text{(ii)} \quad \int_{Q_{\tau,T}} v\theta \leqq \int_{Q_{\tau,T}} \phi g(t, u, v) + \int_\Omega \phi(\cdot, \tau)v(\cdot, \tau) + E_2(\tau, T),$$

$$\text{(iii)} \quad E_1(\tau, T) + E_2(\tau, T) \leqq P_0(q)[1 + \bar{N}_1(\tau, T)](T-\tau)^{1/p}\|\theta\|_{q,\tau,T}$$

where $\bar{N}_1(\tau, T) \equiv \sup\{N_1(t): \tau \leqq t < T\}$ *and* $1/p + 1/q = 1$. *Moreover, if* $T^* < \infty$ *and* $\sup\{N_1(t): 0 \leqq t < T^*\} < \infty$, *then* (i), (ii), *and* (iii) *hold with* $T = T^*$.

Proof. Integrating ϕu_t over Ω, we obtain

$$\begin{aligned} \int_\Omega \phi u_t &= \int_\Omega [d_1\phi\Delta u + \phi f(t, u, v)] \\ &= d_1\int_\Omega u\Delta\phi + \int_\Omega \phi f(t, u, v) + I_1(t) \end{aligned} \tag{3.8}$$

where $I_1(t) = \int_{\partial\Omega}(\phi(\partial u/\partial n) - u(\partial\phi/\partial n))$.

If $\alpha_1 = 0$, then $\alpha_2 = \beta_1 = \beta_2 = 0$ (see (H1)) and so $I_1(t) \equiv 0$. If $\alpha_1 = 1$ then $\alpha_2 = 1$, and so

$$I_1(t) = -\int_{\partial\Omega} u\frac{\partial\phi}{\partial n} = -\beta_1\int_\Omega \Delta\phi \leqq \beta_1\int_\Omega |\Delta\phi|.$$

Finally, if $0 < \alpha_1 < 1$ then $0 < \alpha_2 < 1$, and so (since $\partial\phi/\partial n \leqq 0$)

$$I_1(t)=\int_{\partial\Omega}\left[\frac{(\beta_1-\alpha_1 u)(\alpha_2-1)}{(1-\alpha_1)\alpha_2}\frac{\partial\phi}{\partial n}-u\frac{\partial\phi}{\partial n}\right]$$

$$\leqq\{(\beta_1+\alpha_1\bar{N}_1(\tau,T))(1-\alpha_2)[(1-\alpha_1)\alpha_2]^{-1}+\bar{N}_1(\tau,T)\}\int_{\partial\Omega}-\frac{\partial\phi}{\partial n}$$

$$=-c_0[1+\bar{N}_1(\tau,T)]\int_\Omega\Delta\phi$$

$$\leqq c_0[1+\bar{N}_1(\tau,T)]\int_\Omega|\Delta\phi|.$$

Thus, in each case, $I_1(t)\leqq c_0[1+\bar{N}_1(\tau,T)]\int_\Omega|\Delta\phi|$ for some constant $c_0\geqq 0$.

Now integrating ϕu_t from $t=\tau$ to $t=T$ we obtain

$$\int_\tau^T\phi u_t=-\phi(\cdot,\tau)u(\cdot,\tau)-\int_\tau^T u\phi_t$$

$$=-\phi(\cdot,\tau)u(\cdot,\tau)+d_2\int_\tau^T u\,\Delta\phi+\int_\tau^T u\theta$$

using the fact that $\phi(\cdot,T)\equiv 0$. Now by integrating each side of this equation over Ω and each side of (3.8) from $t=\tau$ to $t=T$ and interchanging the order of integration, we see that (i) holds with

$$E_1=(d_1-d_2)\int_\tau^T\int_\Omega u\Delta\phi+\int_\tau^T I_1(t)\,dt$$

$$\leqq[|d_1-d_2|\bar{N}_1(\tau,T)+c_0(1+\bar{N}_1(\tau,T))]\int_\tau^T\int_\Omega|\Delta\phi|$$

$$\leqq(|d_1-d_2|+c_0)[1+\bar{N}_1(\tau,T)](T-\tau)^{1/p}|\Omega|^{1/p}\|\Delta\phi\|_{q,\tau,T}$$

where $|\Omega|$ is the measure of Ω. Therefore, by Lemma 3, there is a $c_1(q)$ such that

$$E_1\leqq c_1(q)[1+\bar{N}_1(\tau,T)](T-\tau)^{1/p}\|\theta\|_{q,\tau,T}.$$

In the same manner, one can show that (ii) holds with $E_2=\int_\tau^T I_2(t)\,dt$ where $I_2(t)=\int_{\partial\Omega}\phi\,\partial v/\partial n-v\,\partial\phi/\partial n$. Estimates similar to those for $I_1(t)$ reveal that for each boundary condition there is a constant $\bar{c}_0$ such that $I_2(t)\leqq\bar{c}_0\int_\Omega|\Delta\phi|$ for all $t\in(\tau,T)$, and hence there is by (3.3) a $c_2(q)$ such that

$$E_2\leqq c_2(q)(T-\tau)^{1/p}\|\theta\|_{q,\tau,T}.$$

Assertion (iii) now follows, and the proof is complete.

4. Proof of the theorems. Theorems 1 and 2 are proved in this section, and it is assumed throughout that (H1)-(H6) hold and that (u,v) is the noncontinuable solution to (1.1) on $\Omega\times[0,T^*)$ guaranteed by Proposition 1. A crucial estimate is the following.

LEMMA 5. *Suppose that* $0\leqq\tau<T<T^*$ *and* $1<p<\infty$. *Then* $v\in\mathscr{L}^p(Q_{\tau,T})$ *and the estimate*

$$(4.1)\qquad\int_\tau^T|v(\cdot,t)|_p^p\,dt\leqq R_p(\bar{N}_1(\tau,T),C(q,T-\tau))^p[1+|v(\cdot,\tau)|_p+(T-\tau)^{1/p}]^p$$

is valid where $R_p:[0,\infty)^2\to[0,\infty)$ *is continuous,* q *is conjugate to* p*,* C *is as in Lemma* 2, *and*

$$\bar N_1(\tau, T)=\sup\{N_1(t): \tau\leqq t<T\}.$$

Moreover, if $T^*<\infty$ *then* (4.1) *holds with* $T=T^*$.

Proof. Inequalities (i) and (ii) in Lemma 4 imply that

$$\int_{Q_{\tau,T}} u\theta+\int_{Q_{\tau,T}} v\theta\leqq\int_{Q_{\tau,T}}\phi[f(t,u,v)+g(t,u,v)]$$
$$+\int_\Omega\phi(\cdot,\tau)[u(\cdot,\tau)+v(\cdot,\tau)]+E_1(\tau,T)+E_2(\tau,T)$$

where $\theta\in\mathscr{L}^q(Q_{\tau,T})$, $\theta\geqq 0$, and ϕ satisfies (3.1). By (H6) and (1.2),

$$\int_{Q_{\tau,T}} v\theta\leqq\int_{Q_{\tau,T}}\phi\mu_0(t,\bar N_1(\tau,T))+\int_\Omega\phi(\cdot,\tau)N_1(\tau)$$
$$+\int_\Omega\phi(\cdot,\tau)v(\cdot,\tau)+E_1(\tau,T)+E_2(\tau,T).$$

Applying Hölder's inequality,

$$\int_{Q_{\tau},T} v\theta\leqq\left[\int_\tau\int_\Omega\mu_0(t,\bar N_1(\tau,T))^p\,dt\right]^{1/p}\|\phi\|_{q,\tau,T}$$
$$+N_1(\tau)|\Omega|^{1/p}|\phi(\cdot,\tau)|_q+|v(\cdot,\tau)|_p|\phi(\cdot,\tau)|_q+E_1(\tau,T)+E_2(\tau,T).$$

Now using the estimates in Lemma 2 and (iii) in Lemma 4 we have that

$$\int_{Q_{\tau,T}} v\theta\leqq\{R_p(\bar N_1(\tau,T),C(q,T-\tau))[1+|v(\cdot,\tau)|_p+(T-\tau)^{1/p}]\}\|\theta\|_{q,\tau,T}.$$

Since this holds for every $\theta\in\mathscr{L}^q(Q_{\tau,T})$, $\theta\geqq 0$, we have by duality that $v\in\mathscr{L}^p(Q_{\tau,T})$ and that the term in the braces is an estimate for $\|v\|_{p,\tau,T}$. Thus (4.1) holds and since N_1 is assumed bounded on $[0,T^*)$ if $T^*<\infty$, the final assertion is immediate from (4.1).

Combining Lemma 5 with assumption (H5) gives the following estimate:

LEMMA 6. *With the suppositions of Lemma* 5, *suppose that* $r>1$ *and* γ *is as in* (H5). *Then there is a constant* $\bar L_r(T-\tau,\bar N_1(\tau,T))$ *such that*

$$\int_\tau^T|g(t,u(\cdot,t),v(\cdot,t))|_r^r\,dt\leqq\bar L_r(T-\tau,\bar N_1(\tau,T))\left[1+\int_\tau^T|v(\cdot,t)|_{\gamma r}^{\gamma r}\,dt\right]\tag{4.2}$$

whenever $0\leqq\tau<T<T^*$. *Also, this estimate is valid for* $T=T^*$ *if* $T^*<\infty$.

Proof. By (H5),

$$|g(t,u(x,t),v(x,t))|_r^r\leqq L_0(t,\bar N_1(\tau,T))^r[1+|v(x,t)|]^{\gamma r}$$
$$\leqq 2^{\gamma r}L_0(t,\bar N_1(\tau,T))^r[1+|v(x,t)|^{\gamma r}]$$

and, integrating over Ω,

$$|g(t,u(\cdot,t),v(\cdot,t))|_r^r\leqq 2^{\gamma r}L_0(t,\bar N_1(\tau,T))^r[|\Omega|+|v(\cdot,t)|_{\gamma r}^{\gamma r}].$$

Now (4.2) follows by integrating each side of this inequality from $t=\tau$ to $t=T$.

Proof of Theorem 1. Suppose, for contradiction, that $T^*<\infty$. From (H4) and (1.3) it follows that

$$\lim_{t\uparrow T^*}|v(\cdot,t)|_\infty=\infty.\tag{4.3}$$

Let $\alpha \in (0, 1)$ and select $p > 1$ large enough so that $\alpha > n/(2p)$ (see (2.3)) and $\alpha q < 1$ where $p^{-1} + q^{-1} = 1$. Since $T^* < \infty$ and $|v_0|_r \leqq |\Omega|^{1/r} |v_0|_\infty$ for each $r > 1$, we have by (4.1) that

$$\int_0^{T^*} |v(\cdot, t)|_r^r \, dt \leqq M_1(r)^r.$$

Hence, by (4.2) there is an $M_2(p)$ such that

$$\int_0^{T^*} |g(t, u(\cdot, t), v(\cdot, t))| \, dt \leqq M_2(p)^P [1 + M_1(\gamma p)^{\gamma p}]^p. \tag{4.4}$$

Using the second equation in (2.6) and the properties of $B_{2,p}^\alpha$ in Lemma 1 we see that

$$\begin{aligned} |B_{2,p}^\alpha (v(t) - z_2)|_p &\leqq C_{\alpha,p} t^{-\alpha} \, e^{\omega t} |v_0 - z_2|_p + \int_0^t C_{\alpha,p} (t-s)^{-\alpha} |G(s, u(s), v(s))|_p \, ds \\ &\leqq C_{\alpha,p} t^{-\alpha} \, e^{\omega} \tau |v_0 - z_2|_p \\ &\quad + \left[\int_0^t C_{\alpha,p}^q (t-s)^{-\alpha q} \, ds \right]^{1/q} \left[\int_0^t |G(s, u(s), v(s))|_p^p \, ds \right]^{1/p}. \end{aligned}$$

By (4.4) and the definition of G in (2.4)

$$\begin{aligned} \int_0^t |G(s, u(s), v(s))|_p^p \, ds &\leqq \int_0^{T^*} |G(s, u(s), v(s))|_p^p \, ds \\ &\leqq M_2(p)^p [1 + M_1(\gamma p)^{\gamma p}]^p, \end{aligned}$$

and since $\alpha q < 1$ it follows that

$$\limsup_{t \uparrow T^*} |B_{2,p}^\alpha (v(t) - z_2)|_p < \infty.$$

Since $\alpha > n/(2p)$, we have by (2.3) that

$$|v(t) - z_2|_\infty \leqq M_{\alpha,p} |B_{2,p}^\alpha (v(t) - z_2)|_p,$$

which contradicts (4.3). Thus $T^* = \infty$, and the proof of Theorem 1 is complete.

Now assume that the suppositions of Theorem 2 are satisfied, and hence that (u, v) exists on $\Omega \times [0, \infty)$ by Theorem 1. If $\alpha_2 \neq 0$, by Lemma 3 one can assume that $C(q, T - \tau) \leqq \bar{C}(q)$. Since $N_1(t) \leqq \bar{N}_1$ for all $t \geqq 0$, from (4.1) we obtain

$$\int_\tau^T |v(\cdot, t)|_p^p \, dt \leqq \bar{R}_p^p [1 + |v(\cdot, \tau)|_p + (T - \tau)^{1/p}]^p \quad \text{for } T \geqq \tau \geqq 0 \tag{4.5}$$

for some $\bar{R}_p > 0$. If $\alpha_2 = 0$ and $\mu_\infty(r) \equiv 0$, then (4.5) remains valid. Indeed, from Lemma 4, we have

$$\int_{Q_{\tau,T}} v\theta \leqq \int_\Omega \phi(\cdot, \tau)[u(\cdot, \tau) + v(\cdot, \tau)] + E_1(\tau, T) + E_2(\tau, T)$$

where here $E_2 = 0$ and $E_1 \leqq |d_1 - d_2| \int_{Q_{\tau,T}} u |\Delta \phi|$. By Lemma 3,

$$E_1 \leqq |d_1 - d_2| \bar{N}_1 (T - \tau)^{1/p} |\Omega|^{1/p} \bar{C}(q) \|\theta\|_{q,\tau,T}.$$

Using again that $\mu_\infty(r) \equiv 0$ and $\alpha_2 = 0$, we have

$$\int_\Omega [u(+, \tau) + v(\cdot, \tau)] \leqq \int_\Omega u_0 + v_0.$$

Thus, if $\hat{\phi}$ and $P\phi$ are as in Lemma 3, we have

$$\int_\Omega \phi(\cdot,\tau)[u(\cdot,\tau)+v(\cdot,\tau)] = \int_\Omega [\hat{\phi}(\cdot,\tau)+P\phi(\cdot,\tau)][u(\cdot,\tau)+v(\cdot,\tau)]$$

$$\le \hat{\phi}(\cdot,\tau)\int_\Omega [u_0+v_0]+[\bar{N}_1|\Omega|^{1/p}+|v(\cdot,\tau)|_p]|P\phi(+,\tau)|_q$$

where

$$\hat{\phi}(\cdot,\tau)=\int_\tau^T \hat{\theta}(\cdot,s)\,ds \le |\Omega|^{1/p-1}(T-\tau)^{1/p}\|\theta\|_{q,\tau,T}.$$

Recalling Lemma 3 to estimate $|P\phi(\cdot,\tau)|_p$ and combining these inequalities yields (4.5).

LEMMA 7. *Let the suppositions of Theorem 2 be satisfied. For each* $p\in(1,\infty)$ *there are constants* $\Lambda_0(p)$ *and* $\Gamma_0(p)$ *and a sequence* $\{t_k\}_0^\infty$ *in* $[0,\infty)$ *such that* $t_0=0$ *and for each* $k\ge 0$,

(i) $1\le t_{k+1}-t_k\le\Lambda_0(p)$,

(ii) $|v(\cdot,t_k)|_p\le\bar{R}_p+1$,

(iii) $\int_{t_k}^{t_{k+1}}|v(\cdot,t)|_p^p\,dt\le\Gamma_0(p)$

where $\bar{R}_p$ *is as in* (4.5).

Proof. We assume without loss that $\bar{R}_p+1\ge|v_0|_p$ and that $\bar{R}_p\ge 1$. For notational convenience set $\bar{S}_p=\bar{R}_p^p(\bar{R}_p+2)^p$. Note first that

$$\text{(4.6)}\qquad \begin{array}{l}\text{if } \tau\ge 0 \text{ and } |v(\cdot,\tau)|_p\le\bar{R}_p+1, \text{ then there is a}\\ t^*\in(\tau,\tau+S_p] \text{ such that } |v(\cdot,t^*)|_p\le\bar{R}_p+1.\end{array}$$

For if this were not the case, then $|v(\cdot,t)|_p>\bar{R}_p+1$ for all $t\in(\tau,\tau+\bar{S}_p)$ and so

$$\int_\tau^{\tau+\bar{S}_p}|v(\cdot,t)|_p^p\,dt>\bar{S}_p(\bar{R}_p+1)^p.$$

However, by (4.5) with $T=\tau+\bar{S}_p$,

$$\int_\tau^{\tau+\bar{S}_p}|v(\cdot,t)|_p^p\,dt\le\bar{R}_p^p[1+\bar{R}_p+1+\bar{S}_p^{1/p}]^p$$

$$=\bar{R}_p^p[\bar{R}_p+2+\bar{R}_p(\bar{R}_p+2)]^p$$

$$=\bar{S}_p(\bar{R}_p+1)^p$$

which is impossible. Thus (4.6) is valid. Now set $\Lambda_0(p)=2\bar{S}_p$ and inductively define $\{t_k\}_0^\infty$ by $t_0=0$ and

$$t_{k+1}=\sup\{T\ge t_k\colon T-t_k\le 2\bar{S}_p \text{ and } |v(\cdot,T)|_p\le\bar{R}_p+1\}.$$

By the continuity of $t\mapsto|v(\cdot,t)|_p$ we have that $|v(\cdot,t_{k+1})|_p\le\bar{R}_p+1$ and $|v(\cdot,t_{k+1})|_p=\bar{R}_p+1$ if $t_{k+1}<t_k+2\bar{S}_p$. Moreover, $t_{k+1}-t_k\ge\bar{S}_p\ge 1$, for if $t_{k+1}-t_k<\bar{S}_p$ then (4.6) implies that there is a $t^*\in(t_{k+1},t_{k+1}+\bar{S}_p)$ such that $|v(\cdot,t^*)|_p\le\bar{R}_p+1$. But $t^*\le t_{k+1}+\bar{S}_p<t_k+2\bar{S}_p$, and this contradicts the definition of t_{k+1}. Thus $\{t_k\}_0^\infty$ is well defined and satisfies (i) and (ii). Since (4.5) implies that

$$\int_{t_k}^{t_{k+1}}|v(\cdot,t)|_p^p\,dt\le\bar{R}_p^p[1+|v(\cdot,t_k)|_p+(t_{k+1}-t_k)^{1/p}]^p$$

$$\le\bar{R}_p^p[\bar{R}_p+2+(2\bar{S}_p)^{1/p}]^p$$

we see that (iii) holds with $\Gamma_0(p)=\bar{R}_p^p[\bar{R}_p+2+(2\bar{S}_p)^{1/p}]^p$.

LEMMA 8. *Suppose that the assumptions in Theorem* 2 *are satisfied*, $p \in (1, \infty)$, γ *is as in* (H5), *and that* $\{t_k\}_0^\infty$ *is the sequence constructed in Lemma* 7 *corresponding to* γp. *Then there is a* $\Gamma_1(p, \gamma p)$ *such that*

$$\int_{t_k}^{t_{k+1}} |g(t, u(\cdot, t), v(\cdot, t))|_p^p \leqq \Gamma_1(p, \gamma p) \quad \textit{for all } k \geqq 0. \tag{4.7}$$

Proof. The boundedness assumptions on μ_0, L_0, and N_1 applied to (4.2) and Lemma 7 imply that

$$\begin{aligned}\int_{t_k}^{t_{k+1}} |g(t, u(\cdot, t), v(\cdot, t))|_p^p \, dt &\leqq M_3(p)\left[1 + \int_{t_k}^{t_{k+1}} |v(\cdot, t)|_{\gamma p}^{\gamma p} \, dt\right] \\ &\leqq M_3(p)[1 + \Gamma_0(\gamma p)] \\ &= \Gamma_1(p, \gamma p).\end{aligned}$$

$M_3(p)$ and consequently $\Gamma_1(p, \gamma p)$ are independent of k, and the proof is complete.

Proof of Theorem 2. Select $\alpha \in (0, 1)$ and $p \in (1, \infty)$ so that $\alpha > n/(2p)$ and $q < 1$, where $p^{-1} + q^{-1} = 1$. Then $\mathscr{D}(B_{2,p}^\alpha)$ is continuously embedded in $\mathscr{L}^\infty(\Omega)$ (see (2.3)); thus to establish the boundedness of $|v(\cdot, t)|_\infty$ on $[0, \infty)$ it suffices to show that

$$\sup\{|B_{2,p}^\alpha(v(t) - z_2)|_p : t \geqq 0\} < \infty. \tag{4.8}$$

So let $\{t_k\}_0^\infty$ be the sequence constructed in Lemma 7 corresponding to γp and note that for $k \geqq 2$ and $t \in [t_k, t_{k+1}]$,

$$\begin{aligned}B_{2,p}^\alpha(v(t) - z_2) = B_{2,p}^\alpha &\mathscr{T}_{2,p}(t - t_{k-1})(v(t_{k-1}) - z_2) \\ &+ \int_{t_{k-1}}^t B_{2,p}^\alpha \mathscr{T}_{2,p}(t - s) G(s, u(s), v(s)) \, ds\end{aligned}$$

(replace t by $t - t_{k-1}$ and v_0 by $v(t_{k-1})$ in (2.6)). By (ii) in Lemma 1,

$$\begin{aligned}|B_{2,p}^\alpha(v(t) - z_2)|_p \leqq\ &C_{\alpha,p}(t - t_{k-1})^{-\alpha} |v(t_{k-1}) - z_2|_p \\ &+ \int_{t_{k-1}}^t C_{\alpha,p}(t - s)^{-\alpha} |G(s, u(s), v(s))|_p \, ds.\end{aligned}$$

Since $t \geqq t_k \geqq t_{k-1} + 1$, we have from (ii) in Lemma 7 that

$$\begin{aligned}C_{\alpha,p}(t - t_{k-1})^{-\alpha} |v(t_{k-1}) - z_2|_p &\leqq C_{\alpha,p}(|v(t_{k-1})|_p + |z_2|_p) \\ &\leqq C_{\alpha,p}(|v(t_{k-1}) + 1|_{\gamma p}^\gamma + |z_2|_p) \\ &\leqq C_{\alpha,p}[(|v(t_{k-1})|_{\gamma p} + |1|_{\gamma p})^\gamma + |z_2|_p] \\ &\leqq C_{\alpha,p}[(\bar{R}_{\gamma p} + 1 + |\Omega|^{1/\gamma p})^\gamma + |z_2|_p] \\ &\equiv M_4.\end{aligned}$$

Also, by Hölder's inequality and Lemma 8 and since $\alpha q < 1$, we have

$$\begin{aligned}&\int_{t_{k-1}}^t C_{\alpha,p}(t - s)^{-\alpha} |G(s, u(s), v(s))|_p \, ds \\ &\quad\leqq C_{\alpha,p}\left[\int_{t_{k-1}}^t (t - s)^{-\alpha q} \, ds\right]^{1/q} \left[\int_{t_{k-1}}^t |G(s, u(s), v(s))|_p^p \, ds\right]^{1/p} \\ &\quad\leqq M_5\left[\int_{t_{k-1}}^{t_{k+1}} |G(s, u(s), v(s))|_p^p \, ds\right]^{1/p} \\ &\quad\leqq 2^{1/p} M_5 \Gamma_1(p, \gamma p)^{1/p}.\end{aligned}$$

Thus, for $t\in[t_k, t_{k+1}]$ with $k\geqq 2$,

$$|B^{\alpha}_{2,p}(v(t)-z_2)|_p\leqq M_4+2^{1/p}M_5\Gamma_1(p,\gamma p)^{1/p}.$$

So (4.8) holds, and the proof of Theorem 2 is complete.

5. An example and a counterexample. We begin this section by showing how the preceeding results apply to the Brusselator mentioned in the introduction:

$$(5.1)\quad \begin{array}{lll} \text{(a)} & \begin{array}{l} u_t=d_1\Delta u-uv^2+Bv, \\ v_t=d_2\Delta v+uv^2-(B+1)v+A, \end{array} & x\in\Omega,\quad t>0, \\ \text{(b)} & u(x,t)=B/A,\quad v(x,t)=A, & x\in\partial\Omega,\quad t>0, \\ \text{(c)} & u(x,0)=u_0(x),\quad v(x,0)=v_0(x), & x\in\bar{\Omega}. \end{array}$$

It is assumed that A, B, d_1, and d_2 are positive constants and that u_0 and v_0 are nonnegative, uniformly bounded, measurable functions on $\bar{\Omega}$. The following result is valid for the solution to (5.1).

PROPOSITION 2. *The system* (5.1) *has a unique solution* (u, v) *on* $\Omega\times[0,\infty)$, *and there is a constant* $M>0$ *such that*

$$0\leqq u(x,t),\quad v(x,t)\leqq M \quad \textit{for all } t\geqq 0 \textit{ and } x\in\Omega.$$

In order to establish this result we show that (5.1) satisfies the suppositions of Theorem 2 with

$$f(t,\xi,\eta)\equiv-\xi\eta^2+B\eta,\qquad g(t,\xi,\eta)\equiv\xi\eta^2-(B+1)\eta+A.$$

It is immediate that (H1)-(H3) are satisfied, and that (H5) holds with $\gamma=2$ and L_0 independent of $t\geqq 0$. Moreover, since

$$f(t,\xi,\eta)+g(t,\xi,\eta)=A-\eta\leqq A$$

for all t, ξ, $\eta\geqq 0$, we see that (H6) is valid with $\mu_0(t,r)\equiv A$. Therefore, Proposition 2 will be established once it is shown that u remains uniformly bounded on $\Omega\times[0,\infty)$.

Proof of Proposition 2. As asserted in the preceding paragraph, we need only show that $|u(\cdot,t)|_\infty$ remains uniformly bounded so long as the solution to (5.1) exists. Since the suppositions of Proposition 1 are satisfied, there is a $T^*>0$ such that (u, v) exists on $\Omega\times[0, T^*)$. So let N_1, N_2 be as in (1.2) in Proposition 1 and let $0<t_0<T^*$. Define $\delta=\min\{v(x,t_0): x\in\bar{\Omega}\}$ and note that $\delta>0$. For if $\delta=v(x^*,t_0)=0$ then $x^*\in\Omega$, and so $\Delta v(x^*,t_0)\geqq 0$. By substitution of (x^*, t_0) into the second equation in (5.1a) it follows that $v_t(x^*,t_0)\geqq A>0$, which is impossible since this implies $v(x^*,t)<0$ for $t\in[t_0-\varepsilon, t_0)$ where $\varepsilon>0$. To complete the proof we show that

$$(5.2)\quad u(x,t)\leqq\max\{|u(\cdot,t_0)|_\infty, B/\delta, B(B+1)/A\}\equiv N_1(t)\quad \text{for all } (x,t)\in\Omega\times[t_0,T^*).$$

First, however, observe that

$$(5.3)\qquad v(x,t)\geqq\min\{\delta, A/(B+1)\}\quad \text{for all } (x,t)\in\bar{\Omega}\times[t_0,T^*).$$

For otherwise, if $t_0<t_1<T^*$ and

$$\begin{aligned} v(x^*,t^*)&=\min\{v(x,t): (x,t)\in\bar{\Omega}\times[t_0,t_1]\} \\ &<\min\{\delta, A/(B+1)\}, \end{aligned}$$

then $v(x^*,t^*)<A$ and so $x^*\in\Omega$ and $\Delta v(x^*,t^*)\geqq 0$. Thus

$$\begin{aligned} v_t(x^*,t^*)&\geqq-(B+1)v(x^*,t^*)+A \\ &>-(B+1)A/(B+1)+A=0, \end{aligned}$$

and so it follows that $t^*=t_0$. But if $t^*=t_0$ then $v(x^*, t^*) \geqq \delta$ by the definition of δ, and we have a contradiction. This establishes (5.3). Now assuming for contradiction that (5.2) is not true, it follows that there is a $t_1 \in (t_0, T^*)$ such that

$$\begin{aligned} u(x^*, t^*) &= \max \{u(x, t) \colon (x, t) \in \bar{\Omega} \times [t_0, t_1]\} \\ &> \max \{|u(\cdot, t_0)|_\infty, B/\delta, B(B+1)/A\}. \end{aligned}$$

Then $t^* > t_0$ since $u(x^*, t^*) > |u(\cdot, t_0)|_\infty$, and $x^* \in \Omega$ since $u(x^*, t^*) > B/A$. Thus $\Delta u(x^*, t^*) \leqq 0$. Therefore,

$$u_t(x^*, t^*) \leqq -u(x^*, t^*)v(x^*, t^*)^2 + Bv(x^*, t^*) = v(x^*, t^*)^2[Bv(x^*, t^*)^{-1} - u(x^*, t^*)].$$

By (5.3), $\max \{B/\delta, B(B+1)/A\} \geqq Bv(x^*, t^*)^{-1}$. Thus $u(x^*, t^*) > Bv(x^*, t^*)^{-1}$, and so $u_t(x^*, t^*) < 0$. This is not possible, since $t^* > t_0$. Thus (5.2) holds, and hence Theorem 2 implies the boundedness of solutions to (5.1).

We remark here that even though $u_\infty(r) \equiv A > 0$, our techniques can be used to show that the solution to (5.1) exists and is uniformly bounded on $\Omega \times [0, \infty)$ when the boundary conditions are Neumann type rather than Dirichlet (cf. Rothe [11]). This is accomplished by incorporating the $-v$ term in the equation for v into the differential operator. The semigroups $T^{\#}_{2,p}$, $1 < p < \infty$, generated by the corresponding perturbations of the operators $A_{2,p}$ are contractive rather than nonexpansive. Hence the estimates in Lemma 3 hold for the solution σ of

$$\begin{aligned} &\sigma_t = -(d_2\Delta\sigma - \sigma) - \theta && \text{on } Q_{\tau,T}, \\ &\frac{\partial\sigma}{\partial n} = 0 && \text{on } \partial\Omega \times (\tau, T), \\ &\sigma(\cdot, T) = 0 && \text{on } \Omega \end{aligned}$$

without the aid of the operator P. Using σ in place of ϕ in Lemmas 4 and 5 yields similar results, and hence (4.5) holds. The proof then proceeds exactly as before.

Let us now consider the model

$$(5.4) \qquad \begin{aligned} &\left.\begin{aligned} &u_t = d_1\Delta u - uv^\beta, \\ &v_t = d_2\Delta v + uv^\beta, \end{aligned}\right\} && x \in \Omega, \quad t > 0, \\ &\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0, && x \in \partial\Omega, \quad t > 0, \\ &u = u_0, \quad v = v_0, && x \in \Omega, \quad t = 0 \end{aligned}$$

where $0 \leqq u_0$, $v_0 \leqq M$ and d_1, $d_2 > 0$. It is easy to see that this model (with these and other boundary conditions) fits into our result of Theorem 2, since $|u(\cdot, t)|_\infty$ is obviously a priori bounded for $t \geqq 0$. As mentioned in the introduction, global existence and uniform boundedness of solutions to (5.4) were proved by Masuda [8].

It is interesting to notice that the assumption $d_2 > 0$ is essential here. Indeed, although the situation is good when either d_1, $d_2 > 0$ or $d_1 = d_2 = 0$, one can prove that, if $d_2 = 0$ and $\beta > 1$, there is no $\mathscr{L}^\infty$-estimate of the solution of (5.4) in terms of $|u_0|_\infty$ and $|v_0|_\infty$ for t large. More precisely we have

PROPOSITION 3. *Assume $d_2 = 0$, $d_1 > 0$, and $\beta > 1$. Let $a, b > 0$ and suppose that for all u_0, v_0 satisfying*

$$0 \leqq u_0 \leqq 1, \qquad 0 \leqq v_0 \leqq b$$

there exists a solution to (5.4) *on* $[0, T)$ *such that for each* $t \in [0, T)$,

$$|v(\cdot, t)|_\infty \leqq c(T, a, b).$$

Then

$$T \leqq [(\beta-1)b^{\beta-1}a]^{-1}.$$

Proof. Let v_0 be a nonnegative and compactly supported $\mathscr{L}^\infty$ function on Ω. If (5.4) has a solution on $[0, T)$, since $d_2 = 0$, the equation in v can be explicitly solved as

$$v(x,t) = v_0(x)[1-(\beta-1)v(x)^{\beta-1}U(x,t)]^{-1/(\beta-1)} \tag{5.5}$$

where $U(x,t) \equiv \int_0^t u(x,s)\,ds$ (we use here the uniqueness of bounded solutions of (5.4)). Let us now consider (u_k, v_k) the solution of (5.4) with the initial data

$$u_k(x,0) \equiv a \quad \text{for all } x \in \Omega$$

and $v_k(x,0)$ a compactly supported $\mathscr{C}^\infty$ function on Ω satisfying

$$0 \leqq v_k(x,0) \leqq b, \tag{5.6}$$

$$v_k(x_0, 0) = b \quad \text{for some } x_0 \in \Omega \text{ independent of } k, \tag{5.7}$$

$$v_k(x,0) \to 0 \quad \text{for } x \neq x_0 \quad \text{as } k \to \infty. \tag{5.8}$$

Assume that $v_k(x,t)$ remains uniformly bounded on $\Omega \times [0, T)$. Then by (5.5) and (5.7) we have

$$(\beta-1)b^{\beta-1}U_k(x_0, T) \leqq 1. \tag{5.9}$$

Now, by (5.5) and (5.8) and the fact that $U_k(x,t)$ is uniformly bounded on $\Omega \times [0, T)$, $v_k(x,t)$ tends to 0 for all $t \in [0, T)$ and $x \neq x_0$. As u_k and v_k are uniformly bounded, $u_k v_k^\beta$ tends to 0 in any $\mathscr{L}^p(\Omega \times (0, T))$ for $1 \leqq p < \infty$. But for p large enough, the mapping $\zeta \to w$ where

$$w_t = d_1 \Delta w + \zeta,$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

$$w(x,0) = a \quad \text{on } \Omega$$

is compact from $\mathscr{L}^p(\Omega \times (0, T))$ into $\mathscr{C}([0, T] \times \bar{\Omega})$ (see Ladyzenskaja et al. [5]). Therefore, u_k converges uniformly on $\bar{\Omega} \times [0, T]$ to the constant solution $u \equiv a$ of

$$u_t = d_1 \Delta u,$$

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

$$u(x,0) = a \quad \text{on } \Omega.$$

Passing to the limit in (5.9) yields $(\beta-1)b^{\beta-1}Ta \leqq 1$.

Acknowledgment. We would like to thank Professor L. C. Evans for his helpful suggestions regarding the estimates in § 3.

REFERENCES

[1] N. D. ALIKAKOS, *L^p bounds of solutions of reaction-diffusion equations*, Comm. Partial Differential Equations, 4 (1979), pp. 827-868.

[2] J. F. G. AUCHMUTY AND G. NICOLIS, *Bifurcation analysis of nonlinear reaction-diffusion equations—I. Evolution equations and steady state solutions*, Bull. Math. Biol., 37 (1975), pp. 323-365.

[3] J. M. BALL, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford Ser. (2), 28 (1977), pp. 473-486.

[4] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 840, Springer-Verlag, New York, 1981.

[5] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URALCEVA, *Linear and quasilinear equations of parabolic type*, Trans. Math. Monographs Vol. 23, American Mathematical Society, Providence, RI, 1968.

[6] H. LANGE, *Die globale Losbarkeit einer nichtlinearen Reaktionsgleichung*, Math. Nachr., 80 (1977), pp. 165-181.

[7] J. H. LIGHTBOURNE AND R. H. MARTIN, *Relatively continuous nonlinear perturbations of analytic semigroups*, J. Nonlinear Anal.-Theory Meth. Appl., 1 (1977), pp. 277-292.

[8] K. MASUDA, *On the global existence and asymptotic behavior of solutions of reaction-diffusion equations*, Hokkaido Math. J., 12 (1983), pp. 360-370.

[9] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Applied Math. Sciences, 44, Springer-Verlag, New York, 1983.

[10] E. PRIGOGINE AND G. NICOLIS, *Biological order, structure and instabilities*, Quart. Rev. Biophys., 4 (1971), pp. 107-148.

[11] F. ROTHE, *Global solutions of reaction-diffusion systems*, Lecture Notes in Mathematics 1072, Springer-Verlag, Berlin, 1984.

[12] A. M. TURING, *The chemical basis of morphogenesis*, Philos. Trans. Roy. Soc. London Ser. B, 237 (1952), pp. 37-72.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

014

ASYMPTOTICS AND AN ASYMPTOTIC GALERKIN METHOD FOR HYPERBOLIC-PARABOLIC SINGULAR PERTURBATION PROBLEMS*

BENJAMIN F. ESHAM, JR.†

Abstract. A composite asymptotic expansion including initial layer corrections is developed for treating initial boundary value problems for hyperbolic equations with a small parameter multiplying the second-order time derivative term. Proof of uniform asymptotic validity is given a Hilbert space setting. The expansion forms the basis of a continuous-time Galerkin procedure for which error analysis based on the finite element method is included. This work extends recent results of Hsiao and Weinacht in two directions, one toward greater abstraction and one toward greater utility.

Key words. asymptotics, Galerkin method, singular perturbation

AMS (MOS) subject classification. 65

1. Introduction. Let Ω be a bounded domain in R^n, a generic point of which will be denoted by $x=(x_1,\cdots,x_n)$, with smooth boundary $\partial\Omega$ and let $t_0>0$ be arbitrary. We shall be interested in approximating the solution $u(x,t)$ of the initial-boundary value problem (P_ε):

$$\varepsilon^2 u_{tt}(x,t)+u_t(x,t)+Au(x,t)=f(x,t), \qquad \Omega\times(0,t_0), \tag{1.1}$$

$$u(x,t)=0, \qquad \partial\Omega\times[0,t_0], \tag{1.2}$$

$$u(x,0)=u_0(x), \qquad x\in\bar{\Omega}, \tag{1.3}$$

$$\varepsilon u_t(x,0)=u_1(x), \qquad x\in\bar{\Omega}, \tag{1.4}$$

using a combination of asymptotic and numerical techniques.

In (1.1) and (1.4) the parameter ε is assumed to be small, $0<\varepsilon\ll 1$. Here A is the second-order uniformly strongly elliptic operator in $L^2(\Omega)$ defined on $D(A)=H^2(\Omega)\cap H_0^1(\Omega)$ by

$$Au=-\sum_{i,j=1}^{n} D_i(a_{ij}(x)D_ju)+c(x)u, \tag{1.5}$$

in which $a_{ij}=a_{ji}\varepsilon C^\infty(\bar{\Omega})$ for $i,j=1,\cdots,n$, $c(x)\varepsilon C^\infty(\bar{\Omega})$, $c(x)\geqq 0$ and D_i denotes the derivative $\partial/\partial x_i$, $i=1,\cdots,n$. We assume the ellipticity condition

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq \Gamma \sum_{i=1}^{n} \xi_i^2$$

holds for all $\xi=(\xi_1,\cdots,\xi_n)\varepsilon R^n$ and $x\in\bar{\Omega}$. The ε-independent data $\{f,u_0,u_1\}$ is subject to certain regularity conditions specified later (see Theorem 2.1).

Problem (P_ε) is a singular perturbation problem of hyperbolic-parabolic type that has been investigated by many authors. In [21] Zlamal has used Fourier integral methods and in [22] has used eigenfunction expansions. The method of energy integrals has been exploited by Bobisud [2] and Hsiao and Weinacht [9]. In the terminology of singular perturbation theory, there is initial layer behavior in a neighborhood of $t=0$. The solution of (P_ε) is dominated by the solution $U_0(x,t)$ of the reduced problem (P_0), an initial-boundary value problem of parabolic type, obtained by setting $\varepsilon=0$.

* Received by the editors April 8, 1985; accepted for publication (in revised form) February 19, 1986.

† Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, Virginia 23284.

The solution $U_0(x, t)$ cannot generally be expected to satisfy the initial condition (1.4), so one is led to consider initial-layer correction terms, which appear in higher order expansions.

For the proof of uniform validity of the expansion, we reformulate the problem in a Hilbert space setting in § 2. The result to first order for a closely related problem was obtained by Kisynski [13] using semigroups in Hilbert space. For related results see Nur [16] and Schoene [17]. Fattorini [4], [5] has recently used cosine families in Banach space to obtain an N-term composite expansion. Our development extends the work of Hsiao and Weinacht [9] to an abstract setting relying heavily on the positive definiteness of A and the inner product induced norm on the Hilbert space.

For the numerical solution of (P_ε) one may be tempted to solve the singular perturbation problem directly by means of standard methods such as the Crank-Nicolson-Galerkin scheme since this is a linear hyperbolic problem for each $\varepsilon > 0$. However, because of the presence of ε, the usual schemes will not yield meaningful numerical results without reducing the mesh size in the initial layer. This, of course, requires considerable computational effort. Also, the discrete problems involved may become ill-posed numerically when mesh sizes get to be too small. (See [8] and [12].)

In this paper we present a continuous-in-time numerical procedure for treating singularly perturbed problems such as (P_ε), which does not require a very fine mesh. The procedure uses singular perturbation theory to construct problems for the leading terms in the formal asymptotic expansion, which are then solved numerically by the Galerkin method with finite elements as the trial functions for the space variables. This leads to an initial-value problem for a system of ordinary differential equations for which explicit solutions can be constructed. We refer to this approximation as the asymptotic Galerkin approximation. Preliminary numerical implementation of the problems show that very accurate results may be obtained in an efficient way. (For a similar approach to singular perturbation problems of parabolic-elliptic type see [8] and [12].)

In § 2 we prove the uniform validity of an N-term composite asymptotic expansion. In § 3 we formulate and describe the approximation scheme in detail. Section 4 contains error estimates for the continuous-in-time asymptotic Galerkin approximation using finite element methods. The results of some numerical experiments are given in § 5. Throughout this paper, C will denote a generic constant, not necessarily the same in any two places. Also, sums over empty index sets are zero.

2. The asymptotic expansion. In this section a procedure is developed for the construction of an N-term asymptotic expansion with initial layer corrections for the singularly perturbed Cauchy problem (P_ε^*):

$$L_\varepsilon[u] := \varepsilon^2 u''(t) + u'(t) + Au(t) = F(t), \tag{2.1}$$

$$u(0) = u_0, \qquad \varepsilon u'(0) = u_1 \tag{2.2}$$

in a Hilbert space H, with inner product $(\cdot, \cdot)$, norm $\|\cdot\|$, and zero element θ. The linear operator A is assumed to be positive-definite and self-adjoint in H. Problem (P_ε^*) is an abstract analogue of problem (P_ε) specified in (1.1)-(1.4) of § 1.

A formal asymptotic expansion may be obtained in the form of a composite expansion familiar from singular perturbation theory (see [15] and [19]), namely,

$$u_N(t; \varepsilon) = E_0^N(t; \varepsilon) + E_L^{N+1}(t/\varepsilon^2; \varepsilon). \tag{2.3}$$

Here the outer expansion,

$$E_0^N(t; \varepsilon) := \sum_{n=0}^{N} U_n(t)\varepsilon^n, \tag{2.4}$$

provides a uniformly valid $O(\varepsilon^{N+1})$ approximation to $u(t;\,\varepsilon)$ on intervals $[\delta, t_0]$, $\delta>0$, bounded away from $t=0$. The initial layer expansion

$$\text{(2.5)}\qquad E_L^{N+1}(t/\varepsilon^2;\,\varepsilon):=\sum_{n=0}^{N+1} V_n(t/\varepsilon^2;\,\varepsilon)\varepsilon^{n+1}$$

is asymptotically zero except in a neighborhood of the initial point $t=0$. We introduce the stretched variable $\tau=t/\varepsilon^2$ inside the initial layer. The expansion (2.3) contains two terms V_N and V_{N+1}, which are not useful in the final approximation, but are necessary in the proof of uniform validity. The construction of such terms of higher order is typical in singular perturbation problems.

We say that a function $z(t;\,\varepsilon)$ in $C([0, t_0], H)$ is $O(\varepsilon^k)$ as $\varepsilon\to 0^+$ uniformly in $[0, t_0]$ if there are positive constants C and ε_0 such that $\|z(t;\,\varepsilon)\|\leqq C\varepsilon^k$ holds for all t in $[0, t_0]$ and ε in $(0, \varepsilon_0]$.

A calculation shows that

$$\text{(2.6)}\qquad \begin{aligned} L_\varepsilon[u_N]=&\sum_{n=0}^{N}(U_n'(t)+AU_n(t)-U_{n-2}''(t))\varepsilon^n\\ &+\sum_{n=-1}^{N}(\ddot V_{n+1}(\tau)+\dot V_{n+1}(\tau)+AV_{n-1}(\tau))\varepsilon^n+R(t;\,\varepsilon),\end{aligned}$$

where

$$\text{(2.7)}\qquad R(t;\,\varepsilon):=U_{N-1}''(t)\varepsilon^{N+1}+U_N''(t)\varepsilon^{N+2}+AV_N(t/\varepsilon^2)\varepsilon^{N+1}+AV_{N+1}(t/\varepsilon^2)\varepsilon^{N+2}$$

and functions with negative index vanish identically. We shall consistently use primes for t-derivatives and dots for τ-derivatives.

Putting $t=0$ in (2.3) and its derivative we find

$$\text{(2.8)}\qquad u_N(0;\,\varepsilon)=U_0(0)+\sum_{n=1}^{N}(U_n(0)+V_{n-1}(0))\varepsilon^n+V_N(0)\varepsilon^{N+1}+V_{N+1}(0)\varepsilon^{N+2}$$

and

$$\text{(2.9)}\qquad \varepsilon u_N'(0;\,\varepsilon)=\dot V_0(0)+\sum_{n=0}^{N}(U_n'(0)+\dot V_{n+1}(0))\varepsilon^{n+1},$$

which serve to determine initial values for $U_n(t)$ and $\dot V_n(\tau)$, resp., when we impose (2.2).

The formal approximation u_N is now constructed by solving the following set of initial value problems. For the terms of the outer expansion, we have problems (P_n), $0\leqq n\leqq N$, given by

$$U_n'(t)+AU_n(t)=\begin{cases}F(t), & n=0,\\ \theta, & n=1,\\ -U_{n-2}''(t), & 2\leqq n\leqq N,\end{cases}$$

$$U_n(0)=\begin{cases}u_0, & n=0,\\ -V_{n-1}(0), & 1\leqq n\leqq N,\end{cases}$$

while the terms of the initial layer expansion are determined as solutions of the problems $(\tilde P_n)$, $0\leqq n\leqq N+1$,

$$\ddot V_n(\tau)+\dot V_n(\tau)=\begin{cases}\theta, & n=0, 1,\\ -AV_{n-2}(\tau), & 2\leqq n\leqq N+1,\end{cases}$$

$$\dot V_n(0)=\begin{cases}u_1, & n=0,\\ -U_{n-1}'(0), & 1\leqq n\leqq N+1,\end{cases}$$

$$V_n(\tau)\to\theta \text{ as } \tau\to\infty,\qquad 0\leqq n\leqq N+1.$$

It follows that the remainder

$$z(t;\ \varepsilon):=u(t;\ \varepsilon)-u_N(t;\ \varepsilon)$$

satisfies the second-order initial-value problem

$$\varepsilon^2 z''(t)+z'(t)+Az(t)=-R(t;\ \varepsilon), \tag{2.10}$$

$$z(0;\ \varepsilon)=z_0(\varepsilon),\ \varepsilon z'(0;\ \varepsilon)=Z_1(\varepsilon), \tag{2.11}$$

where

$$z_0(\varepsilon):=-V_N(0)\varepsilon^{N+1}-V_{N+1}(0)\varepsilon^{N+2},$$
$$z_1(\varepsilon):=\theta$$

and $R(t;\ \varepsilon)$ is given by (2.7). See [7] for proof of existence and uniqueness of a solution to this problem.

For the solution of problems (P_n), $0\leqq n\leqq N$, and $(\tilde{P}_n)$, $0\leqq n\leqq N+1$, we have the following result.

LEMMA 2.1. *Let N be any nonnegative integer and $t_0>0$. Let A be a positive-definite self-adjoint linear operator in H. Suppose $u_0\in D(A^{N+2})$ and $u_1\in D(A^{N+1})$. Then, provided*

$$F^{(l)}\in C([0,\ t_0],\ D(A^{N-2l})),\qquad 0\leqq l\leqq[N/2],$$

where [ ] denotes the greatest integer function,

(A) *The solution of problem (P_n), $0\leqq n\leqq N$, is given by*

$$U_n(t)=\begin{cases} T(t)P(tA;\ n/2)A^{n/2}u_0+T(t)\sum\limits_{l=0}^{n/2-1}P(tA;\ n/2-l-1)A^{n/2-l-1}F^{(l)}(0) \\ \quad+\displaystyle\int_0^t T(t-s)\sum_{l=0}^{n/2}P((t-s)A;\ n/2-l)A^{n/2-l}F^{(l)}(s)\,ds \quad n \text{ even}, \\ T(t)P(tA;\ (n-1)/2)A^{(n-1)/2}u_1 \quad n \text{ odd}, \end{cases}$$

where $T(t)$ is the analytic semigroup generated by $-A$, and $P(tA;\ k)$ is understood to be a generic polynomial operator in tA of degree k.

(B) *The solution of problem $(\tilde{P}_n)$, $0\leqq n\leqq N+1$, is given by*

$$V_n(\tau)=\begin{cases} e^{-\tau}P(\tau;\ n/2)A^{n/2}u_1 \quad n \text{ even}, \\ e^{-\tau}P(\tau;\ (n-1)/2)A^{(n+1)/2}u_0 \\ \quad+e^{-\tau}\sum\limits_{l=0}^{(n-1)/2}P(\tau;\ (n-1)/2-l)A^{(n-1)/2-l}F^{(l)}(0) \quad n \text{ odd}, \end{cases}$$

where $P(\tau;\ k)$ is a generic polynomial in τ of degree k.

Proof. An induction proof is based on the variation of parameters representation for $U_{n+1}(t)$ using $T(t)$, the analytic semigroup generated by $-A$,

$$U_{n+1}(t)=-T(t)V_n(0)+\int_0^t T(t-s)\{-U''_{n-1}(s)\}\,ds,$$

and a representation for $V_{N+1}(\tau)$ obtained by two Riemann integrations of $(\tilde{P}_{n+1})$, namely,

$$V_{n+1}(\tau)=e^{-\tau}U_n'(0)+\int_0^\tau e^{-(\tau-s)}AV_{n-1}(s)\,ds+\int_\tau^\infty AV_{n-1}(s)\,ds.$$

Remarks. (1) We note that in the above procedure, the zeroth order terms U_0 and V_0 are determined simultaneously from the initial data for problem (P_ε^*), but the higher order terms are determined alternately, as indicated in the diagram.

$$U_0\to V_1\to U_2\to V_3\to U_4\to\cdots,$$
$$V_0\to U_1\to V_2\to U_3\to V_4\to\cdots.$$

In particular, if $u_0=\theta$, the first chain of dependent functions vanishes, while if $u_1=\theta$, the second chain vanishes.

(2) The solution $u(t;\varepsilon)$ does not exhibit initial layer behavior since the first term U_0 of the outer expansion provides a uniformly $O(\varepsilon)$ approximation on $[0, t_0]$, i.e.,

$$u(t;\varepsilon)=U_0(t)+O(\varepsilon).$$

However, for the t-derivative

$$\varepsilon u_t(t;\varepsilon)=\dot{V}_0(\tau)+(U_0'(t)+\dot{V}_1(\tau))\varepsilon+O(\varepsilon^2),$$

so we see that $U_0'(t)$ provides a uniformly $O(\varepsilon)$ approximation in any interval $[\delta, t_0]$, $\delta>0$ where $\dot{V}_0$ and $\dot{V}_1$ are asymptotically zero, but close to the initial point the boundary layer functions V_0 and V_1 are required for a uniform approximation.

The justification of the formal expansion procedure requires that the remainder z be $O(\varepsilon^{N+1})$ on $[0, t_0]$ as $\varepsilon\to 0^+$. A crucial step is the development of an a priori bound for the solution of the second-order initial value problem

$$\varepsilon^2 z''+z'+Az=f(t;\varepsilon), \tag{2.12}$$

$$z(0)=z_0(\varepsilon),\qquad \varepsilon z'(0)=z_1(\varepsilon). \tag{2.13}$$

We formulate the result as follows.

LEMMA 2.2. *Let A be a positive-definite self-adjoint linear operator in H and suppose $z_0(\varepsilon)\in D(A)$, $z_1(\varepsilon)\in D(A^{1/2})$ and $f(\cdot;\varepsilon)\in L^2(0, t_0; H)$, where $0<\varepsilon\leqq\varepsilon_0$ for an arbitrary fixed number ε_0* in $(0, 1)$.

Then the solution $z(t;\varepsilon)$ of (2.12), (2.13) *on* $[0, t_0]$ *satisfies the inequality*

$$\begin{aligned}&\|z(t)\|+\varepsilon\|z'(t)\|+\|A^{1/2}z(t)\|\\&\leqq M\left\{\|z_0\|+\|z_1\|+\|A^{1/2}z_0\|+\left(\int_0^{t_0}\|f(s)\|^2\,ds\right)^{1/2}\right\},\end{aligned}\tag{2.14}$$

in which we have suppressed the ε-dependence of z, z_0, z_1 and f. The constant M is given explicitly by $M=[6(1-\varepsilon_0)^{-1}]^{1/2}\exp[(1-\varepsilon_0)^{-1}t_0]$.

Proof. Forming the inner product

$$(2z+2z', L_\varepsilon[z]-f)=0$$

and transferring derivatives to form perfect t-derivatives, we see that, by introduction of $A^{1/2}$ and use of the positive-definiteness of A, we obtain the differential inequality

$$\begin{aligned}&D_t\{\|z\|^2+2\varepsilon^2(z, z')+\varepsilon^2\|z'\|^2+\|A^{1/2}z\|^2\}\\&\leqq\|z\|^2+2\varepsilon^2\|z'\|^2+2\|f\|^2.\end{aligned}$$

Integrating over the interval $[0, t]$, and using the inequality

$$2\varepsilon^2(z, z')\geqq -\varepsilon\|z\|^2-\varepsilon^3\|z'\|^2$$

which holds for $0<\varepsilon<\varepsilon_0$, where ε_0 lies in $(0, 1)$, we arrive at the inequality

$$\begin{aligned}&\|z(t)\|^2+\varepsilon^2\|z'(t)\|^2+\|A^{1/2}z(t)\|^2\\ &\quad\leqq 2(1-\varepsilon_0)^{-1}\left\{(\|z_0\|^2+\|z_1\|^2+\|A^{1/2}z_0\|^2)+\int_0^{t_0}\|f(s)\|^2\,ds\right\}\\ &\qquad+2(1-\varepsilon_0)^{-1}\int_0^t(\|z(s)\|^2+\varepsilon^2\|z'(s)\|^2+\|A^{1/2}z(s)\|^2)\,ds.\end{aligned}$$

Application of Gronwall's lemma and elementary inequalities produces the result of the lemma.

We can now state the main result of this section.

THEOREM 2.1. *Let a nonnegative integer N and $t_0>0$ be given. Let A be a positive-definite self-adjoint linear operator in a Hilbert space H and suppose that $u_0\in D(A^{N+2})$ and $u_1\in D(A^{N+1})$. Assume $F=F(t)$ is such that*

$$\frac{d^lF}{dt^l}\in C([0, t_0], D(A^{N-2l})),\qquad 0\leqq l\leqq[N/2].$$

Then the solution $u(t; \varepsilon)$ of the initial-value problem (P^*_ε) has an asymptotic expansion of the form (2.3) *such that*

$$u(t;\varepsilon)=u_N(t;\varepsilon)+O(\varepsilon^{N+1}),$$
$$\varepsilon u'(t;\varepsilon)=\varepsilon u'_N(t;\varepsilon)+O(\varepsilon^{N+1}),$$

and

$$A^{1/2}u(t;\varepsilon)=A^{1/2}u_N(t;\varepsilon)+O(\varepsilon^{N+1})$$

as $\varepsilon\to 0^+$ uniformly on $[0, t_0]$.

Proof. The remainder $z(t;\varepsilon)=u(t;\varepsilon)-u_N(t;\varepsilon)$ satisfies (2.10) and (2.11). Since Lemma 2.1 implies

$$\|V_{N+1}(0)\|\leqq\begin{cases}C(\|A^{(N+2)/2}u_0\|;\ \|A^{N/2-l}F^{(l)}(0)\|, 0\leqq l\leqq N/2) & N\text{ even},\\ C(\|A^{(N+1)/2}u_1\|) & N\text{ odd},\end{cases}$$

we see that $\|z_0(\varepsilon)\|<C\varepsilon^{N+1}$. Obviously $\|z_1(\varepsilon)\|=0$.

In order to bound $\|R(t;\varepsilon)\|$, we first note that since $\varepsilon^{-\tau}P(\tau)$ is bounded on $[0,\infty)$ for any polynomial $P(\tau)$, we have from the representations in Lemma 2.1,

$$\|AV_{N+1}(\tau)\|\leqq\begin{cases}C(\|A^{(N+4)/2}u_0\|,\ \|A^{(N+2)/2-l}F^{(l)}(0)\|, 0\leqq l\leqq N/2) & N\text{ even},\\ C(\|A^{(N+3)/2}u_1\|) & N\text{ odd}.\end{cases}$$

Also, the proof of Lemma 2.1 yields an expression for $U''_n(t)$ so that for N odd,

$$\|U''_N(t)\|\leqq C(t_0;\ \|A^{N+1}u_1\|)$$

and for N even, with

$$\|G\|_\infty:=\sup_{0\leqq t\leqq t_0}\|G(t)\|,$$

$$\begin{aligned}\|U''_N(t)\|\leqq C(t_0;\ &\|A^{N+2}u_0\|;\ \|A^{N-2l}F^{(l)}(0)\|, 0\leqq l\leqq(N-2)/2\\ &\|A^{N/2-l}F^{(l+1)}(\cdot)\|_\infty,\ \|A^{N/2-l+1}F^{(l)}(\cdot)\|_\infty,\\ &\|A^{N-2l+2}F^{(l)}(\cdot)\|_\infty, 0\leqq l\leqq N/2).\end{aligned}$$

From these estimates

$$\|R(t;\varepsilon)\| \leqq C(t_0)\varepsilon^{N+1}.$$

Putting these results in the energy inequality (2.14) we obtain

$$\|z(t)\| + \varepsilon\|z'(t)\| + \|A^{1/2}z(t)\| \leqq C(t_0)\varepsilon^{N+1},$$

which establishes the theorem.

Remark. The results of this section hold for any positive-definite self-adjoint linear operator A defined on a Hilbert space H, and are not restricted to the operator defined in (1.5). In particular, if A is of arbitrary order $2m$, the plate problem is included.

3. A numerical procedure. Applying the results of § 2 to the solution $u(x, t)$ of problem (P_ε), given by (1.1) to (1.4), we have the asymptotic expansion

$$u(x,t) = U_0(x,t) + \sum_{n=1}^{N} (U_n(x,t) + V_{n-1}(x, t/\varepsilon^2))\varepsilon^n + O(\varepsilon^{N+1}).$$

Here $U_n(x, t)$ is obtained as the solution of a parabolic problem (P_n):

$$\frac{\partial U_n}{\partial t}(x,t) + AU_n(x,t) = \begin{cases} f(x,t), & n=0, \\ 0, & n=1, \\ -\partial^2 U_{n-2}/\partial t^2, & 2 \leqq n \leqq N \end{cases}$$

$$\text{for } (x,t) \in \Omega \times (0, t_0],$$

$$U_n(x,t) = 0, \quad 0 \leqq n \leqq N, \quad (x,t) \in \partial\Omega \times (0, t_0],$$

$$U_n(x,0) = \begin{cases} u_0(x), & n=0, & x \in \bar{\Omega}, \\ -V_{n-1}(x,0), & 1 \leqq n \leqq N, & x \in \bar{\Omega}, \end{cases}$$

where the elliptic operator A is given by (1.5). To determine the initial-layer correction terms we must solve $(\tilde{P}_n)$:

$$\frac{\partial^2 V_n}{\partial \tau^2}(x,\tau) + \frac{\partial V_n(x,\tau)}{\partial \tau}$$

$$= \begin{cases} 0, & n = 0, 1 \\ -AV_{n-2}(x,\tau), & 2 \leqq n \leqq N-1 \end{cases} \quad \text{for } (x,\tau) \in \Omega \times (0, \infty),$$

$$\frac{\partial V_n}{\partial \tau}(x,0) = \begin{cases} u_1(x), & n=0, & x \in \bar{\Omega}, \\ (-\partial U_{n-1}/\partial t)(x,0), & 1 \leqq n \leqq N-1, & x \in \bar{\Omega}, \end{cases}$$

$$V_n(x,\tau) \to 0 \quad \text{as } \tau \to \infty, \qquad 0 \leqq n \leqq N-1, \quad x \in \bar{\Omega}.$$

In order to develop a numerical procedure based on the Galerkin method, we introduce weak formulations of the problems (P_ε) and (P_n), $0 \leqq n \leqq N$. As usual, we denote by $H^m(\Omega)$ the real Sobolev space of integer order m on Ω equipped with the norm $\|\cdot\|_m$, and by $H_0^m(\Omega)$ the subspace of $H^m(\Omega)$ obtained by completing $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_m$.

By a weak solution of problem (P_ε) we mean a function $u(x, t)$, twice continuously differentiable with respect to t, such that for each fixed t in $[0, t_0]$, $u(t) := u(\cdot, t) \in H_0^1(\Omega)$ and satisfies the integral identities

$$\begin{aligned} &\varepsilon^2(u_{tt}, w) + (u_t, w) + a(u, w) = (f, w), \\ &(u(0) - u_0, w) = 0, \qquad (\varepsilon u_t(0) - u_1, w) = 0 \end{aligned} \tag{3.1}$$

for each $w \in H_0^1(\Omega)$. Here $(\cdot,\cdot)$ denotes the $L_2(\Omega)$ inner product and $a(\cdot,\cdot)$ is the bilinear form

$$a(u, w) = \int_\Omega \left\{ \sum_{i,j=1}^{n} a_{ij}(x) D_j u D_j w + c(x) uw \right\} dx$$

associated with the linear operator A defined in (1.5). Under the assumptions on the coefficients a_{ij} and c given in § 1, the bilinear form $a(\cdot,\cdot)$ is continuous as a map from $H^1(\Omega)\times H^1(\Omega)\to\mathbb{R}$, i.e., there exists a constant $\gamma > 0$ such that

$$a(u, w) \leqq \gamma \|u\|_1 \|w\|_1 \quad \text{for all } u, \quad w \in H^1(\Omega), \tag{3.2}$$

and strongly coercive, i.e., there exists a constant $\lambda > 0$ such that

$$a(u, u) \geqq \lambda \|u\|^2 \quad \text{for all } u \in H_0^1(\Omega). \tag{3.3}$$

These properties are important for the error bounds of our numerical approximation discussed in § 4.

A weak solution of problem (P_n), which we shall also denote $U_n(x, t)$, $0 \leqq n \leqq N$, is a continuously t-differentiable function $U_n(t) := U_n(\cdot, t) \in H_0^1(\Omega)$, for each t in $[0, t_0]$, which satisfies

$$\left(\frac{\partial U_n}{\partial t}, w\right) + a(U_n, w) = \begin{cases} (f, w), & n = 0, \\ 0, & n = 1, \\ (-\partial^2 U_{n-2}/\partial t^2, w), & 2 \leqq n \leqq N, \end{cases}$$

$$(U_n(0) - u_n, w) = 0, \qquad n = 0, 1,$$

$$(U_n(0) + V_{n-1}(0), w) = 0, \qquad 2 \leqq n \leqq N,$$

for each $w \in H_0^1(\Omega)$.

The existence and uniqueness of the weak solutions for (P_ε) and (P_n), $0 \leqq n \leqq N$, follow from standard results for linear hyperbolic and parabolic equations (see, for example, [14]). We note that the terms of the expansion are obtained in a recursive manner so that the problems (P_n), $0 \leqq n \leqq N$, are not solved simultaneously, but successively, together with the layer problems $(\tilde{P}_n)$, $0 \leqq n \leqq N-1$; see the remark following Lemma 2.1.

To describe the asymptotic Galerkin approximation, let us denote by S^h a one-parameter family of $M(h)$-dimensional subspaces of $H_0^1(\Omega)$ possessing a certain interpolation property to be specified later (see (4.1)). Let $\{w_k^h\}_{k=1}^M$ be a basis for S^h. The asymptotic Galerkin approximation is defined by

$$u_*^h(x, t) = U_0^h(x, t) + \sum_{n=1}^{N} (U_n^h(x, t) + V_{n-1}^h(x, t/\varepsilon^2))\varepsilon^n, \tag{3.4}$$

in which $U_n^h(\cdot, t) \in S^h$ is the continuous-in-time Galerkin approximation of the weak solution $U_n(x, t)$, which satisfies

$$\left(\frac{\partial U_n^h}{\partial t}, w^h\right) + a(U_n^h, w^h) = \begin{cases} (f(\cdot, t), w^h), & n = 0, \\ 0, & n = 1, \\ (-\partial^2 U_{n-2}^h/\partial t^2, w^h), & 2 \leqq n \leqq N, \end{cases}$$

$$(U_n^h(\cdot, 0) - u_n, w^h) = 0, \qquad n = 0, 1,$$

$$(U_n^h(\cdot, 0) + V_{n-1}^h(\cdot, 0), w^h) = 0, \qquad 2 \leqq n \leqq N,$$

for each $w^h \in S^h$.

The functions V_n^h, $0 \leqq n \leqq N-1$, are the L_2-projections of the layer functions $V_n(x, \tau)$ so that

$$(V_n(\cdot, \tau), w^h) = (V_n^h(\quad \cdot, \tau), w^h)$$

for all $w^h \in S^h$.

In terms of the basis $\{w_k^h\}_{k=1}^M$ for S^h, the outer approximations have the explicit representations

$$U_n^h(x, t) = \sum_{k=1}^M \alpha_k^n(t) w_k^h(x).$$

Here the functions $\{\alpha_k^n\}_{k=1}^M$, $n = 0, 1, \cdots, N$, are determined as solutions of an initial-value problem for a system of ordinary differential equations

$$\sum_{k=1}^M \left\{\frac{d}{dt}\alpha_k^n(t)(w_k^h, w_l^h) + \alpha_k^n(t) a(w_k^h, w_l^h)\right\}$$

$$= \begin{cases} (f(\cdot, t), w_l^h), & n = 0, \\ 0, & n = 1, \\ \displaystyle\sum_{k=1}^M \frac{d^2}{dt^2}\alpha_k^{n-2}(t)(w_k^h, w_l^h), & 2 \leqq n \leqq N, \end{cases} \tag{3.5}$$

$$\sum_{k=1}^M \alpha_k^n(0)(w_k^h, w_l^h) = \begin{cases} (u_n, w_l^h), & n = 0, 1, \\ (-V_{n-1}(0), w_l^h), & 2 \leqq n \leqq N \end{cases}$$

for $0 \leqq l \leqq M$.

The system is uniquely solvable for each $n = 0, 1, \cdots, N$, since the Gram matrix $G = [(w_k^h, w_l^h)]$ is positive definite. Let $S = [a(w_k^h, w_l^h)]$.

The solution vectors

$$\boldsymbol{\alpha}^n(t) = (\alpha_1^n, (t), \cdots, \alpha_M^n(t))^T,$$

$n = 0, 1, \cdots, N$, are given by the explicit representation

$$\boldsymbol{\alpha}^n(t) = \begin{cases} \exp\{-tG^{-1}S\}G^{-1}\mathbf{u}_0 + \displaystyle\int_0^t \exp\{-(t-t_1)G^{-1}S\}\mathbf{f}(t_1)\,dt_1, & n = 0, \\ \exp\{-tG^{-1}S\}G^{-1}\mathbf{u}_1, & n = 1, \\ \exp\{-tG^{-1}S\}G^{-1}\mathbf{V}_n(0) + \displaystyle\int_0^t \exp\{-(t-t_1)G^{-1}S\}\frac{d^2\boldsymbol{\alpha}^{n-2}(t_1)}{dt^2}\,dt_1, & 2 \leqq n \leqq N, \end{cases}$$

with

$$\mathbf{f}(t) = ((f(\cdot, t), w_1^h), \cdots, (f(\cdot, t), w_M^h))^T,$$

$$\mathbf{u}_n = ((u_n, w_1^h), \cdots, (u_n, w_M^h))^T, \qquad n = 0, 1$$

and

$$\mathbf{V}_n(0) = ((-V_{n-1}(\cdot, 0), w_1^h), \cdots, (-V_{n-1}(\cdot, 0), w_M^h))^T, \qquad 2 \leqq n \leqq N.$$

The asymptotic Galerkin approximation defined in (3.4) admits the explicit representation

$$u_*^h(x, t) = \sum_{n=0}^N \sum_{k=1}^M \beta_k^n(t) w_k^h(x) \varepsilon^n, \tag{3.6}$$

where

$$\beta_k^0(t) = \boldsymbol{\alpha}^0(t) \cdot \mathbf{e}_k \tag{3.7}$$

and

$$\beta_k^n(t) = \{\boldsymbol{\alpha}^n(t) - \mathbf{V}_{n-1}(\tau)\} \cdot \mathbf{e}_k, \qquad 1 \leqq n \leqq N,$$

in which $\mathbf{e}_k$ is the kth unit vector in $\mathbb{R}^M$, and

$$V_{n-1}(\tau) = ((V_{n-1}(\cdot, \tau), w_1^h), \cdots, (V_{n-1}(\cdot, \tau), w_M^h))^T$$

for $n = 1, \cdots, N$.

We comment that if we solve (3.1) directly by the Galerkin method and approximate the weak solution u by the Galerkin approximation

$$g^h(x, t) = \sum_{k=1}^{M} \gamma_k(t) w_k^h(x),$$

then $\boldsymbol{\gamma}(t) = (\gamma_1(t), \cdots, \gamma_M(t))^T$ satisfies the singularly perturbed second-order ordinary differential equation with initial data:

$$\varepsilon^2 G\boldsymbol{\gamma}''(t) + G\boldsymbol{\gamma}'(t) + S\boldsymbol{\gamma}(t) = \mathbf{f}(t), \qquad 0 < t \leqq t_0$$

$$\boldsymbol{\gamma}(0) = G^{-1}((u_0, w_1^h), \cdots, (u_0, w_M^h))^T,$$

$$\varepsilon\boldsymbol{\gamma}'(0) = G^{-1}((u_1, w_1^h), \cdots, (u_1, w_M^h))^T.$$

The vectors $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^1$ defined in (3.7) are the zeroth order term and the $O(\varepsilon)$-term in the expansion of $\boldsymbol{\gamma}$. Hence the asymptotic Galerkin approximation is a combination of asymptotic and numerical approximations.

The coefficients $\beta_k^n(t)$, $1 \leqq k \leqq M$, defined by (3.7) are computable in principle. The exponential matrix is best handled in practice by difference approximations or Padé approximations. The implementation of the procedure using discrete schemes follows along well developed lines. Preliminary computations indicate that the procedure is very promising.

4. Error estimates. In this section we derive error estimates for the asymptotic Galerkin approximation u_*^h, the continuous-time Galerkin procedure developed in § 3. Since the accuracy of the procedure depends on the properties of the approximating subspaces, we follow [3] in assuming that for a fixed integer $m \geqq 2$, the finite-dimensional subspaces S^h of $H_0^1(\Omega)$ possess the following approximation property: For any $u \in H^s(\Omega) \cap H_0^1(\Omega)$, $1 \leqq s \leqq m$, there is a constant Q, independent of h and u, such that

$$\inf \{\|u - V^h\| + h\|u - V^h\|_1 \colon V^h \in S^h\} \leqq Qh^s\|u\|_s. \tag{4.1}$$

The space of piecewise linear polynomials is known to satisfy (4.1) for $s = 2$.

We introduce the elliptic projection of u onto S^h, denoted e^h, defined as the solution of the variational elliptic boundary value problem

$$a(e^h(t), w^h) = a(u(t), w^h) \tag{4.2}$$

for all $w^h \in S^h$ and for each fixed t in $[0, t_0]$, treated as a steady-state problem with parameter t. The existence of e^h is guaranteed by the Lax-Milgram theorem. The elliptic projection is the best approximation to u in S^h with respect to the H_1-norm, i.e., for some constant ν

$$\|u - e^h\|_1 \leqq \nu\|u - w^h\|_1 \quad \text{for all } w^h \in S^h.$$

Moreover, it is easily shown that if for some constant $k \geqq 0$, $u \in C^k([0, t_0]; H^s(\Omega) \cap H_0^1(\Omega))$, then for $1 \leqq s \leqq m$, and for some constant C, independent of h and u,

$$\|D_t(u - e^h)\| \leqq Ch^s \|D_t^k u\|_s. \tag{4.3}$$

We also introduce the continuous-time Galerkin solution of (3.1), $g^h \in C^2((0, t_0]; S^h)$, satisfying

$$\begin{gathered} \varepsilon^2(g_{tt}^h, w^h) + (g_t^h, w^h) + a(g^h, w^h) = (f, w^h), \\ (g^h(\cdot, 0) - u_0(\cdot), w^h) = 0, \\ (\varepsilon g_t^h(\cdot, 0) - u_1(\cdot), w^h) = 0 \end{gathered} \tag{4.4}$$

for $0 < t \leqq t_0$ and for all $w^h \in S^h$.

The elliptic projection e^h and the Galerkin solution g^h are used only to derive the desired error estimates and do not appear in the numerical procedure (see [3], [8] and [20]).

If $\underline{X}$ is a Hilbert space with norm $\|\cdot\|_{\underline{X}}$, then

$$\|u\|_{L^\infty(\underline{X})} := \sup_{t \in [0, t_0]} \|u(\cdot, t)\|_{\underline{X}}$$

and

$$\|u\|_{L^2(\underline{X})} := \left(\int_0^{t_0} \|u(\cdot, t)\|_{\underline{X}}^2 \, dt \right)^{1/2}.$$

We obtain error estimates with respect to the following norm

$$\|u\|_E := \|u\|_{L^\infty(H^1)} + \varepsilon \|u_t\|_{L^\infty(L^2)}.$$

THEOREM 4.1. *Let $\{S^h\}$, $0 < h \leqq 1$, be a family of finite-dimensional subspaces of $H_0^1(\Omega)$ satisfying the approximation property* (4.1). *Denote by*

u, *the weak solution of problem* (P_ε),

g^h, *the Galerkin approximation of u in S^h,*

e^h, *the elliptic projection of u into S^h, and*

u_*^h, *the asymptotic Galerkin approximation.*

Assume that $u_0 \in D(A^{N+2})$ and $u_1 \in D(A^{N+1})$. If $u_{tt} \in L^2(0, t_0; H^s(\Omega))$, and

$$\|e^h(\cdot, 0) - g^h(\cdot, 0)\|_1 \leqq Ch^s, \tag{4.5}$$

$$\|e_t^h(\cdot, 0) - g_t^h(\cdot, 0)\| \leqq Ch^s, \tag{4.6}$$

then the bound

$$\|u - u_*^h\|_E \leqq C\{h^{s+1} + \varepsilon^{N+1}\},$$

holds for $0 < t \leqq t_0$, where the constant C is independent of ε, h and u.

Proof. To measure $\|u - u_*^h\|_E$ we introduce the elliptic projection e^h and the Galerkin solution g^h,

$$u - u_*^h = (u - e^h) + (e^h - g^h) + (g^h - u_*^h),$$

and proceed by bounding each term on the right-hand side. The appropriate estimates are

(i) $\|u-e^h\|_E \leqq Ch^{s-1}\{\|u\|_{L^\infty(H^s)}+\varepsilon h\|u_t\|_{L^\infty(H^s)}\}$,
(ii) $\|e^h-g^h\|_E \leqq Ch^s\{1+\varepsilon+\varepsilon\|u_t\|_{L^\infty(H^s)}+\varepsilon^2\|u_{tt}\|_{L^2(H^s)}\}$,
(iii) $\|g^h-u_*^h\|_E \leqq C\varepsilon^{N+1}$.

By the best approximation property of e^h and the property of convergence of the family of subspaces S^h, we see that if $u \in H^s(\Omega) \cap H_0^1(\Omega)$, then

$$\|u-e^h\|_1 \leqq Ch^{s-1}\|u\|_s,$$

where $C=Q\nu$. Applying this argument to $u_t \in H^s(\Omega) \cap H_0^1(\Omega)$ together with Nitsche's trick (see [18]), we get

$$\|u_t-e_t^h\| \leqq Ch^s\|u_t\|_s.$$

The inequality (i) follows immediately.

For the proof of (ii), let $\psi^h := g^h - e^h$ and $\eta := u-e^h$, then we easily find, using (3.1), (4.2) and (4.4), that for each $w^h \in S^h$,

$$\varepsilon^2(\psi_{tt}^h, w^h)+(\psi_t^h, w^h)+a(\psi^h, w^h)=\varepsilon^2(\eta_{tt}, w^h)+(\eta_t, w^h).$$

We combine the first pair of terms on each side as follows

$$\varepsilon^2(e^{-t/\varepsilon^2}D_t(\psi_t^h e^{t/\varepsilon^2}), w^h)+a(\eta, w^h)=\varepsilon^2(e^{-t/\varepsilon^2}D_t(\eta_t e^{t/\varepsilon^2}), w^h).$$

For each fixed t in $(0, t_0]$ we choose the particular element of S^h given by $w^h(t)=\psi_t^h(t)e^{t/\varepsilon^2}$, so that we obtain

$$\tfrac{1}{2}\varepsilon^2 D_t\|\psi_t^h e^{t/\varepsilon^2}\|^2+\tfrac{1}{2}e^{2t/\varepsilon^2}D_t a(\psi^h, \psi^h)=\varepsilon^2(D_t(\eta_t e^{t/\varepsilon^2}), \psi_t^h e^{t/\varepsilon^2}).$$

Integration from $t=0$ to $t=\xi$ yields, after integration by parts and use of the coercivity and continuity of the bilinear form $a(\cdot\,.\,\cdot)$, specified in (3.2) and (3.3), the inequality

$$\begin{aligned}\lambda\|\psi^h(\xi)\|_1^2+\varepsilon^2\|\psi_t^h(\xi)\|^2 &\leqq e^{-2\xi/\varepsilon^2}(\gamma\|\psi^h(0)\|_1^2+\varepsilon^2\|\psi_t^h(0)\|^2)\\ &\quad+2\gamma/\varepsilon^2\int_0^\xi e^{-2(\xi-t)/\varepsilon^2}\|\psi(t)\|_1^2\,dt\\ &\quad+2\varepsilon^2\int_0^\xi e^{-(2\xi-t)/\varepsilon^2}(D_t(\eta_t e^{t/\varepsilon^2}), \psi_t^h)\,dt.\end{aligned}\tag{4.7}$$

Consider that

$$\begin{aligned}&2\varepsilon^2\int_0^\xi e^{-(2\xi-t)/\varepsilon^2}(D_t(\eta_t e^{t/\varepsilon^2}), \psi_t^h)\,dt\\ &\quad\leqq \varepsilon^4\int_0^\xi e^{-2(\xi-t)/\varepsilon^2}\|\eta_{tt}\|^2\,dt+2/\varepsilon^2\int_0^\xi e^{-2(\xi-t)/\varepsilon^2}\varepsilon^2\|\psi_t^h\|^2\,dt\\ &\quad\quad+\int_0^\xi e^{-2(\xi-t)/\varepsilon^2}\|\eta_t\|^2\,dt.\end{aligned}$$

Thus (4.7) becomes

$$\begin{aligned}&\|\psi^h(\xi)\|_1^2+\varepsilon^2\|\psi_t^h(\xi)\|^2\\ &\quad\leqq C_1\{\|\psi^h(0)\|_1^2+\varepsilon\|\psi_t^h(0)\|^2+\varepsilon^4\|\eta_{tt}\|_{L^2(L^2(\Omega))}^2+\varepsilon^2\|\eta_t\|_{L^\infty(L^2(\Omega))}^2\}\\ &\quad\quad+C_2\varepsilon^{-2}\int_0^\xi e^{-2(\xi-t)/\varepsilon^2}\{\|\psi^h(t)\|_1^2+\varepsilon^2\|\psi_t^h(t)\|^2\}\,dt.\end{aligned}$$

By Gronwall's lemma and elementary inequalities, the following is seen to hold:

$$\|\psi^h(t)\|_1+\varepsilon\|\psi_t^h(t)\|\leqq C\{\|\psi^h(0)\|_1+\varepsilon\|\psi_t^h(0)\| \\ +\varepsilon\|\eta_{tt}\|_{L^\infty(L^2(\Omega))}+\varepsilon^2\|\eta_{tt}\|_{L^2(L^2(\Omega))}\}.$$

From (4.3) and the hypotheses (4.5), (4.6),

$$\|e^h-g^h\|_E\leqq Ch^s\{1+\varepsilon+\varepsilon\|u_t\|_{L\infty(H^s)}+\varepsilon^2\|u_{tt}\|_{L^2(H^2)}\}$$

and (ii) is established.

Applying the inequality of Lemma 2.2 to $z(t)=g^h(t)-u_*^h(t)$, we find

$$\|A^{1/2}(g^h-u_*^h)\|+\varepsilon\|(g^h-u_*^h)_t\|\leq C\varepsilon^{N+1},$$

where $\|\cdot\|$ denotes the L_2-norm. From [6] we know that the norm $\|A^{1/2}\cdot\|$ is equivalent to the norm $\|\cdot\|_1$ so that

$$\|g^h(t)-u_*^h(t)\|_1+\varepsilon\|g_t^h(t)-u_*^h(t)\|\leqq C\varepsilon^{N+1}$$

and

$$\|g^h-u_*^h\|_E\leqq C\varepsilon^{N+1}.$$

Thus (iii) is seen to hold.

5. Numerical experiments. Here we present some numerical evidence that the method provides good numerical approximations to the solution of the singularly perturbed problem. We consider the following model problem (P_ε):

$$\varepsilon^2u_{tt}+u_t-u_{xx}=0, \qquad 0<x<1,\ t>0,$$

$$u(x,0)=u_0(x), \qquad \varepsilon u_t(x,0)=u_1(x), \quad 0<x<1,$$

$$u(0,t)=u(1,t)=0, \qquad t>0,$$

with $u_0(x)=u_1(x)=\sin\pi x$. The exact solution of (P_ε) can be found explicitly: when $0<\varepsilon<\frac{1}{2}\pi$, $u(x,t;\varepsilon)=\{A(\varepsilon)\,e^{r_1(\varepsilon)t}+B(\varepsilon)\,e^{r_2(\varepsilon)t}\}\sin\pi x$ where

$$r_1(\varepsilon)=\frac{-1+\sqrt{1-4\pi^2\varepsilon^2}}{2\varepsilon^2}, \qquad r_2(\varepsilon)=\frac{-1-\sqrt{1-4\pi^2\varepsilon^2}}{2\varepsilon^2},$$

$$A(\varepsilon)=\frac{2\varepsilon+1+\sqrt{1-4\pi^2\varepsilon^2}}{2\sqrt{1-4\pi^2\varepsilon^2}}, \qquad B(\varepsilon)=\frac{\sqrt{1-4\pi^2\varepsilon^2}-1-2\varepsilon}{2\sqrt{1-4\pi^2\varepsilon^2}},$$

and, when $\varepsilon>\frac{1}{2}\pi$,

$$u(x,t;\varepsilon)=e^{-t/2\varepsilon^2}\{\cos r(\varepsilon)t+C(\varepsilon)\sin r(\varepsilon)t\}\sin\pi x$$

where

$$r(\varepsilon)=\frac{\sqrt{4\pi^2\varepsilon^2-1}}{2\varepsilon^2} \quad\text{and}\quad C(\varepsilon)=\frac{1+2\varepsilon}{\sqrt{4\pi^2\varepsilon^2-1}}.$$

Attacking the hyperbolic problem directly we first write it as a system (see Baker [1]):

$$u_t = v, \qquad \varepsilon^2 v_t = -v + u_{xx}.$$

With $U^n = u(\cdot, t_n)$ and $Q^n = v(\cdot, t_n)$, we introduce the scheme

$$U^{n+1} = U^n + \tfrac{1}{2}\tau(Q^{n+1} + Q^n), \tag{5.1}$$

$$\left(\frac{\varepsilon^2}{\tau} + \frac{1}{2}\right)(Q^{n+1}, \chi) + \left(\frac{1}{2} - \frac{\varepsilon^2}{\tau}\right)(Q^n, \chi) + a(U^n, \chi) + \frac{1}{4}\tau a(Q^{n+1} + Q^n, \chi) = 0, \tag{5.2}$$

where τ is the time step and $\chi \in S^h$, the finite-dimensional subspace of piecewise linear polynomials. At each time step the new value of u_t is computed by means of (5.2) and then used to update u using (5.1). To see that this Crank-Nicolson-Galerkin scheme is well suited to the problem (P_ε) when ε is large, we set $\varepsilon = 1$, $\Delta x = \Delta t = .01$ and generated the data in Table 1 at $x = .5$. When $\varepsilon = .01$, however, the approximate time derivative oscillates wildly as shown in Table 2. Since these values are used to update u, we have a very poor approximation. In order to achieve the same accuracy as indicated in Table 1 it was necessary to reduce the time step to $\Delta t = .00001$.

TABLE 1
$\varepsilon = 1$, $\Delta x = \Delta t = .01$, $x = .5$

t	Approx Sol	Exact Sol	Rel Error
.01	1.009455	1.009455	0.0
.05	1.036466	1.036459	6.46×10^{-6}
.1	1.046283	1.046249	3.28×10^{-5}
.5	.392190	.392168	5.66×10^{-5}
1	−.594303	−.594301	3.37×10^{-6}

TABLE 2
$\varepsilon = .01$, $\Delta x = \Delta t = .01$, $x = .5$

t	Computed u_t	Exact u_t	Rel Error
.01	−114.72	−9.04	1.2×10^{1}
.02	93.32	−8.19	1.2×10^{1}
.03	−104.96	−7.42	1.3×10^{1}
.04	86.97	−6.72	1.4×10^{1}
.05	−96.11	−6.09	1.5×10^{1}

TABLE 3
$\Delta x = \Delta t = .01$, $x = .5$, $t = .01$

ε^2	Asymp Sol	Exact Sol	Rel Error
10^{-2}	.95974	1.02616	6.5×10^{-2}
10^{-3}	.93457	.94352	9.4×10^{-3}
10^{-4}	.91499	.91590	9.9×10^{-4}
10^{-5}	.90879	.90896	1.9×10^{-4}
10^{-6}	.90683	.90693	1.1×10^{-4}
10^{-7}	.90621	.90630	9.8×10^{-5}
10^{-8}	.90602	.90610	9.7×10^{-5}

The asymptotic expansion to two terms is

$$u_1(x, t) = U_0(x, t) + \varepsilon(U_1(x, t) + V_0(x_1, t/\varepsilon^2)),$$

where U_0 and U_1 both satisfy the parabolic IBVP

$$U_t - U_{xx} = 0,$$
$$U(x, 0) = \sin \pi x,$$
$$U(0, t) = U(1, t) = 0$$

and $V_0(x, t/\varepsilon^2) = -e^{-t/\varepsilon^2} \sin \pi x$. We have used a standard Crank-Nicolson-Galerkin scheme to generate the solution to the parabolic problem. Table 3 shows that the asymptotic Galerkin approximation is in close agreement with the exact solution well beyond the point at which the Crank-Nicolson-Galerkin scheme has failed. Note that this has been accomplished using the same space-time mesh sizes in both cases.

REFERENCES

[1] G. A. BAKER, *Error estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564-576.

[2] L. BOBISUD, *On the behavior of the solution of the telegraphist's equation for large velocities*, Pacific J. Math., 22:2 (1967), pp. 213-219.

[3] G. FAIRWEATHER, *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, New York, 1978.

[4] H. O. FATTORINI, *Singular perturbation and boundary layer for an abstract Cauchy problem*, J. Math. Anal. Appl., 97 (1983), pp. 529-571.

[5] ———, *Second Order Linear Differential Equations in Banach Spaces*, North-Holland, New York, 1985.

[6] J. A. GOLDSTEIN, *Semigroups and second-order differential equations*, J. Funct. Anal., 4 (1969), pp. 50-70.

[7] E. HEINZ AND W. VON WAHL, *Zu Einem Satz von F. E. Browder über nichtlineare Wellengleichen*, Math. Z., 141 (1975), pp. 33-45.

[8] G. C. HSIAO AND K. E. JORDAN, *A finite element method for singularly perturbed parabolic equations*, in Boundary and Interior Layers—Computation and Asymptotic Methods, J. J. H. Miller, ed., Boole Press, Dublin, 1980, pp. 317-321.

[9] G. C. HSIAO AND R. J. WEINACHT, *A singularly perturbed Cauchy problem*, J. Math. Anal. Appl., 71:1 (1979), pp. 242-250.

[10] ———, *Singular perturbations for a weakly nonlinear hyperbolic equation*, Appl. Anal., 10 (1980), pp. 221-229.

[11] ———, *Singular perturbations for a semi-linear hyperbolic equation*, this Journal, 14 (1983), pp. 1168-1179.

[12] K. E. JORDAN, *A numerical treatment of singularly perturbed boundary and initial-boundary value problems*, Ph.D., dissertation, Univ. of Delaware, 1980.

[13] K. KISYNSKI, *Sur les equations hyperboliques avec petit parametre*, Colloq. Math., 10 (1963), pp. 331-343.

[14] V. P. MIKHAILOV, *Partial Differential Equations*, MIR Publishers, Moscow, 1978.

[15] A. H. NAYFEH, *Perturbation Methods*, John Wiley, New York, 1973.

[16] H. S. NUR, *Singular perturbations of differential equations in abstract spaces*, Pacific J. Math., 36 (1971), pp. 775-780.

[17] A. Y. SCHOENE, *Semi-groups and a class of singular perturbation problems*, Indiana Univ. Math. J., 78 (1963) pp. 247-263.

[18] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[19] M. I. VISIK AND L. A. LYUSTERNIK, *Regular degeneration and boundary-layer for linear differential equations with small parameter*, Uspekhi Mat. Nauk., 12 (1957), pp. 3-122.

[20] M. WHEELER, *A Priori L_2 Error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723-759.

[21] M. ZLAMAL, *Sur l'équation des telegraphistes avec un petit parametre*, Atti Accad. Naz. Lincei Rend. C1. Sci. Fis. Mat. Natur., 27 (1959), pp. 324-332.

[22] ———, *On a singular perturbation problem concerning hyperbolic equations*, Lecture Series 45, The Institute for Fluid Dynamics and Applied Mathematics, Univ. of Maryland, College Park, MD, November, 1964.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

015

IDENTIFICATION OF THE CONDUCTIVITY COEFFICIENT IN AN ELLIPTIC EQUATION*

AVNER FRIEDMAN† AND BJÖRN GUSTAFSSON‡

Abstract. Consider an elliptic equation in a two-dimensional domain Ω with conductivity coefficient $a=1+k\chi_D$ $(k\neq 0)$ where D is a subdomain of Ω. From the measurements of a pair of Dirichlet and Neumann data one wishes to identify D. It is proved that this problem is stable in some local sense.

Key words. elliptic equations, conductivity coefficient, identification problem, electrical prospecting

AMS(MOS) subject classifications. Primary 35R30; secondary 35J25, 35R05

Introduction. Consider an elliptic equation

$$\operatorname{div}(a\nabla u)=0 \quad \text{in } \Omega \tag{0.1}$$

with Dirichlet data

$$u=f \quad \text{on } \partial\Omega, \quad f\not\equiv \text{const}, \tag{0.2}$$

and with coefficient $a=1+k\chi_D$ $(-1<k<\infty,\ k\neq 0)$, where D is an unknown subdomain of Ω. We seek to determine D by measurements of the Neumann data

$$\frac{\partial u}{\partial \nu}=g \quad \text{on } \partial\Omega. \tag{0.3}$$

This identification problem arises in electrical prospecting, whereby one wishes to discover the location of metals or fluid reservoirs inside the earth.

Let $D(t)$ be a 1-parameter monotone family of domains with $D(0)=D$ such that

$$\left.\frac{d}{dt}\chi_{D(t)}\right|_{t=0}\neq 0 \quad \text{in } \mathcal{D}', \tag{0.4}$$

and denote by $u(t)\equiv u(x,t)$ the solution of (0.1), (0.2) corresponding to $a=1+k\chi_{D(t)}$. Our main result asserts that, in case the $D(t)$ are affine transformations of D, C^1 in t, for all t with $|t|$ small enough, there holds

$$\left\|\frac{\partial}{\partial \nu}[u(t)-u(0)]\right\|_{L^2(\partial\Omega)}\geqq c|t| \tag{0.5}$$

where c is a positive constant.

If we denote by Φ the mapping from a to g (when f is fixed) then (0.5) means formally that $d\Phi/da\neq 0$; thus, if $\Phi(a_1)=g_1$, $\Phi(a_2)=g_2$ and $\|a_2-a_1\|$ is small, then

$$\|a_2-a_1\|\leqq C\|g_2-g_1\| \quad \text{where } \frac{1}{\|d\Phi/da\|}\leqq C<\infty. \tag{0.6}$$

This means that the computation of D among a monotone family of domains is stable with respect to small errors in the measurement of the Neumann data; for more details on the significance of a result of this type see [12] and § 1 below.

* Received by the editors July 24, 1985; accepted for publication April 8, 1986. This work was partially supported by National Science Foundation grant MCS-8300293.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

‡ Royal Institute of Technology, Stockholm, Sweden.

There are other versions of identification problems. In [2], [4]-[6], [11] one measures the quadratic form

$$Q_a = \int_\Omega a|\nabla u|^2$$

for *all* f and shows that this determines $a = a(x)$ in Ω, provided that either $a(x)$ is piecewise analytic [4], [6] or $\|a-1\|$ is small enough [11]. For some special domains the identification problem can be resolved by separation of variables [3], [8] or by explicit representation of u by means of Green's function [9].

In another version (0.1) is replaced by

$$\operatorname{div}(a\nabla u) = l \quad \text{in } \Omega \quad (l \text{ is given})$$

and one wishes to find a, given the knowledge of u throughout *all* of Ω; see [1], [10] and the references given there. This problem is unstable.

References to physical models and numerical computations of identification problems are given in [1], [5].

1. The main result. Let Ω be a bounded simply connected domain in $\mathbb{R}^2$ with $C^{1,\alpha}$ boundary $\partial\Omega$ $(0<\alpha<1)$ and let D be a bounded subdomain of Ω with $C^{2,\alpha}$ boundary ∂D, $\bar{D} \subset \Omega$. We shall designate points in $\mathbb{R}^2$ by $x = (x_1, x_2)$.

Denote by χ_A the characteristic function of a set A.

We assume that D is star-shaped with respect to any point x^* of some nonempty subset D^* of D.

For any $x^* \in D^*$ introduce the 1-parameter family of domains

$$D(t) = \{x^* + (1-t)(x-x^*), x \in D\} \quad (-1<t<1). \tag{1.1}$$

Then $D(t) \subset D(t')$ if $t > t'$. Also

$$\frac{\partial}{\partial t}\chi_{D(t)}\bigg|_{t=0} = \beta \otimes \delta_{\partial D} \quad \text{in } \mathscr{D}', \tag{1.2}$$

that is

$$\frac{\partial}{\partial t}\left[\iint_\Omega \chi_{D(t)}\phi\right]\bigg|_{t=0} = \int_{\partial D} \beta\phi$$

for any $\phi \in C_0^0(\Omega)$, and β is a continuous and strictly negative function on ∂D; $\beta \in C^{1,\alpha}$.

Set $D_e(t) = \Omega\backslash\overline{D(t)}$. We shall use the notation w^e (or w^i) to denote the value of a function w on $\partial D(t)$ taken as a limit from $D_e(t)$ (or $D(t)$).

Let k be a fixed number, $-1<k<0$ or $k>0$, and set

$$a(x,t) = 1 + k\chi_{D(t)}(x), \qquad a(x) = a(x,0). \tag{1.3}$$

Consider the elliptic equation

$$\operatorname{div}(a(x,t)\nabla u) = 0 \quad \text{in } \Omega \tag{1.4}$$

with the Dirichlet condition

$$u = f \quad \text{on} \quad \partial\Omega \tag{1.5}$$

where $f = f(x)$ is in $C^{1,\alpha}(\partial\Omega)$.

It is well known [7] that the solution u of this diffraction problem is in $C^{0,\beta}(\bar{\Omega}) \cap H^1(\Omega)$ for some $0 < \beta < 1$, as well as in $C^{2,\alpha}(\overline{D(t)})$ and in $C^{2,\alpha}(\overline{D_e(t)} \backslash \partial\Omega)$, and that

$$\frac{\partial u^e}{\partial \nu} = (k+1)\frac{\partial u^i}{\partial \nu} \quad \text{on } \partial D(t) \tag{1.6}$$

where ν is the outward normal to $\partial D(t)$.

Set

$$g(x, t) = \frac{\partial u(x, t)}{\partial \nu}, \qquad x \in \partial\Omega \tag{1.7}$$

where ν is the outward normal to $\partial\Omega$. Then $g \in C^\alpha$.

We would like to determine the conductivity coefficient $a(x)$ from measurements of $g(x) \equiv g(x, 0)$. Since in real terms we can only measure $g(x)$ with some error, we would like to ensure that if the measurements give us a function $g(x, t)$ "close" to $g(x)$ then the corresponding $a(x, t)$ is also "close" to the true coefficient $a(x)$. If that is the case, then by compiling a catalog of various g's corresponding to various a's we can have an effective way of determining the true conductivity: We simply correspond to a function $\tilde{g}$ that we obtained by actual measurements the coefficient a which fits to that g in our catalog that is "nearest" to $\tilde{g}$. This point of view is quite common in inverse problems [12].

If $f \equiv$ const. then $u \equiv$ const. for any choice of $a(x, t)$ and thus $g(x, t) \equiv 0$. This means that we cannot gain any information on the coefficient a. Thus we must henceforth assume that

$$f \not\equiv \text{const.} \tag{1.8}$$

THEOREM 1.1. *If* (1.8) *holds then there exists a positive constant c such that*

$$\|g(\cdot, h) - g(\cdot)\|_{L^2(\partial\Omega)} \geqq c|h| \tag{1.9}$$

if $|h|$ *is small enough.*

Theorem 1.1 extends to more general monotone families of domains $D(t)$; see § 3.

Theorem 1.1 means that we can effectively determine D by the procedure outlined in the paragraph following (1.7), provided D is known to be imbedded in a monotone family of domains.

As we shall see in § 3 (Remark 3.2), Theorem 1.1 is generally false if $D(t)$ is not a monotone family (at least in one space dimension, or for Ω an annulus).

The remainder of this paper is devoted to the proof of Theorem 1.1; some generalizations are mentioned at the end of § 3.

2. Proof of Theorem 1.1. Set $g(t) \equiv g(x, t)$. To prove the theorem it suffices to assume that

$$\left\| \frac{g(h) - g(0)}{h} \right\|_{L^2(\partial\Omega)} \to 0 \quad \text{for some sequence } h \to 0 \tag{2.1}$$

and derive a contradiction. From now on h will be restricted to this sequence.

Consider first the case where $0 < h < 1$, so that

$$D(h) \subset D, \tag{2.2}$$

and set

$$a(t)\equiv a(x,t), \qquad u(t)\equiv u(x,t),$$
$$a_h=\frac{a(h)-a(0)}{h}, \qquad v_h=\frac{u(h)-u(0)}{h}.$$

From (1.4) we get

$$\operatorname{div}(a(h)\nabla u(h)-a(0)\nabla u(0))=0, \tag{2.3}$$

which implies that there exists a function w^h in $H^1(\Omega)$ such that

$$\frac{1}{h}[a(h)\nabla u(h)-a(0)\nabla u(0)]=\operatorname{curl} w^h; \tag{2.4}$$

here curl $w=(w_{x_2},-w_{x_1})$. We normalize w^h so that

$$w^h(x^0)=0 \quad \text{at some point } x^0\in\partial\Omega. \tag{2.5}$$

Since

$$\frac{1}{h}[a(h)\nabla u(h)-a(0)\nabla u(0)]=a(0)\nabla u_h+a_h\nabla u(h) \tag{2.6}$$

we can rewrite (2.4) in the form

$$a(0)\frac{\partial}{\partial\bar z}v_h+i\frac{\partial}{\partial\bar z}w^h=-a_h\frac{\partial u(h)}{\partial\bar z}. \tag{2.7}$$

Introduce the function

$$f^h=a(0)v_h+iw^h. \tag{2.8}$$

Then

$$\frac{\partial f^h}{\partial\bar z}=\frac{\partial a(0)}{\partial\bar z}v_h+a(0)\frac{\partial v_h}{\partial\bar z}+i\frac{\partial w^h}{\partial\bar z} \quad \text{in } \mathcal{D}' \tag{2.9}$$

and, using (2.7),

$$\frac{\partial f^h}{\partial\bar z}=\frac{\partial a(0)}{\partial\bar z}v_h-a_h\frac{\partial u(h)}{\partial\bar z}: \tag{2.10}$$

the right-hand side is a measure.

It follows by Cauchy's formula that

$$\begin{aligned} f^h(\zeta)=&-\frac{1}{\pi}\iint_\Omega\left(\frac{\partial a(0)}{\partial\bar z}v_h-a_h\frac{\partial u(h)}{\partial\bar z}\right)(z)\frac{1}{z-\zeta}\,dx\,dy\\ &+\frac{1}{2\pi i}\int_{\partial\Omega}\frac{f^h(z)}{z-\zeta}\,dz, \qquad \zeta\in\Omega.\end{aligned} \tag{2.11}$$

LEMMA 2.1. *As* $h\to 0$,

$$D_x^\gamma v_h\to 0 \quad \textit{uniformly in compact subsets of } D_e, \quad 0\leqq|\gamma|\leqq 1, \tag{2.12}$$

and

$$\int_{\partial D}|v_h|\,ds\to 0. \tag{2.13}$$

Proof. Take for simplicity $x^*=0$. Define

$$\Omega_h=\left\{x\in\Omega \text{ and } \frac{x}{1-h}\in\Omega\right\}=\Omega\cap(1-h)\Omega,$$

$$u(x)=u(x,0),$$

$$U^h(x)=\frac{1}{h}\left[u\left(\frac{x}{1-h}\right)-u(x)\right] \quad \text{in } \Omega_h,$$

$$V^h(x)=\frac{1}{h}\left[u(x,h)-u\left(\frac{x}{1-h}\right)\right] \quad \text{in } \Omega_h.$$

Then

$$v_h(x)=U^h(x)+V^h(x) \quad \text{in } \Omega_h.$$

By the $C^{2,\alpha}$ regularity of u in $D_e\cup\partial D$,

$$|U^h|+|\nabla U^h|\leqq C \quad \text{in } D_e\cap\Omega_h. \tag{2.14}$$

Next,

$$\operatorname{div}(a(h)\nabla V^h)=\frac{1}{h}\left\{0-\operatorname{div}_x\left[a(h)\nabla_x u\left(\frac{x}{1-h}\right)\right]\right\}.$$

Since

$$a(h)=1+k\chi_{D(h)}(x)=1+k\chi_{(1-h)D}(x)=1+k\chi_D\left(\frac{x}{1-h}\right),$$

setting $y=x/(1-h)$, $\nabla_x=\nabla_y/(1-h)$, the expression in braces becomes

$$-\frac{1}{(1-h)^2}\operatorname{div}_y[1+k\chi_D(y)\nabla_y u(y)]=0 \quad \text{in } \Omega_h,$$

i.e.,

$$\operatorname{div}(a(h)\nabla V^h)=0 \quad \text{in } \Omega_h. \tag{2.15}$$

Further,

$$V^h(x)=\frac{u(x,h)-f(x/(1-h))}{h} \quad \text{if } x\in\partial\Omega_h\cap\partial((1-h)\Omega),$$

$$V^h(x)=\frac{f(x)-u(x/(1-h))}{h} \quad \text{if } x\in\partial\Omega_h\cap\partial\Omega;$$

since $u(x)$ and $u(x,h)$ are in $C^{1,\alpha}$ in some neighborhood of $\partial\Omega$, it follows that in both cases

$$|V^h|\leqq C,$$

i.e., $|V^h|\leqq C$ on $\partial\Omega_h$. Hence, by the maximum principle,

$$|V^h|\leqq C \quad \text{in } \Omega_h. \tag{2.16}$$

Recalling (2.14) we conclude that

$$|v_h|\leqq C \quad \text{in } D_e\backslash B_\delta(\partial\Omega)$$

for any $\delta > 0$ and h small enough, where $B_\delta(A)$ denotes a δ-neighborhood of a set A. Since v_h is harmonic in D_e and $v_h = 0$ on $\partial\Omega$, we then also have that

$$|v_h| \leqq C \quad \text{in } D_e. \tag{2.17}$$

Hence, for a subsequence,

$$v_h \to v \text{ uniformly on compact subsets of } D_e \cup \partial\Omega \tag{2.18}$$

where v is harmonic in D_e and

$$v = \frac{\partial v}{\partial \nu} = 0 \quad \text{on } \partial\Omega; \tag{2.19}$$

here (2.1) was used. It follows that the zero function is a harmonic extension of v into $\mathbb{R}^2 \backslash \Omega$ and therefore $v \equiv 0$ in D_e. Clearly (2.12) now follows from (2.18) and Harnack's theorem.

To prove (2.13) we multiply (2.15) by V^h and integrate over $\Omega' = \Omega \backslash B_\eta(\partial\Omega)$, where $0 < \eta < \operatorname{dist}(D, \partial\Omega)$. We obtain, for small h,

$$\iint_{\Omega'} a(h)|\nabla V^h|^2 \leqq \int_{\partial\Omega'} a(h)|\nabla V^h|\,|V^h|.$$

Since V^h is harmonic and bounded in $B_{\delta_0}(\partial\Omega')$ (δ_0 is independent of h) the right-hand side is uniformly bounded; hence

$$\int_{\Omega'} |\nabla V^h|^2 \leqq C.$$

Recalling (2.14) we deduce that

$$\int_{\Omega' \cap D_e} |\nabla v_h|^2 \leqq C. \tag{2.20}$$

Now, for any small $\delta > 0$,

$$\int_{\partial D} |v_h| \leqq C \int_{D_e \cap B_\delta(\partial D)} |\nabla v_h| + C \int_{(\Omega' \cap D_e) \backslash B_\delta(\partial D)} |\nabla v_h| + C \int_{\partial\Omega'} |v_h|.$$

The last two integrals on the right-hand side converge to zero as $h \to 0$, whereas the first integral is bounded by $C\delta^{1/2}$ (by (2.20)). It follows that

$$\limsup_{h \to 0} \int_{\partial D} |v_h| \leqq C\delta^{1/2},$$

and, since δ is arbitrary, (2.13) follows.

From (2.12), (2.5), (2.7) we conclude that

$$f^h \to 0, \ \nabla f^h \to 0 \text{ uniformly in closed subsets of } D_e \cup \partial\Omega. \tag{2.21}$$

Taking $h \to 0$ in (2.11) we see that if

$$I = \lim_{h \to 0} \left(-\frac{1}{\pi}\right) \iint_\Omega \frac{\partial a(0)}{\partial \bar{z}} \frac{v_h}{z - \zeta}\, dx\, dy, \tag{2.22}$$

$$J = \lim_{h \to 0} \left(-\frac{1}{\pi}\right) \iint_\Omega a_h \frac{\partial u(h)}{\partial \bar{z}} \frac{1}{z - \zeta}\, dx\, dy \tag{2.23}$$

exist for any $\zeta \in D \cup D_e$, then $f^h(\zeta) \to f^0(\zeta)$, where

$$f^0(\zeta) = I - J \quad \text{if } \zeta \in D \cup D_e; \tag{2.24}$$

from (2.21) we also have

$$f^0(\zeta)=0 \quad \text{if } \zeta \in D_e. \tag{2.25}$$

Now clearly

$$\frac{\partial a(0)}{\partial \bar{z}}=\gamma \otimes \delta_{\partial D} \quad \text{in } \mathscr{D}'$$

where γ is a $C^{1,\alpha}$ function on ∂D. Therefore

$$\lim_{h\to 0}\iint_{\Omega}\frac{\partial a(0)}{\partial \bar{z}}\,\frac{v_h}{z-\zeta}\,dx\,dy=\lim_{h\to 0}\int_{\partial D}\frac{\gamma}{z-\zeta}v_h(z)\,ds=0$$

by (2.13), i.e.,

$$I=0. \tag{2.26}$$

Next, by (1.6) and the fact that $u(t)$ is in $C^1(\overline{D(t)})$ and in $C^1(\overline{D_e(t)})$ with moduli of continuity independent of t, it follows that

$$\lim_{h\to 0}\iint_{\Omega} a_h\frac{\partial u(h)}{\partial \bar{z}}\,\frac{1}{z-\zeta}\,dx\,dy=\int_{\partial D} k\beta\left(\frac{\partial u(0)}{\partial \bar{z}}\right)^e\frac{1}{z-\zeta}\,ds;$$

here we used (1.2) and (2.2). We conclude that J exists and, by (2.24), (2.26),

$$f^0(\zeta)=\frac{k}{\pi}\int_{\partial D}\beta\left(\frac{\partial u}{\partial \bar{z}}\right)^e\frac{1}{z-\zeta}\,ds \quad \text{if } \zeta\in D\cup D_e; \tag{2.27}$$

here $u=u(x,0)$.

Let $T(z)$ be the positively oriented tangent vector to ∂D at z. Then

$$dz=T(z)\,ds \quad \text{along } \partial D.$$

Using this in (2.27) we get

$$f^0(\zeta)=\frac{k}{\pi}\int_{\partial D}\frac{\beta}{T(z)}\left(\frac{\partial u}{\partial \bar{z}}\right)^e\frac{dz}{z-\zeta} \qquad (\zeta\in D\cup D_e). \tag{2.28}$$

In view of (2.25) and the standard jump relation of the integral in (2.28) across ∂D, we then have

$$f^0(z)=2ik\frac{\beta(z)}{T(z)}\left(\frac{\partial u}{\partial \bar{z}}\right)^e \quad \text{on } \partial D \tag{2.29}$$

where $f^0(z)=(f(z))^i$ is the limit of f^0 from D. Observe also from (2.28) that

$$f^0(z) \quad \text{is holomorphic in } D. \tag{2.30}$$

LEMMA 2.2. *There holds*

$$\beta\left(\frac{\partial u}{\partial \bar{z}}\right)^e=0 \quad \textit{on } \partial D. \tag{2.31}$$

The proof is given in § 3. Assuming its validity, we shall now proceed to complete the proof of Theorem 1.1. Since $\beta\neq 0$ along ∂D,

$$\left(\frac{\partial u}{\partial \bar{z}}\right)^e=0 \quad \text{on } \partial D.$$

Recalling the jump relation (1.6) we deduce that for some constant c, the function $U = u - c$ vanishes on ∂D together with its first derivatives. By the argument following (2.19) it then follows that $U \equiv 0$ in $B_\delta(\partial D)$ and, by analytic continuation, $u \equiv c$ in D_e which contradicts (1.8).

So far we have assumed that (2.2) holds. If $-1 < h < 0$, so that

$$D(h) \supset D, \tag{2.32}$$

then we replace (2.6) by

$$\frac{1}{h}[a(h)\nabla u(h) - a(0)\nabla u(0)] = a(h)\nabla v_h + a_h \nabla u(0)$$

and proceed as above (with minor changes) to establish (2.24) with the corresponding I vanishing and with J being the same as before.

3. Proof of Lemma 2.2. Set

$$u_1 = u|_{D_e}, \qquad u_2 = u|_D.$$

Then

$$u_1 = u_2 \quad \text{on } \partial D, \tag{3.1}$$

$$\frac{\partial u_1}{\partial \nu} = (k+1)\frac{\partial u_2}{\partial \nu} \quad \text{on } \partial D. \tag{3.2}$$

Notice that the function $\partial u_2/\partial z$ is homomorphic in D. Multiplication of both sides of (2.29) by $\partial u_2/\partial z$ gives

$$F'(z)T(z) = 2ik\beta(z)\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z} \qquad (z \in \partial D) \tag{3.3}$$

where F is a holomorphic function in D, namely, the primitive of $f^0(z)\partial u_2/\partial z$.

Along ∂D we have

$$2\frac{\partial}{\partial z} = \frac{\partial}{\partial x} - i\frac{\partial}{\partial y} = e^{i\omega}\left(\frac{\partial}{\partial \nu} - i\frac{\partial}{\partial s}\right),$$

$$2\frac{\partial}{\partial \bar z} = \frac{\partial}{\partial x} + i\frac{\partial}{\partial y} = e^{-i\omega}\left(\frac{\partial}{\partial \nu} + i\frac{\partial}{\partial s}\right)$$

where ν is the outward normal, $\partial/\partial s$ is in the tangential direction obtained from $\partial/\partial \nu$ by rotation counterclockwise by $\pi/2$ and ω is a real valued function. Therefore by (3.1), (3.2) we easily obtain

$$4\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z} = (k+1)\left(\frac{\partial u_2}{\partial \nu}\right)^2 + \left(\frac{\partial u_2}{\partial s}\right)^2 - ik\frac{\partial u_2}{\partial \nu}\frac{\partial u_2}{\partial s}. \tag{3.4}$$

Hence, if $k > 0$,

$$4\left|\operatorname{Im}\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z}\right| = k\left|\frac{\partial u_2}{\partial \nu}\frac{\partial u_2}{\partial s}\right| \leqq \frac{k}{2}\left[\left(\frac{\partial u_2}{\partial \nu}\right)^2 + \left(\frac{\partial u_2}{\partial s}\right)^2\right] \leqq 2k\left[\operatorname{Re}\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z}\right].$$

Similarly, if $-1 < k < 0$,

$$4\left|\operatorname{Im}\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z}\right| \leqq \frac{|k|}{2(k+1)}\left[(k+1)\left(\frac{\partial u_2}{\partial \nu}\right)^2 + \left(\frac{\partial u_1}{\partial s}\right)^2\right] = \frac{2|k|}{k+1}\left[\operatorname{Re}\frac{\partial u_1}{\partial \bar z}\frac{\partial u_2}{\partial z}\right].$$

Since β is real valued it follows that in both cases, for any $z \in \partial D$,

$$F'(z)T(z) \in G \equiv \{z = x_1 + ix_2;\ |x_2| \leqq C|x_1|\} \tag{3.5}$$

where

$$C=\frac{2|k|}{\min\{1,k+1\}}.$$

Writing the holomorphic function F in the form $F=V+iW$ we have

$$\frac{dF}{ds}=V_s+iW_s=V_s+iV_\nu \quad \text{along}\,\partial D.$$

Since also

$$\frac{dF}{ds}=\frac{dF}{dz}\frac{dz}{ds}=F'(z)T(z),$$

we conclude from (3.4) that

$$|V_\nu|\leqq C|V_s| \quad \text{along}\,\partial D. \tag{3.6}$$

Suppose $V\not\equiv$ const. in D. Then V must attain its maximum in $\bar{D}$ at a point $x^0\in\partial D$ and $V_\nu(x^0)>0$. Since also $V_s(x^0)=0$, we get a contradiction to (3.6). We have thus proved that $V\equiv$ const. and therefore also $F\equiv$ const. From (3.3) it then follows that

$$\beta\frac{\partial u_1}{\partial\bar{z}}\frac{\partial u_2}{\partial z}=0 \quad \text{on}\,\partial D$$

which, in view of (3.4) and (3.2), implies (2.31).

Remark 3.1. Theorem 1.1 extends (with minor changes in the proof) to the case where the domains $D(t)$ are conformal affine transformations of D varying in C^2 manner and monotonically in t, provided $\beta\not\equiv 0$ on ∂D. The theorem also extends to the case where f depends on t, say $f=f(x,t)$, provided

$$\frac{1}{h}[f(\cdot,h)-f(\cdot,0)]\to 0 \quad \text{in}\,C^{1,\alpha}(\partial D)\text{-norm}$$

as $h\to 0$. If the $D(t)$ do not vary monotonically in t, then Lemma 2.1 is still valid with (2.13) replaced by

$$\int_{\partial(D\cup D(h))}|v_h|\,ds\to 0.$$

But this is not sufficient for proving (2.26); see also next remark.

Remark 3.2. Consider the case where Ω is one-dimensional, say $\Omega=\{0<x<1\}$. The solution of (1.1) with $u(0)=\alpha$, $u(1)=\beta$, $u'(0)=\alpha'$, $u'(1)=\beta'$ is given by

$$u(x)=\alpha+u(0)\alpha'\int_0^x\frac{dy}{a(y)} \tag{3.7}$$

where

$$u(0)\alpha'=(\beta-\alpha)\Big/\int_0^1\frac{dy}{a(y)},\qquad \beta'=\frac{a(0)\alpha'}{a(1)}. \tag{3.8}$$

For any other conductivity $\tilde{a}(x)$ with

$$\tilde{a}(0)=a(0),\quad \tilde{a}(1)=a(1),\quad \int_0^1\frac{dy}{a(y)}=\int_0^1\frac{dy}{\tilde{a}(y)}, \tag{3.9}$$

the Neumann data g corresponding to the Dirchlet data f are the same as for a. Clearly (3.9) is satisfied if $\tilde{a}(t)=1+k\chi_{D(t)}$, $a=\tilde{a}(0)$ whenever $D(t)$ is a translation of D. In this example the mapping $a\rightarrow g$ is thus nonunique; furthermore, the assertion of Theorem 1.1 is not valid if $D(t)$ is a translation of D. If however $D(t)$ is monotone in t then the assertion of Theorem 1.1 is valid, as can be verified directly by means of (3.9). Similarly, if Ω is an annulus $\Omega=\{r_1<|x|<r_2\}$ and $f=c_i$ on $\{|x|=r_i\}$, c_i constants, then the assertion of Theorem 1.1 is valid for a family of annuli $D(t)=\{d_1(t)<|x|<d_2(t)\}$ provided the family is monotone in t, but it is generally false if the $D(t)$ do not vary monotonically in t (note however that Ω is not simply connected, as required in Theorem 1.1).

Remark 3.3. Let $\tilde{z}=\phi(z)$ be a conformal mapping of $\bar{\Omega}$ onto the closure of a domain $\tilde{\Omega}$ and set $\tilde{D}(t)=\phi(D(t))$, $a=\tilde{a}\circ\phi$, $f=\tilde{f}\circ\phi$, $u=\tilde{u}\circ\phi$, $\tilde{g}=|\phi'|\tilde{g}\circ\phi$. Then (1.4), (1.5) and (1.7) are equivalent to

$$\operatorname{div}(\tilde{a}\nabla\tilde{u})=0 \quad \text{in } \tilde{\Omega}, \qquad \tilde{u}=\tilde{f} \quad \text{on } \partial\tilde{\Omega},$$

$$\frac{\partial\tilde{u}}{\partial\tilde{\nu}}=\tilde{g} \quad \text{on } \partial\tilde{\Omega}.$$

Since

$$0<c\leqq\frac{\|\tilde{g}(t)-\tilde{g}(0)\|_{L^2}}{\|g(t)-g(0)\|_{L^2}}\leqq C<\infty,$$

Theorem 1.1 extends to the family $\tilde{D}(t)$ of subdomains of $\tilde{\Omega}$.

Remark 3.4. Theorem 1.1 extends to inhomogeneous equations

$$\operatorname{div}(a\nabla u)=l(x) \quad \text{in } \Omega$$

provided $l\in C^{1,\alpha}$ and $S\equiv\operatorname{supp} l$ satisfies: $S\subset D_e$ and $D_e\backslash S$ is connected; if $l\not\equiv 0$ and $l>0$, then the condition (1.8) is not needed. The function l may also be taken to depend on t.

Remark 3.5. The results of this paper extend with minor changes to the case where the Neumann data (1.7) are prescribed, whereas the Dirichlet data $f=f(t,x)$ are measured; here it is assumed that $\int_{\partial\Omega} g=0$ and u is normalized, say, by $\int_{\partial\Omega} u=0$. The assertion (1.9) is replaced by

$$\|f(\cdot,h)-f(\cdot)\|_{L^2(\partial\Omega)}\geqq c|h|$$

where c is a positive constant.

REFERENCES

[1] G. ALESSANDRINI, *On the identification of the leading coefficient of an elliptic equation*, Serie "Problemi non ben posti ed inverse," no. 17, University of Florence, 1984.

[2] A. P. CALDERON, *On an inverse boundary value problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, 1980, pp. 65-73.

[3] J. R. CANNON, J. DOUGLAS AND B. F. JONES, *Determination of the diffusivity of an isotropic medium*, Internat. J. Engrg. Sci., 1 (1963), pp. 453-455.

[4] R. KOHN AND M. VOGELIUS, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289-298.

[5] ———, *Identification of an unknown conductivity by means of measurements at the boundary* in Inverse Problems, D. W. McLaughlin, ed., Society for Industrial and Applied Mathematics-American Mathematical Society Proc. 14, 1984, pp. 113-123.

[6] R. KOHN AND M. VOGELIUS, *Determining conductivity by boundary measurements* II, *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643-668.

[7] O. A. LADYZHENSKAJA AND N. N. URAL'TZEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, London, 1968.

[8] R. E. LANGER, *An inverse problem in differential equations*, Bull. Amer. Math. Soc., 39 (1933), pp. 814-820.

[9] A. LORENZI AND C. D. PAGANI, *On the stability of the surface separating two homogeneous media with different thermal conductivities*, Technical Report, University of Milano, 1979.

[10] G. R. RICHTER, *An inverse problem for the steady state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210-221.

[11] J. SYLVESTER AND G. A. UHLMAN, *A uniqueness theorem for an inverse boundary value problem in electrical prospection*, Comm. Pure Appl. Math., 39 (1985), pp. 95-112.

[12] G. TALENTI, *Sui problemi mal posti*, Boll. Un. Mat. Ital., 15-A (1978), 1-29.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

016

THE ELECTROPAINTING PROBLEM WITH OVERPOTENTIALS*

VIVIANA MARQUEZ† AND MEIR SHILLOR‡

Abstract. Existence, uniqueness and regularity are proved for the electropainting problem with overpotentials. The problem consists of finding a pair $\{\varphi(x, t), h(x, t)\}$ such that $\varphi(x, t)$ is a time dependent family of harmonic functions, representing the electric potential in a domain, and $h(x, t)$ is related to the paint thickness on the part of the boundary being painted. The boundary condition on this part is $\varphi_n = G(\varphi, h)$ where φ_n is the inward normal derivative. h is determined from the history of the process. The assumption on the overpotential $\sigma(x)$ implies $h \geqq \sigma(x) > 0$ and thus the boundary condition is nondegenerate. We show that the process is monotone; there is no paint dissolution. Then we consider the explicit time discretization of the problem. Letting the time step shrink to zero leads to the above mentioned results. Then the $t \to \infty$ limit is considered, existence, uniqueness and regularity are proved. Moreover it is shown that this asymptotic limit can be recast as a Signorini variational inequality with an obstacle constructed from $\varphi(x, t)$, $t < \infty$. Finally the degenerate case $\sigma(x) \equiv 0$ is considered and the existence of a weak solution is proved, in a convex geometry, using monotonicity arguments.

Key words. electropainting model, overpotentials, evolution problem for the Laplacian, Signorini problem

AMS(MOS) subject classifications. 35J65, 35R35

1. Introduction.

We consider an evolution problem associated with an electropainting process with overpotentials. The problem is to find a pair $\{\varphi(x, t), h(x, t)\}$ such that in an annular region $\Omega \subset \mathbb{R}^n$ $(n \geqq 2)$ with outer boundary S and inner boundary Γ there holds

$$(1.1) \qquad \Delta\varphi = 0 \quad \text{in } \Omega, \quad 0 \leqq t \leqq T,$$

$$(1.2) \qquad \varphi = 1 \quad \text{on S}, \quad 0 \leqq t \leqq T,$$

$$(1.3) \qquad \varphi_n = G(\varphi, h) \quad \text{on } \Gamma, \quad 0 \leqq t \leqq T,$$

$$(1.4) \qquad h(x, t) = \sigma(x) + \int_0^t g((\varphi_n(x, \tau) - \varepsilon)^+)\, d\tau, \qquad x \in \Gamma, \quad 0 \leqq t \leqq T$$

where φ_n is the inward normal derivative on Γ, $(z)^+ = \max\{0, z\}$, G, g and σ are given functions, $\varepsilon > 0$ and $T > 0$ are given constants.

Such problems were considered by Hansen and McGeough [6], Aitchison, Lacey and Shillor [1] and Caffarelli and Friedman [2]. The first two deal with the modeling aspects of the electropainting process and numerical experiments. The third deals with mathematical analysis of the model. In all these papers it was assumed that $g(s) = s$, $G(\varphi, h) = \varphi/h$ and $\sigma = 0$, thus the problem considered here is more general mathematically as well as from the practical point of view.

A problem of the type (1.1)-(1.4) can be considered (see [1] or [6]) as a model for the following process. A metal body with an outer surface Γ, to be painted, is immersed in a bath with an electrolytic solution. The solution occupies the region Ω such that $\partial\Omega = \Gamma \cup S$, where S is the inner surface of the bath. The metal part is

* Received by the editors September 30, 1985; accepted for publication (in revised form) May 8, 1986.

† Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina.

‡ Instituto Argentino de Matemática-CONICET, Viamonte 1636, 1055 Buenos Aires, and Departamento de Matemática, Facultad de Ingeniería, Universidad de Buenos Aires, 1000 Buenos Aires, Argentina. Present address, Department of Mathematics, Imperial College of Science and Technology, London, England SW7 2BZ.

connected to an electric potential source, the bath itself (S) serves as the other electrode and as a result of the flow of the electric current in the solution and into Γ the process of paint deposition takes place on Γ. The existence of a cutoff current $\varepsilon > 0$, that was postulated in [1], assures that there is paint deposition only at those points of Γ where the current φ_n satisfies $\varphi_n > \varepsilon$. Indeed one of the main purposes of the construction of the model in [1] was to be able to predict which parts of Γ become painted and which remain bare. The model that was proposed in [6] is the same as in [1] but with $\varepsilon = 0$ and therefore in this case Γ becomes completely covered by paint for any $t > 0$. Thus the numerical experiments in [1] were done with the purpose of showing painted and unpainted regions in the Signorini problem that corresponds to the steady state and the evolution of the paint layer in the time dependent problem. The numerical experiments in [6] were to study the saturation and leveling effects.

Some mathematical analysis of the model in [1] was performed in [2]. First they showed that under the conditions of [1] there is no paint dissolution, i.e. the process is monotone, and therefore the boundary condition that was given in [1] for that possibility is redundant. Then they considered a time discretized version of the model and proved the existence and uniqueness of the discretized solution and its convergence to the steady state Signorini problem that was conjectured in [1].

We consider a more general problem. It is well known (see e.g. Levich [9] or McGeough [10]) that when electric current is passing in a solution some chemical reactions take place near the electrodes that lead to a nonlinear relationship between the potential and the current. This is represented above in (1.3). Moreover by taking $\sigma(x) > 0$ we allow for the existence of a resistance to current besides the paint layer. More specifically, there may exist initially on Γ a thin coating of a different material, which is a common occurrence in the industrial process; also the production of gas bubbles and a buildup of different reaction products may block the way of the paint molecules and thus form a resistance on Γ. As far as we know there exists no description of what takes place near Γ, and $\sigma(x)$ should be found experimentally. Since in cases of practical interest some or all of these phenomena are likely to be present the fact that most of our results are obtained for $\sigma(x) \geqq \sigma_* > 0$ does not restrict the applicability of these results. In the numerical experiments in [1] it was $\sigma \equiv 0$ and this did not cause any trouble, but the method of computation was not particularly sensitive to weak divergence of the normal derivative. In our case the (nondimensional) paint thickness is $h_*(x, t) = h(x, t) - \sigma(x)$ at $x \in \Gamma$, $0 \leqq t$. These two phenomena are referred to as "overpotentials" (see [9], [10] and Lacey [7]). In addition we consider a more general process of paint deposition, represented by g. Although the main practical interest is the paint thickness h_* as a function of space and time it turns out that the total resistance h is the dominant factor in the model.

We prove existence, uniqueness and asymptotic behaviour of the solution to (1.1)-(1.4) and obtain some regularity.

But first, following the ideas of [2], we too prove, in § 2, that the smooth solution to (1.1)-(1.4) with an appropriate condition for paint dissolution ($h_t < 0$) is monotone (i.e. $h_t \geqq 0$) and therefore such a condition is not needed. We stress that this is so under the given condition (1.2) on S. It is clear mathematically and from the practical point of view that under different conditions dissolution can take place. Indeed if the voltage is switched off, i.e. $\varphi = 0$ on S after some time $t \geqq t_0 > 0$, or the voltage φ on S is an oscillating function with both positive and negative values that are sufficiently large, then one should find that $h_t < 0$ on a portion of Γ. It is also shown that the process tends to an asymptotic limit, that is considered in § 5, but cannot attain it in finite time. This is in contrast to the result in [2], where such a possibility was not ruled out.

The existence and uniqueness of the weak solution are proved in § 4 where under additional assumptions on G it is shown that the weak solution is a smooth solution. The proofs are based on an explicit time discretization of the problem that is considered in § 3. Existence, uniqueness and regularity are proved for the discretized problem and then some necessary bounds, independent of the time step δ, are derived. The results of § 4 follow from the $\delta \to 0$ limit and these bounds.

In § 5 we prove that the process tends asymptotically to a unique steady state. Moreover using the fact that this steady problem is the limit of an evolution problem we are able to give a variational inequality formulation, a Signorini-like problem, where the obstacle is constructed from $\varphi(x, t)$, $t < \infty$.

It may be of interest to investigate what types of elliptic boundary value problems, like the problem of the steady state, can be given variational inequality formulations by their imbedding as asymptotic limits of evolution problems. In their analysis in [2] they used implicit time discretization and introduced $\sigma_* > 0$ as a regularizing parameter, we use an explicit one in § 3. Since they considered the limit $\sigma_* \to 0$ they were unable to consider the limit $\delta \to 0$ and hence proved the existence and uniqueness of a time discretized solution. We took the limit $\delta \to 0$ for $\sigma_* > 0$ and obtained the above mentioned results. Also the steady state in [2] is the Signorini problem but ours is more complicated.

Finally the nondegeneracy condition $\sigma(x) \geqq \sigma_* > 0$ is essential for all the proofs mentioned above. Nevertheless we are able to prove, in § 6, the existence of a weak solution to the degenerate problem with $\sigma(x) \equiv 0$ in a convex geometry as a monotone limit of solutions with $\sigma(x) \equiv \sigma > 0$, as $\sigma \to 0$. Thus in this geometry, where S is taken as the inner electrode while Γ, assumed to be convex, is the outer boundary, there exists a solution to the problem that was considered in [1] and [2].

The regularity of the free boundary, the boundary of the set in Γ where $h > \sigma$ and the regularity of the solution with $\sigma = 0$ remain open questions.

2. No paint dissolution.

The model for the electropaint process that was proposed in [1] included a condition for the case of the dissolution of the paint layer. But it was proved in [2] that any smooth solution can be modified smoothly so that the solution and the paint layer are nondecreasing in time. We give the generalization of the model in [1] to include overpotentials and nonlinear growth condition for the paint thickness and then prove that any smooth solution is monotone increasing with time. So under such conditions, $\varphi = 1$ on S, the paint dissolution does not occur and therefore the appropriate condition can be omitted from the model.

The fact that there is no paint dissolution has some importance from the mathematical point of view as well as from the practical one. Indeed it seems that if dissolution can take place then the free boundary (i.e. the boundary on Γ of the region where there is no paint deposition) is likely to be linearly unstable and so it is likely that the model will break down leading to an ill posed problem (see e.g. [3] and [4]).

We consider the process in $\Omega \subset R^n$, $n \geqq 2$, a doubly connected region with outer boundary S and inner boundary Γ. Everywhere below it is assumed that S and Γ are in $C^{1+\alpha}$ for any $\alpha \in (0, 1)$.

The modified model is to find a pair $\{\varphi(x, t), h(x, t)\}$ such that

$$\Delta\varphi = 0 \quad \text{in } \Omega, \quad 0 \leqq t, \tag{2.1}$$

$$\varphi = 1 \quad \text{on } S, \quad 0 \leqq t, \tag{2.2}$$

$$\varphi_n = G(\varphi, h) \quad \text{on } \Gamma, \quad 0 \leqq t, \tag{2.3}$$

$$h_t = g((\varphi_n - \varepsilon)^+) \quad \text{on } \Gamma, \quad 0 \leqq t \quad \text{if } h = \sigma(x), \tag{2.4$'$}$$

$$(2.4)'' \qquad h_t = g((\varphi_n - \varepsilon)) \quad \text{on } \Gamma, \quad 0 \leqq t \quad \text{if } h > \sigma(x),$$

$$(2.4)''' \qquad h(x, 0) = \sigma(x) \quad \text{on } \Gamma$$

where φ_n is the inward normal derivative on Γ and ε is a positive constant. G is prescribed and g is an odd function with $g(s) > 0$ for $s > 0$. $\sigma(x)$ is a given overpotential and the paint thickness at a point $x \in \Gamma$ at time t is given by $h(x, t) - \sigma(x)$. Condition (2.4)′ means that at $x \in \Gamma$ if there is no paint, i.e., $h = \sigma$, then paint deposition will start only if $\varphi_n > \varepsilon$. On the other hand at a point $x \in \Gamma$ where there is paint, i.e., $h > \sigma$, then by (2.4)″, if $\varphi_n > \varepsilon$ then there is paint deposition and when $\varphi_n < \varepsilon$ there is paint dissolution. It follows from (2.4)′, (2.4)″ and (2.4)‴ that

$$(2.5) \qquad h(x, t) \geqq \sigma(x), \qquad x \in \Gamma, \quad 0 \leqq t.$$

We shall need the following assumptions:

$$(2.6) \qquad \sigma(x) \in C^{0,1}(\Gamma), \qquad 0 < \sigma_* \leqq \sigma(x) \leqq \sigma^* < K_2/\varepsilon,$$

$$(2.7) \qquad g(s) \in C^1(\mathbb{R}), \quad 0 < k_* \leqq g' \leqq k, \quad g(0) = 0,$$

$$(2.8) \qquad \begin{aligned} &G(s, p) \in C^1([0, 1] \times \mathbb{R}_+), \quad G(0, p) = 0, \quad G(s, p) > 0 \quad \text{if } s > 0, \\ &G(1, p) < K_2/p, \qquad p \in \mathbb{R}_+, \end{aligned}$$

$\partial G/\partial s > 0$, $\partial G/\partial p \leqq 0$ and both bounded on compact subsets of $[0, 1] \times \mathbb{R}_+$, $\partial G/\partial p < 0$ if $s > 0$. They are used everywhere below (except in § 6).

Denote by ω_* the solution of

$$(2.9) \qquad \begin{aligned} &\Delta\omega_* = 0 \quad \text{in } \Omega, \quad \omega_* = 1 \quad \text{on } S, \\ &\omega_{*n} = G(\omega_*, \sigma) \quad \text{on } \Gamma. \end{aligned}$$

So $\omega_*(x) = \varphi(x, 0)$, where φ is a solution to (2.1)-(2.4)‴. If

$$(2.10) \qquad \omega_{*n} \leqq \varepsilon \quad \text{on } \Gamma$$

then $\varphi(x, t) \equiv \omega_*(x)$ together with $h(x, t) \equiv \sigma(x)$ form a solution of (2.1)-(2.4)‴.

In order to exclude this trivial case we shall assume that the geometry and the data are such that

$$(2.11) \qquad \omega_{*n} > \varepsilon \quad \text{for some points } x \in \Gamma.$$

DEFINITION 2.1. By a *smooth solution* of (2.1)-(2.4)‴ we mean a solution (φ, h) such that φ, φ_t, $\nabla\varphi$ are continuous in $\bar{\Omega} \times [0, \infty)$ and h, h_t are continuous on $\Gamma \times [0, \infty)$.

DEFINITION 2.2. Denote by t_0 the supremum of all s such that $h(x, t)$ is nondecreasing for all $x \in \Gamma$, $0 \leqq t < s$.

THEOREM 2.3. *Let* (φ, h) *be a smooth solution of* (2.1)-(2.4)‴ *and let* (2.11) *hold. Then* $t_0 = \infty$.

This result is stronger than the one in [2], where the authors were able to assert only that if $t_0 < \infty$ then one can join a solution continuously to the steady solution at $t = t_0$ in order to obtain a nondecreasing solution for $0 \leqq t < \infty$. This is so, as well as the uniqueness of the smooth solution, due to the fact that $0 < \sigma_* \leqq \sigma(x)$.

Proof. We follow [2] in part. By (2.3)

$$\varphi_n = G(\varphi, h) \quad \text{on } \Gamma.$$

Since $h \geqq \sigma > 0$ we can use the strong maximum principle and (2.8) to deduce that φ cannot take a positive maximum or a nonpositive minimum on Γ and therefore

$$(2.12) \qquad c_* \leqq \varphi(x, t) \leqq 1, \qquad x \in \bar{\Omega}, \quad 0 \leqq t$$

for some $c_* > 0$.

Since this holds for any function $G(s,p)$ on $\mathbb{R}\times\mathbb{R}_+$ such that $G(s,p)>0$, $s\neq 0$, $s\in\mathbb{R}$, the restriction of $s\in[0,1]$ in (2.8) leads to no loss of generality.

Thus by (2.3) and (2.8)

$$\varphi_n(x,t)>0 \quad \text{on } \Gamma. \tag{2.13}$$

LEMMA 2.4. *If $h(x,t_1)\geqq h(x,t_2)$ and $h(x,t_1)\not\equiv h(x,t_2)$ on Γ then*

$$\varphi(x,t_1)>\varphi(x,t_2) \quad \textit{in } \Omega.$$

Proof. The function $\psi(x)=\varphi(x,t_1)-\varphi(x,t_2)$ is harmonic in Ω, vanishes on S and satisfies

$$\begin{aligned}\psi_n &= G(\varphi(x,t_1),h(x,t_1))-G(\varphi(x,t_2),h(x,t_2))\\ &= G_\varphi(*)(\varphi(x,t_1)-\varphi(x,t_2))+G_h(*)(h(x,t_1)-h(x,t_2))\end{aligned}$$

where here and below, $G_\varphi(*)$ and $G_h(*)$ stand for the partial derivatives evaluated at some value by the mean value theorem, and since $\varphi>0$ then $G_h(*)<0$ by (2.8). But $h(x,t_1)\geqq h(x,t_2)$ so

$$\psi_n\leqq G_\varphi(*)\psi, \qquad G_\varphi>0$$

by (2.8). Applying the strong maximum principle we show that $\psi>0$ in Ω.

Assume that $t_0<\infty$. Then there exists a sequence (x_i,t_i) with $x_i\in\Gamma$, $t_i>t_0$ such that

$$t_i\to t_0, \qquad x_i\to x_0 \quad \text{as } i\to\infty$$

and $h_t(x_i,t_i)<0$; consequently also $h(x_i,t_i)>\sigma(x_i)$. Clearly $h_t(x_0,t_0)=0$.

LEMMA 2.5. *There holds*: $\varphi_{nt}(x_0,t_0)\leqq 0$.

Proof. If $\varphi_n(x_i,t_0)\geqq\varepsilon$ then, since $\varphi_n(x_i,t_i)<\varepsilon$,

$$\varphi_{nt}(x_i,\tilde{t}_i)<0 \quad \text{for some } t_0<\tilde{t}_i<t_i. \tag{2.14}$$

If, on the other hand, $\varphi_n(x_i,t_0)<\varepsilon$ then $h(x_i,t_0)=\sigma(x_i)$ (since $h(x_i,t_0)>\sigma(x_i)$ implies $h_t(x_i,t_0)=g(\varphi_n(x_i,t_0)-\varepsilon)<0$, a contradiction to the definition of t_0). Since further $h(x_i,t_i)>\sigma(x_i)$ we conclude that $h_t(x_i,\hat{t}_i)>0$ for some $t_0<\hat{t}_i<t_i$ and thus $\varphi_n(x_i,\hat{t}_i)>\varepsilon$, which yields (since $\varphi_n(x_i,t_i)<\varepsilon$)

$$\varphi_{nt}(x_i,\tilde{\tilde{t}}_i)<0 \quad \text{for some } \hat{t}_i<\tilde{\tilde{t}}_i<t_i. \tag{2.15}$$

The lemma now follows from (2.14), (2.15) upon taking $i\to\infty$.

From Lemma 2.4 we have

$$\varphi_t\geqq 0 \quad \text{if } 0<t\leqq t_0. \tag{2.16}$$

LEMMA 2.6. *There holds* $h_t(x,t_0)\equiv 0$.

Proof. Suppose that $h_t(x,t_0)\not\equiv 0$. For $t=t_0$

$$\varphi_{nt}=G_\varphi\varphi_t+G_h h_t \quad \text{on } \Gamma; \tag{2.17}$$

hence

$$\varphi_{nt}\leqq G_\varphi\varphi_t \qquad (G_h<0).$$

We can use the strong maximum principle to deduce that φ_t cannot take the minimum on Γ and so

$$\varphi_t>0 \quad \text{on } \Gamma. \tag{2.18}$$

At (x_0, t_0), $h_t=0$ and $\varphi_t>0$ by (2.18). Then (2.17) implies that $\varphi_{nt}(x_0, t_0)>0$, a contradiction to Lemma 2.5.

Set $\Gamma_0=\Gamma\cap\{\varphi_n(x, 0)>\varepsilon\}$ and for any $0<t_1<t_2<t_0$ let

$$\varphi^i(x)=\varphi(x, t_i) \quad \text{and} \quad h^i(x)=h(x, t_i), \quad i=1, 2.$$

Then

$$\varphi_n^2-\varphi_n^1=G_\varphi(*)(\varphi^2-\varphi^1)+G_h(*)(h^2-h^1) \tag{2.19}$$

rearranging

$$-G_h(*)(h^2-h^1)+(\varphi_n^2-\varphi_n^1)=G_\varphi(*)(\varphi^2-\varphi^1) \quad \text{on } \Gamma.$$

By Lemma 2.4, $\varphi^2\geqq\varepsilon^1$; hence

$$-G_h(*)\int_{t_1}^{t_2} h_t\,dt\geqq-(\varphi_n^2-\varphi_n^1).$$

On Γ_0, $h_t=g(\varphi_n-\varepsilon)\leqq k(\varphi_n-\varepsilon)$ if $t<t_0$; hence

$$c\int_{t_1}^{t_2}(\varphi_n-\varepsilon)\,dt\geqq-((\varphi_n^2-\varepsilon)-(\varphi_n^1-\varepsilon)), \qquad c>0 \quad \text{on } \Gamma_0. \tag{2.20}$$

Integrating (2.20) over Γ_0 we find

$$c\int_{t_1}^{t_2}\left(\int_{\Gamma_0}(\varphi_n-\varepsilon)\right)dt\geqq-\left(\int_{\Gamma_0}(\varphi_n^2-\varepsilon)-\int_{\Gamma_0}(\varphi_n^1-\varepsilon)\right),$$

and setting

$$\psi(t)=\int_{\Gamma_0}(\varphi_n(x, t)-\varepsilon),$$

we arrive at the inequality

$$c\int_{t_1}^{t_2}\psi\,dt\geqq-(\psi(t_2)-\psi(t_1))$$

or $\dot\psi+c\psi\geqq0$. So ψe^{ct} is monotone nondecreasing for $0<t<t_0$. Since $\psi(0)>0$ it follows that $\psi(t_0)>0$, which is a contradiction to Lemma 2.6, i.e., $h_t(x, t_0)\equiv0$. Hence $t_0<\infty$ is impossible and so the theorem is proved.

COROLLARY 2.7. *If $h(\bar x, s)>\sigma(\bar x)$ for some $\bar x\in\Gamma$, $s>0$, then $h_t(\bar x, t)>0$ for all $t>s$.*

Thus, once $h(\bar x, t)$ becomes greater than $\sigma(\bar x)$, i.e. once paint deposition starts at $\bar x$, it does not stop and h continues to grow at a strictly positive rate.

To prove the corollary we proceed by contradiction and consider the number t_1 such that $h_t(\bar x, t)>0$ if $\bar s\leqq t<t_1$ and $h_t(\bar x, t_1)=0$ where $\bar s\leqq s$ is such that $\varphi_n(\bar x, \bar s)>\varepsilon$. Applying the argument of Lemma 2.6, we deduce that $h_t(x, t_1)\equiv0$, which is impossible from the proof of Theorem 2.3.

Theorem 2.3 shows that, under the condition $\varphi=1$ on S, there is no paint dissolution and therefore (2.4)′-(2.4)‴ in the model can be replaced by

$$h(x, t)=\sigma(x)+\int_0^t g((\varphi_n-\varepsilon)^+)\,d\tau, \qquad x\in\Gamma, \quad 0\leqq t. \tag{2.4}$$

Finally we prove the following.

THEOREM 2.8. *The smooth solution is unique.*

Proof. Let (φ^1, h^1) and (φ^2, h^2) be two smooth solutions.

Subtracting condition (2.3) for $i=1$ from the one for $i=2$ we find (2.19), only for the same time. From (2.8) $G_\varphi>0$ and $G_h\leqq0$. Now a point of positive maximum of $\varphi^2-\varphi^1$ can be only on Γ and then $\varphi_n^2-\varphi_n^1<0$ by the strong maximum principle. Hence it follows from (2.19) that

$$G_\varphi(*)(\varphi^2-\varphi^1)\leqq -G_h(*)(h^2-h^1) \tag{2.21}$$

and since a similar argument applies to a negative minimum it follows that

$$|\varphi^2-\varphi^1|_\infty\leqq K|h^2-h^1|_\infty. \tag{2.22}$$

On the other hand

$$\begin{aligned}|h^2-h^1|&\leqq\int_0^\delta |g((\varphi_n^2-\varepsilon)^+)-g((\varphi_n^1-\varepsilon)^+)|\,dt\\&\leqq k\delta|\varphi_n^2-\varphi_n^1|_\infty.\end{aligned} \tag{2.23}$$

Combining (2.19), (2.22) and (2.23) we get

$$|\varphi_n^2-\varphi_n^1|\leqq K|h^2-h^1|_\infty\leqq K\delta|\varphi_n^2-\varphi_n^1|_\infty,$$

which implies that $\varphi_n^2\equiv\varphi_n^1$ if we take $\delta<K^{-1}$. But this implies that $\varphi^1\equiv\varphi^2$, both being harmonic functions, and also that $h^1\equiv h^2$ by (2.4).

3. The time discretized problem. We consider the time discretized version of the evolutionary process (2.1)-(2.4). We retain all the assumptions on the data and on Ω from § 2. We use explicit time discretization in order to obtain approximate solutions and then we obtain the necessary uniform bounds.

For any $T>0$ and a large integer $M>0$ let

$$\delta=\frac{T}{M}, \tag{3.1}$$

and set $t_m=m\delta$, $m=0,1,\cdots,M$. If we replace $\varphi(x,t_m)$ by $\varphi^m(x)$ in (2.1)-(2.4) we obtain the following explicit finite differences system

$$\Delta\varphi^m=0 \quad \text{in } \Omega, \tag{3.2}$$

$$\varphi^m=1 \quad \text{on } S, \tag{3.3}$$

$$\varphi_n^m=G(\varphi^m,h^m) \quad \text{on } \Gamma, \tag{3.4}$$

$$h^m=\sigma+\delta\sum_{i=0}^{m-1}g((\varphi_n^i-\varepsilon)^+) \quad \text{on } \Gamma, \tag{3.5}$$

where $m=0,1,2,\cdots,M$.

LEMMA 3.1. *There exists a unique solution* $\{\varphi^m(x),h^m(x)\}_{m=0}^M$ *of* (3.2)-(3.5) *with* $\varphi^m\in C^{1+\alpha}(\bar\Omega)$, $h^m\in C^\alpha(\Gamma)$ *any* $\alpha\in(0,1)$, $0\leqq m\leqq M$.

Proof. Proceeding by induction it is enough to prove that the elliptic problem

$$\Delta u=0 \quad \text{in } \Omega, \tag{3.6}$$

$$u=1 \quad \text{on S}, \tag{3.7}$$

$$u_n=G(u,\gamma) \quad \text{on } \Gamma \tag{3.8}$$

has a unique solution in $C^{1+\alpha}(\bar\Omega)$ any $\alpha\in(0,1)$ provided $0<\gamma$ and $\gamma\in C^\alpha(\Gamma)$. By the maximum principle and by (2.8) it is seen that a priori $0<u\leqq1$ in $\bar\Omega$. Let $\beta\in(0,1)$ and let

$$V_M=\{v\in C^\beta(\bar\Omega),\ \|v\|_{C^\beta(\bar\Omega)}\leqq M\},\qquad M>0. \tag{3.9}$$

For any $v \in V_M$ we solve (3.6) and (3.7) together with

$$u_n = G(v, \gamma) \quad \text{on } \Gamma.$$

But by (2.8) and the assumptions on v and γ, $G(v, \gamma)$ is in $C^\beta(\Gamma)$; hence $u_n \in C^\beta(\Gamma)$ and therefore $u \in C^{1+\beta}(\bar{\Omega})$ (see e.g. [5, p. 117]) and $\|u\|_{C^{\beta'}(\bar{\Omega})} \leqq M_0$ for any $\beta \leqq \beta' < 1$, where M_0 depends only on the data and β' but independent of γ. If we choose $M = M_0$ in (3.9) and define a mapping $u = Tv$ on V_M then $T: V_{M_0} \to V_{M_0}$ is continuous. Moreover $u \in C^{1+\beta}(\bar{\Omega})$; hence T has a compact range in V_{M_0} and so by Schauder's fixed point theorem T has a fixed point u such that $u = Tu$ which is the solution to (3.6)-(3.8); moreover $u \in C^{1+\beta}(\bar{\Omega})$. This argument applies to any $\beta \in (0, 1)$, hence $u \in C^{1+\alpha}(\bar{\Omega})$, any $\alpha \in (0, 1)$. Now returning to the discretized system (3.2)-(3.5) we see that if $\varphi^m \in C^{1+\alpha}(\bar{\Omega})$ then $\varphi^m_n \in C^\alpha(\Gamma)$ and so $(\varphi^m_n - \varepsilon)^+ \in C^\alpha(\Gamma)$ and therefore so is $g((\varphi^m_n - \varepsilon)^+)$ and hence, by (3.5), $h^{m+1} \in C^\alpha(\Gamma)$. Clearly this holds for any $\alpha \in (0, 1)$. To prove uniqueness we take the difference of any two solutions and apply to it the maximum principle.

We note that since $h^i \geqq h^{i-1}$ the proof of Lemma 2.4 gives

$$\varphi^i \geqq \varphi^{i-1} \quad \text{in } \Omega, \quad i \geqq 1. \tag{3.10}$$

Recall also that by the maximum principle

$$0 < \varphi^i \leqq 1 \quad \text{in } \bar{\Omega}. \tag{3.11}$$

For any $x_0 \in \Gamma$ denote by $j_0 = j(x_0)$ the first integer $j_0 \geqq 0$, if existing, such that

$$\begin{aligned} &\varphi^i_n \leqq \varepsilon \quad \text{if } i \leqq j_0 - 1, \\ &\varphi^{j_0}_n > \varepsilon. \end{aligned} \tag{3.12}$$

Notice that $h^i(x_0) = \sigma(x_0)$ if $i \leqq j_0$ and $h^{j_0+1}(x_0) > h^{j_0}(x_0)$.

Remark 3.2. We choose to work with the explicit scheme, i.e. the summation in (3.5) is $0 \leqq i \leqq m-1$. The proof of Lemma 3.1 is thus more direct and simple than in an implicit scheme where questions of the invertibility of the equation

$$\varphi_n = G(v, \gamma + \delta g((\varphi_n - \varepsilon)^+))$$

would arise. Moreover from the numerical point of view the explicit scheme is simpler to consider and to program since no iterations are needed. On the other hand the explicit scheme has an unattractive feature—we cannot assure, as was done in Corollary 2.7, that once $h^i(x_0) > \sigma(x_0)$ then $h^{j+1}(x_0) > h^j(x_0)$ holds for every $j \geqq i$; i.e., the computed h may not be strictly monotone increasing. But nevertheless the explicit scheme does introduce a correction in the cases of overshooting. Indeed let $m \geqq j_0$ be an integer such that

$$\varphi^m_n(x_0) > \varepsilon, \qquad \varphi^{m+1}_n(x_0) \leqq \varepsilon. \tag{3.13}$$

Then we find from (3.5) that

$$h^{m+1}(x_0) > h^m(x_0), \qquad h^{m+2}(x_0) = h^{m+1}(x_0). \tag{3.14}$$

Since

$$-G_h(*)(h^{m+2} - h^{m+1}) + (\varphi^{m+2}_n - \varphi^{m+1}_n) = G_\varphi(*)(\varphi^{m+2} - \varphi^{m+1})$$

and so (3.14) and (3.10) imply that $\varphi^{m+2}_n \geqq \varphi^{m+1}_n$ at x_0, i.e. the normal derivative increases in the correct way at one time step later. On the other hand it is easy to show that in an implicit scheme $h^{i+1}(x_0) > h^i(x_0)$ if $i > j_0$, i.e. the analogue of Corollary 2.7.

We proceed to obtain uniform bounds on the solution to (3.2)-(3.5).

LEMMA 3.3. *There holds*

$$\sigma(x) \leqq h^m(x) \leqq \frac{K_2}{\varepsilon} + 1 \quad \forall x \in \Gamma, \quad m \geqq 0 \tag{3.15}$$

for all $\delta < \sigma_* / kK_2$.

Proof. From (3.5) the left inequality is immediate. Now let $m \geqq 0$, $x \in \Gamma$ and suppose that there exists $l \geqq m$ such that $\varphi_n^l(x) \geqq \varepsilon$. Then by (3.4) and (2.8)

$$\varphi_n^l(x) = G(\varphi^l(x), h^l(x)) \leqq G(1, h^l(x)) \leqq \frac{K_2}{h^l(x)} \tag{3.16}$$

and so

$$h^l(x) \leqq \frac{K_2}{\varphi_n^l(x)} \leqq \frac{K_2}{\varepsilon}. \tag{3.17}$$

Since $h^m(x) \leqq h^l(x)$ for every $l \geqq m$, the other inequality holds in this case.

Now suppose that $\varphi_n^l(x) < \varepsilon$ for every $l \geqq m$. If $h^m(x) > (K_2/\varepsilon) + 1$ then by (2.6) $h^m(x) > \sigma(x)$ and so there exists $j \leqq m-1$ such that $\varphi_n^j(x) > \varepsilon$ and $\varphi_n^i(x) \leqq \varepsilon$ for every $j < i \leqq m-1$; then

$$h^m(x) = h^{j+1}(x) > h^j(x).$$

As (3.17) is true also for $l = j$ we find that $h^{j+1}(x) - h^j(x) \geqq 1$ and so

$$\delta k(\varphi_n^j - \varepsilon) \geqq \delta g(\varphi_n^j - \varepsilon) \geqq 1.$$

Hence

$$\varphi_n^j(x) \geqq \frac{1}{\delta k}. \tag{3.18}$$

Combining (3.16) with $l = j$ and (3.18) we obtain

$$h^j(x) \leqq \frac{K_2}{\varphi_n^j(x)} \leqq K_2 k \delta < \sigma_*,$$

which is impossible.

LEMMA 3.4. *For every* $\alpha \in (0, 1)$ *there exists a constant* $C > 0$ *such that for any* $0 < \sigma^* < K_2/\varepsilon$, $0 < \delta < \sigma_*/kK_2$, $m \geqq 0$

$$\|\varphi^m\|_{C^\alpha(\bar{\Omega})} \leqq C. \tag{3.19}$$

Proof. This follows from Theorem 4.2 [8, Chap. 5, p. 333].

LEMMA 3.5. *There holds*

$$0 \leqq \varphi_n^m \leqq G_* \tag{3.20}$$

where G_* *is independent of* δ *and of* T.

Proof. From (2.8), $G(s, p)$ is positive for $s > 0$ and is monotone increasing in s and monotone decreasing in p. Therefore the fact that $0 < \varphi^m \leqq 1$ and Lemma 3.3 imply that

$$\varphi_n^m = G(\varphi^m, h^m) \leqq G(1, \sigma_*) \equiv G_*, \qquad m \geqq 0. \tag{3.21}$$

LEMMA 3.6. *There exists a constant* $C > 0$, *independent of* δ *and* T, *such that*

$$\|\varphi^m\|_{H^1(\Omega)} \leqq C, \qquad m \geqq 0. \tag{3.22}$$

Proof. If we multiply (3.2) by (φ^m-1) and use Green's theorem we obtain

$$\int_\Omega |\nabla\varphi^m|^2=\int_\Gamma (1-\varphi^m)\varphi_n^m \leqq \operatorname{meas}\Gamma\cdot G_*, \qquad m\geqq 0.$$

LEMMA 3.7. *For every $\alpha\in(0,1)$ there exists a constant $C>0$, that depends on the data, on α and on T, but is independent of δ, such that*

$$\|\varphi_n^m\|_{C^\alpha(\Gamma)}, \|h^m\|_{C^\alpha(\Gamma)}\leqq C, \qquad m\geqq 0. \tag{3.23}$$

Proof. For any $x, y\in\Gamma$, $m>0$, we have as in (2.19) that

$$\varphi_n^m(x)-\varphi_n^m(y)=G_\varphi(*)(\varphi^m(x)-\varphi^m(y))+G_h(*)(h^m(x)-h^m(y)).$$

If we divide both sides by $|x-y|^\alpha$, take the supremum over all $x, y\in\Gamma$, use (2.8) and the fact that $|G_h(*)|\leqq K_3$ and $|G_\varphi(*)|<K_1$ which follows from Lemma 3.3, we find

$$H_\alpha(\varphi_n^m)\leqq K_1H_\alpha(\varphi^m)+K_3H_\alpha(h^m) \tag{3.24}$$

where $H_\alpha(f)$ stands for the Hölder constant of f. From Lemma 3.4 we have that $K_1H_\alpha(\varphi^m)\leqq K_4$ and K_4 is independent of δ and T. From (3.5) and (2.7) we see that

$$|h^m(x)-h^m(y)|\leqq|\sigma(x)-\sigma(y)|+\delta k\sum_{i=0}^{m-1}|\varphi_n^i(x)-\varphi_n^i(y)|,$$

and therefore

$$H_\alpha(h^m)\leqq H_\alpha(\sigma)+\delta k\sum_{i=0}^{m-1}H_\alpha(\varphi_n^i). \tag{3.25}$$

Inserting (3.25) in (3.24), we have

$$H_\alpha(\varphi_n^m)\leqq K_4+K_3H_\alpha(\sigma)+\delta kK_3\sum_{i=0}^{m-1}H_\alpha(\varphi_n^i).$$

In order to simplify the notation let $b_m=H_\alpha(\varphi_n^m)$, $m\geqq 0$, $D=kK_3$ and $a=K_4+K_3H_\alpha(\sigma)$. Then the last inequality can be written in the form

$$b_m\leqq a+\delta D\sum_{i=0}^{m-1}b_i, \qquad m\geqq 1. \tag{3.26}$$

Let $B=a+\delta Db_0$ then one can show by induction that

$$b_m\leqq(1+\delta D)^{m-1}B, \qquad m\geqq 1,$$

and since $\delta=T/M$ we find that

$$b_M\leqq\left(1+\frac{TD}{M}\right)^M B\leqq B\lim_{M\to\infty}\left(1+\frac{TD}{M}\right)^M=(a+Db_0)\,e^{TD}.$$

Therefore $H_\alpha(\varphi_n^m)$, for every $m\geqq 0$ is bounded for $T<\infty$ and it follows from (3.25) that $H_\alpha(h^m)$ is similarly bounded, hence (3.23) is satisfied.

LEMMA 3.8. *There exist positive constants C_1, C_2 independent of δ and of T such that*

$$\left|\frac{\varphi^{m+1}(x)-\varphi^m(x)}{\delta}\right|\leqq C_1 \quad \forall x\in\bar\Omega, \quad m\geqq 0 \tag{3.27}$$

and

$$\left|\frac{\varphi_n^{m+1}(x)-\varphi_n^m(x)}{\delta}\right|\leqq C_2 \quad \forall x\in\Gamma, \quad m\geqq 0. \tag{3.28}$$

Proof. We have that $\varphi^{m+1}-\varphi^m>0$ in Ω, (if $\varphi^{m+1}\equiv\varphi^m$ in $\bar\Omega$ then the assertion is clear), is harmonic and vanishes on S. Therefore its positive maximum is obtained on Γ, say at $x_0\in\Gamma$. Clearly at that point $\varphi_n^{m+1}-\varphi_n^m<0$. Therefore it follows from

$$\varphi_n^{m+1}-\varphi_n^m=G_\varphi(*)(\varphi^{m+1}-\varphi^m)+G_h(*)(h^{m+1}-h^m) \tag{3.29}$$

that at x_0, using (3.5) and (2.8),

$$G_\varphi(*)(\varphi^{m+1}-\varphi^m)<|G_h|\delta g((\varphi_n^m-\varepsilon)^+).$$

But $K\leqq G_\varphi$ for some $K>0$, from (2.8) and in view of Lemma 3.3 and $|G_h|<K_3$ and $g'\leqq k$; hence

$$|\varphi^{m+1}-\varphi^m|\leqq\left(\frac{K_3}{K}kG_*\right)\delta$$

where Lemma 3.5 was used, and so (3.27) is proved. Now (3.28) follows from (3.29) if we use (3.27), the necessary bounds on G_φ, G_h and $g((\varphi_n^m-\varepsilon)^+)<kG_*$.

LEMMA 3.9. *There exists a positive constant $C>0$, independent of δ and of T, such that*

$$\int_\Omega\left|\nabla\left(\frac{\varphi^{m+1}-\varphi^m}{\delta}\right)\right|^2\leqq C,\qquad m\geqq 0. \tag{3.30}$$

Proof. From

$$0=\delta^{-2}\int_\Omega(\varphi^{m+1}-\varphi^m)\Delta(\varphi^{m+1}-\varphi^m)$$

we obtain

$$\int_\Omega\left|\nabla\left(\frac{\varphi^{m+1}-\varphi^m}{\delta}\right)\right|^2\leqq\int_\Gamma\left|\frac{\varphi^{m+1}-\varphi^m}{\delta}\right|\left|\frac{\varphi_n^{m+1}-\varphi_n^m}{\delta}\right|$$

and the right-hand side is bounded by (meas Γ) $\cdot\, C_1\cdot C_2$ in view of Lemma 3.8.

4. The weak solution. In this section we use the results of the time discretized problem, take the limit $\delta\to 0$ and deduce the existence of a weak solution to (2.1)-(2.4).

DEFINITION 4.1. A *weak solution* to (2.1)-(2.4) is a pair (φ, h) of functions such that for any $0<T<\infty$:

(i)
$$\varphi\in W^{1,2}(0,T;H^1(\Omega))\cap C^\alpha(\bar\Omega\times[0,T]),$$
$$\varphi\in C^{1+\alpha}(\bar\Omega),\qquad 0\leqq t\leqq T,$$
$$h,h_t\in C^\alpha(\Gamma\times[0,T])$$

for any $\alpha\in(0,1)$;

(ii)
$$\Delta\varphi=0\quad\text{in }\Omega,\quad 0\leqq t\leqq T,$$
$$\varphi=1\quad\text{on }S\times[0,T];$$

(iii) The (inward) normal derivative on Γ, φ_n, satisfies

$$\varphi_n\in C^\alpha(\Gamma\times[0,T])\quad\text{for any }\alpha\in(0,1),$$

and

$$\varphi_n=G(\varphi,h)\quad\text{on }\Gamma\times[0,T];$$

(iv)
$$h(x,t)=\sigma(x)+\int_0^t g((\varphi_n(x,\tau)-\varepsilon)^+)\,d\tau\quad\text{on }\Gamma\times[0,T].$$

Thus (2.1)-(2.4) are satisfied but the weak solution lacks some regularity with respect to t.

THEOREM 4.2. *There exists a unique weak solution to* (2.1)-(2.4) *for any* $T>0$. *Moreover*

$$\varphi \in W^{1,\infty}(0, T; H^1(\Omega)). \tag{4.1}$$

Proof. Let φ^m be the solution of the time discretized problem (3.2)-(3.5) with $\delta = T/M$, such that $\delta < \sigma_*/kK_2$, and let $\varphi_\delta(x, t)$ denote the linear interpolation, in time, of the φ^m's, that is

$$\begin{aligned}\varphi_\delta(x, t) = [\varphi^m(x) - \varphi^{m-1}(x)][t-(m-1)\delta]\delta^{-1} + \varphi^{m-1}(x),&\\ x \in \bar{\Omega}, \qquad (m-1)\delta \leqq t \leqq m\delta.&\end{aligned} \tag{4.2}$$

By Lemma 3.1 $\varphi^m \in C^{1+\alpha}(\bar{\Omega})$ for all $m \geqq 0$ and therefore $\varphi_\delta \in C^{1+\alpha}(\bar{\Omega})$, and by Lemma 3.7

$$\|\varphi_\delta\|_{C^{1+\alpha}(\bar{\Omega})} \leqq C, \qquad 0 \leqq t \leqq T \quad \text{for any } \alpha \in (0, 1) \tag{4.3}$$

where C depends on α, and, here and below, C is a constant independent of δ.

From Lemma 3.8 we have that

$$\left|\frac{\partial \varphi_\delta}{\partial t}\right| \leqq C \quad \text{in } \bar{\Omega} \times [0, T] \tag{4.4}$$

where C is independent of T. Therefore there exists a constant C such that for any $\alpha \in (0, 1)$

$$\|\varphi_\delta\|_{C^\alpha(\bar{\Omega}\times[0, T])} \leqq C \tag{4.5}$$

and C depends on α but is independent of T. From Lemma 3.6 it follows that

$$\|\varphi_\delta\|_{H^1(\Omega)} \leqq \frac{C}{\sigma_*}, \qquad 0 \leqq t \leqq T \tag{4.6}$$

and C is independent of T. From (4.4) we find that

$$\|\varphi_\delta\|_{H^1(\Omega\times(0,T))} \leqq C. \tag{4.7}$$

Let $\delta \to 0$. For simplicity we use the same notation for subsequences. Then we have that

$$\varphi_\delta \to \varphi \begin{cases} \text{uniformly in } C^\alpha(\bar{\Omega} \times [0, T]) \quad \text{for any } \alpha \in (0, 1), \\ \text{uniformly in } C^{1+\alpha}(\bar{\Omega}), \qquad 0 \leqq t \leqq T. \end{cases} \tag{4.8}$$

Clearly the limit function satisfies $\varphi = 1$ on S. Let $\zeta \in C_0^\infty(\Omega)$; then

$$\int_\Omega \nabla\zeta \cdot \nabla\varphi_\delta = 0 \quad \forall \delta > 0,$$

so (4.8) implies that

$$\int_\Omega \nabla\zeta \cdot \nabla\varphi = 0,$$

and therefore $\Delta\varphi = 0$ in Ω, $t \in [0, T]$, so (ii) of the definition is satisfied. Now from Lemma 3.9 we have that

$$\int_\Omega \left|\nabla \frac{\partial \varphi_\delta}{\partial t}\right|^2 dx \leqq C, \qquad t \in (0, T), \tag{4.9}$$

$C>0$ independent of T; hence together with (4.4), (4.6) and (4.8) it follows that as $\delta \to 0$, $\varphi \in W^{1,\infty}(0, T; H^1(\Omega))$, i.e. (4.1) holds.

Let $\varphi_{\delta,n}$ be the linear interpolations, in time, of the φ_n^m,

$$\varphi_{\delta,n}(x,t)=[\varphi_n^{m+1}(x)-\varphi_n^m(x)][t-m\delta]\delta^{-1}+\varphi_n^m(x), \qquad x\in\Gamma, \quad m\delta\leqq t\leqq(m+1)\delta \tag{4.10}$$

then $\varphi_{\delta,n}\in C^\alpha(\Gamma)$, for any $\alpha\in(0,1)$, all $t\in[0,T]$, by Lemma 3.1. And

$$\|\varphi_{\delta,n}\|_{C^\alpha(\Gamma)}\leqq C,$$

then Lemma 3.8 implies that

$$\left|\frac{\partial\varphi_{\delta,n}}{\partial t}\right|\leqq C, \qquad 0\leqq t\leqq T$$

where $C>0$ is independent of T; therefore $\varphi_{\delta,n}$ are uniformly bounded in $C^\alpha(\Gamma\times[0,T])$ for any $\alpha\in(0,1)$.

It follows that

$$\varphi_{\delta,n}\to\psi \quad \text{uniformly in } C^\alpha(\Gamma\times[0,T]). \tag{4.11}$$

But for any $\zeta\in H^1(\Omega)$ with $\zeta=0$ on S we have

$$\int_\Omega \nabla\zeta\cdot\nabla\varphi_\delta=-\int_\Gamma \zeta\varphi_{\delta,n}, \qquad 0\leqq t\leqq T;$$

then as $\delta\to 0$, in view of (4.8), we find that

$$\int_\Omega \nabla\zeta\cdot\nabla\varphi=-\int_\Gamma \zeta\psi, \qquad 0\leqq t\leqq T.$$

It follows from (4.8) that the normal derivative φ_n satisfies

$$\varphi_n=\psi \quad \text{a.e. on } \Gamma, \quad t\in(0,T),$$

but $\varphi_n\in C^\alpha(\Gamma)$, $0\leqq t\leqq T$; hence $\varphi_n\in C^\alpha(\Gamma\times[0,T])$.

Let

$$\begin{aligned} h_\delta(x,t)&=\sigma(x)+\int_0^t g((\varphi_{\delta,n}(x,\tau)-\varepsilon)^+)\,d\tau \\ &=\sigma(x)+\delta\sum_{i=0}^{m-1} g((\varphi_{\delta,n}(*)-\varepsilon)^+)+[t-m\delta]g((\varphi_{\delta,n}(*)-\varepsilon)^+) \end{aligned} \tag{4.12}$$

for $m\delta\leqq t\leqq(m+1)\delta$. We used $\varphi(*)$ to denote the values of the function at some intermediate points $t_i<t_i^*<t_{i+1}$. So

$$\begin{aligned} h^m(x)-h_\delta(x,t)=\delta\sum_0^{m-1}\{g((\varphi_n^i-\varepsilon)^+)-g((\varphi_{\delta,n}(*)-\varepsilon)^+)\} \\ -[t-m\delta]g((\varphi_{\delta,n}(*)-\varepsilon)^+). \end{aligned}$$

But from the properties of g, (2.7), we find that

$$|g((\varphi_n^i-\varepsilon)^+)-g((\varphi_{\delta,n}(*)-\varepsilon)^+)|\leqq k|\varphi_n^i-\varphi_{\delta,n}(*)|, \qquad 0\leqq i\leqq m-1,$$

$$g((\varphi_{\delta,n}(*)-\varepsilon)^+)\leqq k\varphi_{\delta,n}(*).$$

But from (4.10) and Lemmas 3.5 and 3.8, we obtain

$$\begin{aligned} &|\varphi_n^i - \varphi_{\delta,n}(*)| \leqq |\varphi_n^{i+1} - \varphi_n^i| \leqq C\delta, && 0 \leqq i \leqq m-1, \\ &|t - m\delta| g((\varphi_{\delta,n}(*) - \varepsilon)^+) \leqq \delta k G_*, && m\delta \leqq t \leqq (m+1)\delta; \end{aligned} \tag{4.13}$$

hence

$$|h^m(x) - h_\delta(x, t)| \leqq C\delta \quad \text{on } \Gamma \times [0, T]. \tag{4.14}$$

Clearly $h_\delta(x, t) \in C^\alpha(\Gamma \times [0, T])$, since $\varphi_{\delta,n} \in C^\alpha(\Gamma \times [0, T])$. Moreover

$$h_{\delta,t} = g((\varphi_{\delta,n} - \varepsilon)^+),$$

which means that

$$h_{\delta,t} \in C^\alpha(\Gamma \times [0, T]).$$

As $\delta \to 0$

$$g((\varphi_{\delta,n} - \varepsilon)^+) \to g((\varphi_n - \varepsilon)^+) \quad \text{uniformly in } C^\alpha(\Gamma \times [0, T]) \tag{4.15}$$

and therefore

$$\left.\begin{aligned} h_\delta(x, t) &\to h(x, t) \\ h_{\delta,t}(x, t) &\to h_t(x, t) \end{aligned}\right\} \quad \text{uniformly in } C^\alpha(\Gamma \times [0, T]). \tag{4.16}$$

So, since (4.15) holds, we obtain that

$$\{h^m(x)\}_{m=0}^M \to h(x, t) \quad \text{as } \delta \to 0 \tag{4.17}$$

and clearly (4.12) implies that

$$h(x, t) = \sigma(x) + \int_0^t g((\varphi_n - \varepsilon)^+)\, d\tau \quad \text{on } \Gamma \times [0, T], \tag{4.18}$$

so (iv) is satisfied. Also it follows from (4.16) that

$$h, h_t \in C^\alpha(\Gamma \times [0, T]).$$

So (i) is satisfied and we are left with the second part of (iii), which follows from the uniform convergence in (4.8), (4.16), from (3.27) and the assumptions on G. More specifically, since

$$\varphi_n^m = G(\varphi^m, h^m) \to G(\varphi . h) \quad \text{as } \delta \to 0 \tag{4.19}$$

where the limit is uniform on $\Gamma \times [0, T]$ and in view of (4.13), $\varphi_n^m \to \varphi_n$ and therefore $\varphi_n = G(\varphi, h)$ on $\Gamma \times [0, T]$, i.e. (iii) is satisfied. The proof of the uniqueness is very similar to that of Theorem 2.8.

For later references we collect some of the facts above.

LEMMA 4.3. *The weak solution* (φ, h) *satisfies*

(i) $\varphi(x, t_1) \leqq \varphi(x, t_2)$ *in* $\bar{\Omega}$ *if* $t_1 < t_2$,

(ii) $\varphi \in C^\alpha(\bar{\Omega} \times [0, \infty))$ *any* $\alpha \in (0, 1)$,

(iii) $\|\varphi\|_{H^1(\Omega)} \leqq C, \quad 0 \leqq t \leqq T$,

C independent of T,

(iv) $\|\nabla \varphi_t\|_{L^2(\Omega)} \leqq C, \quad 0 \leqq t \leqq T$,

C independent of T,

(v) $\sigma_* \leqq h(x, t) \leqq 1 + K_2/\varepsilon$.

Proof. (i) follows from (3.10). (ii) follows from (4.5) and (4.8). (4.6) and (4.8) imply (iii). (iv) is a consequence of (4.9). (v) follows from Lemma 3.3.

In order to obtain better regularity we assume, in addition, that

(4.20) The partial derivatives $G_s(s,p)$ and $G_p(s,p)$ are C^α functions (any $\alpha\in(0,1)$) on compact subsets of $[0,1]\times\mathbb{R}_+$.

Then we can assert that under the assumptions of Theorem 4.2 and (4.20) we have for $0<T<\infty$ that the weak solution is a smooth solution (in the sense of § 2).

THEOREM 4.4. *If in addition* (4.20) *holds then the weak solution* (φ, h) *is a smooth solution,* $\nabla\varphi$ *is continuous in* $\bar\Omega\times[0,T]$ *and moreover*

$$\varphi_t(\cdot,t)\in C^{1+\alpha}(\bar\Omega),\qquad 0\leqq t\leqq T,\tag{4.21}$$
$$\varphi_t\in C^\alpha(\bar\Omega\times[0,T]),$$

$$\varphi_{nt}\in C^\alpha(\Gamma\times[0,T]).\tag{4.22}$$

Proof. Let ω be a solution of

$$\begin{aligned}&\Delta\omega=0\quad\text{in }\Omega,\\ &\omega=0\quad\text{on }S,\\ &\omega_n-G_s(\varphi,h)\omega=G_p(\varphi,h)h_t\quad\text{on }\Gamma.\end{aligned}\tag{4.23}$$

As, for $t\in[0,T]$, $G_p(\varphi,h)h_t$ and $G_s(\varphi,h)$ are in $C^\alpha(\Gamma)$ then $\omega\in C^{1+\alpha}(\bar\Omega)$. Since the solution to (4.23) is unique ($G_s\geqq 0$) we find that $\omega=\varphi_t$ and therefore $\varphi_t\in C^{1+\alpha}(\bar\Omega)$, $0\leqq t\leqq T$. To show continuity in t let $0\leqq t_1$, $t_2\leqq T$ and set

$$\varphi_t^i=\varphi_t(x,t_i),\qquad h_t^i=h_t(x,t_i),$$
$$G_s^i=G_s(\varphi(x,t_i),h(x,t_i)),\qquad G_p^i=G_p(\varphi(x,t_i),h(x,t_i)),\quad i=1,2,$$

and let $v=\varphi_t^1-\varphi_t^2$. Then v satisfies $\Delta v=0$ in Ω, $v=0$ on S and

$$v_n-(G_s^1\varphi_t^1-G_s^2\varphi_t^2)=G_p^1h_t^1-G_p^2h_t^2,\qquad x\in\Gamma.$$

The last equality can be written as

$$G_s^1v=v_n-(G_p^1h_t^1-G_p^2h_t^2)-(G_s^1-G_s^2)\varphi_t^2.\tag{4.24}$$

If $v>0$ at some point then it attains its positive maximum on Γ where, by the maximum principle, $v_n<0$. So

$$0<G_s^1v<|G_p^1h_t^1-G_p^2h_t^2|+|G_s^1-G_s^2||\varphi_t^2|\tag{4.25}$$

and a similar argument at a point of negative minimum (on Γ) gives

$$0<-G_s^1v<|G_p^1h_t^1-G_p^2h_t^2|+|G_s^1-G_s^2||\varphi_t^2|.\tag{4.26}$$

By (4.25), (4.26) and (2.8) we have

$$|v|=|\varphi_t^1-\varphi_t^2|\leqq K(\|G_p^1h_t^1-G_p^2h_t^2\|_{L^\infty(\Omega)}+\|G_s^1-G_s^2\|_{L^\infty(\Omega)}\|\varphi_t^2\|_{L^\infty(\Omega)}),\tag{4.27}$$

and the right-hand side converges to zero as $|t_2-t_1|\to 0$; hence φ_t is uniformly continuous. Now we divide (4.27) by $|t_2-t_1|^\alpha$, $\alpha\in(0,1)$, use the Hölder continuity of G_p, G_s, (4.20), and of h_t and write

$$G_p^1h_t^1-G_p^2h_t^2=(G_p^1-G_p^2)h_t^1+(h_t^1-h_t^2)G_p^2.$$

It follows that

$$\frac{|\varphi_t^1-\varphi_t^2|}{|t_1-t_2|^\alpha}\leqq K,\tag{4.28}$$

where K depends on the data and on T and is uniform in $\bar{\Omega}$. Hence $\varphi_t \in C^\alpha([0, T])$ uniformly in $\bar{\Omega}$ for any $\alpha \in (0, 1)$. In (4.20) we used the fact that $\sigma_* \leqq h \leqq 1 + K_2/\varepsilon$, (v) in Lemma 4.3. Combining this with $\varphi_t \in C^{1+\alpha}(\bar{\Omega})$ leads to the second part of (4.21).

Since $\varphi_{nt} = (\partial/\partial t)\varphi_n = G_s\varphi_t + G_p h_t$ then by the continuity we have $\varphi_{nt} = \varphi_{tn}$. Moreover the right-hand side is Hölder on Γ, $0 \leqq t \leqq T$ and in view of the second part of (4.21) it is Hölder on $[0, T]$ uniformly on Γ; hence (4.22) follows.

It remains to prove the continuity of $\nabla\varphi$ in $\bar{\Omega} \times [0, T]$. Let e be a unit vector in some (spatial) direction and denote by φ_e the derivative of φ in this direction.

By contradiction, if

$$|\varphi_e(x, t_0 + \delta t) - \varphi_e(x_0, t_0)| \geqq \gamma > 0$$

for $x - x_0 = \delta x$ small, for some $(x_0, t_0) \in \bar{\Omega} \times [0, \infty)$, then rewriting and using the fact that $\varphi_e \in C^\alpha$, it is enough to assume that

$$|\varphi_e(x_0, t_0 + \delta t) - \varphi_e(x_0, t_0)| \geqq \gamma. \tag{4.29}$$

Let $\delta t = 1/m$ then since $\varphi_m(x) = \varphi(x, t_0 + 1/m)$ is bounded uniformly in $C^{1+\alpha}$ then there exists a subsequence, also denoted by m, such that $\varphi_m(x) \to \bar{\varphi}(x)$ in $C^{1+\alpha}$ but $\varphi(x, t)$ are in $C^\alpha(\bar{\Omega} \times [0, T])$, hence $\bar{\varphi}(x) = \varphi(x, t_0)$. But clearly $\varphi_{m,e}(x) \to \bar{\varphi}_e(x)$ uniformly hence $\varphi_{m,e}(x) = \varphi_e(x, t_0 + 1/m) \to \varphi_e(x, t_0)$ which is a contradiction to (4.29), at $x = x_0$.

This holds for any direction so $\nabla\varphi$ is continuous in $\bar{\Omega} \times [0, T]$.

If (4.20) holds then the solution satisfies the following corollary.

COROLLARY 4.5. *It holds that*

$$\varphi(x, t_1) < \varphi(x, t_2), \qquad x \in \Omega \cup \Gamma \quad \textit{if } t_1 < t_2; \tag{4.30}$$

if for some $\bar{x} \in \Gamma$, $r > 0$, $h(\bar{x}, r) > \sigma(\bar{x})$ *then*

$$h_t(\bar{x}, t) > 0 \quad \forall t \geqq r. \tag{4.31}$$

Proof. Since (φ, h) is a smooth solution, it follows from Corollary 2.7. It follows from (4.31) and (2.4) that

$$\varphi_n(\bar{x}, t) > \varepsilon \quad \forall t \geqq r. \tag{4.32}$$

5. The asymptotic limit. We consider the asymptotic behavior of the solution (φ, h) of (2.1)-(2.4). For simplicity we assume (4.20) and so this is the smooth solution. It turns out that the steady state limit (as $t \to \infty$), which can be obtained formally by setting $h_t = 0$ in (2.4) (or more precisely in (2.4)′-(2.4)″), can be characterized as a Signorini problem with an obstacle that depends on $\varphi(x, t)$. Moreover the interest in the limit from the practical point of view is obvious since the thickness of the paint coat after a long time can give an indication of the areas that will not be painted at all.

The steady-state problem associated with (2.1)-(2.4) is to find a pair $\{\bar{\varphi}(x), \bar{h}(x)\}$, $\bar{\varphi} \in C^1(\bar{\Omega})$, $\bar{h} \in L^\infty(\Gamma)$ and such that

$$\Delta\bar{\varphi} = 0 \quad \text{a.e. in } \Omega, \tag{5.1}$$

$$\bar{\varphi}_n = G(\bar{\varphi}, \bar{h}) \quad \text{a.e. on } \Gamma, \qquad \bar{\varphi} = 1 \quad \text{on } S, \tag{5.2}$$

$$\left.\begin{aligned} \bar{\varphi}_n(x) &= \varepsilon \quad \text{if } \bar{h}(x) > \sigma(x), \\ \bar{\varphi}_n(x) &\leqq \varepsilon \quad \text{if } \bar{h}(x) = \sigma(x), \end{aligned}\right\} \text{a.e. on } \Gamma. \tag{5.3}$$

If we denote by $(\varphi(x, t), h(x, t))$ the smooth solution of (2.1)-(2.4) and $(\bar{\varphi}(x), \bar{h}(x))$ the solution of (5.1)-(5.3) then we have the following theorem.

THEOREM 5.1. *$\varphi(x,t) \nearrow \bar\varphi(x)$ uniformly in $C^\alpha(\bar\Omega)$, any $\alpha \in (0,1)$, and weakly in $H^1(\Omega)$ and $h(x,t) \nearrow \bar h(x)$ pointwise on Γ as $t \to \infty$. Moreover $\bar h(x) \leqq K_2/\varepsilon$.*

Proof. First suppose that $h(x,t) > \sigma(x)$ for some $x \in \Gamma$, $t > 0$. Hence $\varphi_n(x,s) > \varepsilon$ for some $0 < s \leq t$. Let $\tau = \sup\{s \leqq t;\ \varphi_n(x,s) > \varepsilon\}$. Then $\varphi_n(x,\tau) \geqq \varepsilon$, and then it follows from (2.3) and (2.8) that

$$\varepsilon \leqq \varphi_n(x,\tau) \leqq G(1,h) < \frac{K_2}{h}. \tag{5.4}$$

Therefore $h(x,\tau) < K_2/\varepsilon$ and clearly $h(x,t) < K_2/\varepsilon$ too. By Lemma 4.3(ii) $\varphi(x,t)$ is uniformly bounded in $C^\alpha(\bar\Omega \times [0,\infty))$. Let $\{t_i\}$ be a sequence such that $t_i \to \infty$ as $i \to \infty$. Consider the sequence of harmonic functions $\varphi_i(x) = \varphi(x,t_i)$. This sequence converges uniformly in $\bar\Omega$ to $\varphi_*(x)$. Clearly $\varphi_*(x) = 1$ on S. By Lemma 4.3(iii) the $\varphi_i(x)$ are uniformly bounded in $H^1(\Omega)$ and hence $\varphi_i(x) \to \varphi_*(x)$ as $i \to \infty$ (a subsequence if necessary) weakly in $H^1(\Omega)$. Hence $\varphi_*(x)$ is a harmonic function in Ω. By (2.3) and (2.8)

$$0 < \varphi_n = G(\varphi,h) < G(1,\sigma_*) = G_* \quad \text{on } \Gamma,$$

which implies that $\sup|\varphi_n(x,t)| \leqq G_* < \infty$ where the supremum is in $(x,t) \in \Gamma \times [0,\infty)$. Therefore there exists a subsequence φ_{in} that converges to ψ_* weak star in $L^\infty(\Gamma)$, i.e. for every $\zeta \in H^1(\Omega)$, $\zeta = 0$ on S,

$$\int_\Gamma \zeta\varphi_{in} \to \int_\Gamma \zeta\psi_*. \tag{5.5}$$

But since

$$\int_\Gamma \zeta\varphi_{in} = -\int_\Omega \nabla\zeta \cdot \nabla\varphi_i \to -\int_\Omega \nabla\zeta \cdot \nabla\varphi_* \tag{5.6}$$

it follows from (5.5) and (5.6) that ψ_* is a generalized (inward) normal derivative of φ_*, so we denote it by φ_{*n}. Now it is clear that $h_i(x) \equiv h(x,t_i) \geqq h_{i-1}(x)$ and since $h_i(x) < K_2/\varepsilon$ it follows that $h_i(x) \nearrow \bar h(x)$ pointwise in Γ and $\bar h(x) \leqq K_2/\varepsilon$. But then, using Lebesgue's theorem we obtain that

$$\int_\Gamma \zeta G(\varphi_i, h_i) \to \int_\Gamma \zeta G(\varphi_*, \bar h) \tag{5.7}$$

where (2.8) was used, and this in turn together with (5.6) implies that

$$\varphi_{*n} = G(\varphi_*, \bar h) \quad \text{a.e. on } \Gamma. \tag{5.8}$$

We can identify $\varphi_*(x)$ with $\bar\varphi(x)$, the unique solution of (5.1)-(5.2). Indeed if φ_1 and φ_2 satisfy (5.1), (5.2) with the same $\bar h$ then, from the monotonicity of G with φ, it follows that if we multiply $\Delta(\varphi_1 - \varphi_2) = 0$ by $(\varphi_1 - \varphi_2)$ and integrate over Ω we find

$$\begin{aligned} \int_\Omega |\nabla(\varphi_1 - \varphi_2)|^2 &= -\int_\Gamma (\varphi_{1n} - \varphi_{2n})(\varphi_1 - \varphi_2) \\ &= -\int_\Gamma (G_1 - G_2)(\varphi_1 - \varphi_2) \leqq 0, \end{aligned} \tag{5.9}$$

i.e., $\varphi_1 \equiv \varphi_2$ since $\varphi_1 = \varphi_2 = 1$ on S. It remains to show (5.3).

If $h_i(x) = \sigma(x)$ for some $x \in \Gamma$, all $i \geqq 1$, then from (2.4) it follows that $\varphi_{in}(x) \leqq \varepsilon$ all $i \geqq 1$, and by the pointwise convergence $\bar h(x) = \sigma(x)$ and $\bar\varphi_n \leqq \varepsilon$ a.e. on $\Gamma \cap \{\bar h(x) = \sigma(x)\}$. On the other hand if $h_i(x) > \sigma(x)$ for some $i > 1$ then by (4.32) $\varphi_{jn}(x) > \varepsilon$ all $j \geqq i$ but since the integral in (2.4) converges uniformly $(\bar h(x) \leqq K_2/\varepsilon)$ hence $\varphi_{in}(x) \searrow \varepsilon$ as $i \to \infty$ a.e. on $\Gamma \cap \{\bar h(x) > \sigma(x)\}$ thus (5.3) is satisfied a.e. on Γ.

We have

LEMMA 5.2. *The solution to* (5.1)-(5.3) *is unique.*

Proof. Let (φ_1, h_1) and (φ_2, h_2) be two solutions. As in (5.9) above we find that

$$\int_\Omega |\nabla(\varphi_1-\varphi_2)|^2 = -\int_\Gamma (G(\varphi_1, h_1)-G(\varphi_2, h_2))(\varphi_1-\varphi_2). \tag{5.10}$$

It is enough to show that the right-hand side of (5.10) is nonpositive. Indeed if $h_1, h_2 > \sigma(x)$ then $\varphi_{1n}(x) = \varepsilon = \varphi_{2n}(x)$.

If $h_1(x) > \sigma(x)$ and $h_2(x) = \sigma(x)$ then $\varphi_{1n}(x) = \varepsilon$ and $\varphi_{2n}(x) \leqq \varepsilon$, and this means that $G_1 \equiv G(\varphi_1, h_1) = \varepsilon \geqq G(\varphi_2, h_2) \equiv G_2$. So we want to show that $\varphi_1 \geqq \varphi_2$. Assume that $\varphi_1(x) < \varphi_2(x)$ at $x \in \Gamma$, then

$$\begin{aligned} 0 \geqq (G_1-G_2)(\varphi_1-\varphi_2) &= (G(\varphi_1, h_1)-G(\varphi_2, h_1))(\varphi_1-\varphi_2) \\ &\quad + (G(\varphi_2, h_1)-G(\varphi_2, h_2))(\varphi_1-\varphi_2) \\ &> (G(\varphi_2, h_1)-G(\varphi_2, h_2))(\varphi_1-\varphi_2), \end{aligned} \tag{5.11}$$

since G is strictly monotone increasing in φ, but G is monotone decreasing in h and $h_1 > h_2$; hence $G(\varphi_2, h_1)-G(\varphi_2, h_2) \leqq 0$ and so the assumption $(\varphi_1-\varphi_2) < 0$ leads to a contradiction in (5.11). Hence $\varphi_1 - \varphi_2 \geqq 0$ and therefore $(G_1-G_2)(\varphi_1-\varphi_2) \geqq 0$. The argument in the case when $h_1 = \sigma(x)$ and $h_2 > \sigma(x)$ is similar. Finally when $h_1 = \sigma(x) = h_2$ then $(G_1-G_2)(\varphi_1-\varphi_2) \geqq 0$ follows from the monotonicity of G in φ.

From Theorem 5.1 and Lemma 5.2 we have the following theorem.

THEOREM 5.3. *Assume that* $\partial\Omega$ *is in* $C^{1+\alpha}$, *any* $\alpha \in (0, 1)$, (2.6)-(2.8) *and* (4.20) *hold. Then there exists a unique solution* $(\bar\varphi, \bar h)$ *to the problem* (5.1)-(5.3). *Moreover*

$$\bar\varphi \in C^{1+\alpha}(\bar\Omega), \tag{5.12}$$

$$\bar h \in C^\alpha(\Gamma) \quad \textit{for any } \alpha \in (0, 1). \tag{5.13}$$

Proof. The uniqueness follows from Lemma 5.2. In addition we have that $\bar\varphi \in H^1(\Omega)$. Recall that $\sigma_* \leqq \bar h \leqq K_2/\varepsilon$ and $0 < c_* = \min_\Gamma \varphi(x, 0) \leqq \min_{\Gamma\times[0,\infty)} \varphi(x, t)$ (see (2.12)). Define

$$\Gamma^+ = \{x \in \Gamma, \bar h(x) > \sigma(x)\}, \tag{5.14}$$

then $\Gamma^+ \neq \varnothing$ by the assumption (2.11) and is open in Γ by the continuity of $h(x, t)$ and the fact that $h(x, t)$ is monotone nondecreasing in t on Γ^+. Let $x, y \in \Gamma^+$ then $G = \varepsilon$ a.e. at these points and hence

$$\begin{aligned} 0 &= G(\bar\varphi(x), \bar h(x)) - G(\bar\varphi(y), \bar h(y)) \\ &= G_\varphi(*)(\bar\varphi(x)-\bar\varphi(y)) + G_h(*)(\bar h(x)-\bar h(y)) \quad \text{a.e. on } \Gamma^+ \end{aligned} \tag{5.15}$$

where $G_\varphi(*)$ and $G_h(*)$ were evaluated by the mean value theorem. Now since $c_* > 0$ it follows from (2.8) that $0 < k' < |G_h(*)| < k''$ and therefore (5.15) can be rewritten as

$$k'|\bar h(x)-\bar h(y)| \leqq K_1|\bar\varphi(x)-\bar\varphi(y)| \quad \text{a.e. on } \Gamma^+. \tag{5.16}$$

But the right-hand side is Hölder continuous with exponent α in $\bar\Gamma^+$ hence so is the left-hand side, i.e. $\bar h(x) \in C^\alpha(\bar\Gamma^+)$. Therefore $\bar h \in C^\alpha(\Gamma)$ since on $\Gamma\backslash\Gamma^+$ we have $h(x) = \sigma(x)$. But then, since $\bar\varphi \in C^\alpha(\bar\Omega)$, we obtain by (5.2) that $\bar\varphi_n \in C^\alpha(\Gamma)$.

Let ω be a solution of

$$\begin{aligned} \Delta\omega &= 0 \quad \text{in } \Omega, \\ \omega &= 1 \quad \text{on } S, \\ \omega_n &= \bar\varphi_n \quad \text{on } \Gamma. \end{aligned} \tag{5.17}$$

Then $\omega \in C^{1+\alpha}(\bar{\Omega})$ and since the solution to (5.17) is unique we find that $\omega = \bar{\varphi}$ and therefore $\bar{\varphi} \in C^{1+\alpha}(\bar{\Omega})$, any $\alpha \in (0, 1)$.

We are able to show that the solution $(\bar{\varphi}, \bar{h})$ of (5.1)-(5.3) is also a solution of a Signorini problem (see e.g. [5, p. 111]). To this end let the obstacle be defined by

$$(5.18) \qquad Z(x) = \lim_{t \nearrow \tau(x)} \varphi(x, t), \qquad x \in \Gamma$$

where

$$(5.19) \qquad \tau(x) = \sup\{0 \leqq t;\ h(x, t) = \sigma(x)\}, \qquad x \in \Gamma.$$

Since h is monotone nondecreasing in t, $\tau(x)$ is well defined and it can have the value $+\infty$.

We have the following lemma.

LEMMA 5.4. *$\tau(x)$ is continuous from above on Γ^+.*

Proof. Clearly $\Gamma^+ = \{x \in \Gamma, \tau(x) < \infty\}$, since if $x \in \Gamma^+$ then $\bar{h}(x) > \sigma(x)$. Then for some $t < \infty$, $h(x, t) > \sigma(x)$, hence $\tau(x) < t < \infty$ and vice versa.

Now let $x_0 \in \Gamma^+$, $\tau(x_0) = M$ and by (5.19) $h(x_0, \tau(x_0)) = \sigma(x_0)$ but $h(x_0, t) > \sigma(x_0)$ for all $t > \tau(x_0)$. Consider the solution (φ, h) in $0 \leqq t \leqq 2M$ and, by Theorem 4.2 $h \in C^\alpha(\Gamma \times [0, 2M])$. Then for any x in a δ-neighbourhood $N_\delta \subset \Gamma$ of x_0, $\tau(x) < 3M/2$, $h(x, t) > \sigma(x)$ for $t > \tau(x)$, $x \in N_\delta$. Also, if $\tau(x) > \tau(x_0)$, we have

$$h(x_0, \tau(x)) - h(x, \tau(x)) = \sigma(x_0) - \sigma(x) + \int_{\tau(x_0)}^{\tau(x)} g(\varphi_n(x_0, \tau) - \varepsilon)\, d\tau$$

where we used (2.4) and the fact that $\varphi_n > \varepsilon$ for $t > \tau(x_0)$ at x_0, by (4.32). Hence

$$0 < \int_{\tau(x_0)}^{\tau(x)} g(\varphi_n(x_0, \tau) - \varepsilon)\, d\tau \leqq |\sigma(x) - \sigma(x_0)| + |h(x_0, \tau(x)) - h(x, \tau(x))|,$$

but the right-hand side converges to zero as $x \to x_0$. Hence, since $g > 0$ for $\varphi_n(x_0, t) > \varepsilon$, $t > \tau(x_0)$, it follows that $\tau(x) \searrow \tau(x_0)$.

Define

$$(5.20) \qquad \mathbb{K} = \{\zeta \in H^1(\Omega);\ \zeta = 1 \text{ on } S,\ \zeta \geqq Z \text{ on } \Gamma\},$$

then we prove the following.

THEOREM 5.5. *If (4.20) is satisfied then $(\bar{\varphi}, \bar{h})$ is a solution to (5.1)-(5.3) if and only if $\bar{\varphi}$ satisfies the variational inequality*

$$(5.21) \qquad \int_\Omega \nabla\bar{\varphi} \cdot \nabla(\zeta - \bar{\varphi}) + \varepsilon \int_\Gamma (\zeta - \bar{\varphi}) \geqq 0 \quad \forall \zeta \in \mathbb{K}, \quad \bar{\varphi} \in \mathbb{K}.$$

Remark 5.6. The fact that $(\bar{\varphi}, \bar{h})$ is a solution to a variational inequality is a result of its being the asymptotic limit of an evolutionary problem and indeed the obstacle $Z(x)$ is constructed from the time dependent solution. It may be interesting to investigate what classes of problems of the type (5.1)-(5.3) can be recast as variational inequalities via their imbedding in appropriate evolutionary problems.

Proof. It is enough to show that (5.21) is satisfied by $\varphi(x, t)$ as $t \to \infty$, where (φ, h) is the solution to (2.1)-(2.4). For any $\zeta \in \mathbb{K}$, $0 < t < \infty$, we have

$$(5.22) \quad \int_\Omega \nabla\varphi(x, t) \cdot \nabla(\zeta - \varphi) + \varepsilon \int_\Gamma (\zeta - \varphi) \geqq -\int_\Gamma (\varphi_n - \varepsilon)(\zeta - \varphi) \geqq -\int_{\Gamma_t} (\varphi_n - \varepsilon)(\zeta - \varphi)$$

where $\Gamma_t = \Gamma \cap \{x;\ \varphi_n(x, t) > \varepsilon\}$ and Lemma 4.5 was used. Clearly $\Gamma_t \nearrow \Gamma_\infty$ if $t \to \infty$ by (4.31) and so for any $\eta > 0$ small and $M > 0$ large there is a θ (depending only on η,

$\theta \to \infty$ if $\eta \to 0$) such that

$$\int_{\theta}^{\theta+M} \int_{\Gamma_t} |\varphi_n - \varepsilon| = \int_{\theta}^{\theta+M} \int_{\Gamma_t} (\varphi_n - \varepsilon)^+$$
$$\leqq \frac{1}{k_*} \int_{\theta}^{\theta+M} \int_{\Gamma_\theta} g((\varphi_n - \varepsilon)^+) + \eta M \tag{5.23}$$
$$\leqq K \int_{\Gamma_\theta} h(x, \theta + M) + \eta M \leqq \frac{K}{\varepsilon} + \eta M$$

where we used (2.7), the fact that $h \leqq K_2/\varepsilon$ and that $\Gamma_\infty \backslash \Gamma_\theta \to 0$ if $\theta \to \infty$, with an appropriate K.

We integrate (5.22) with respect to $\theta < t < \theta + M$, divide by M, use (5.23) and let $M \to \infty$, $\eta \to 0$. We find that $\bar{\varphi}(x) \overset{=}{\cdot} \lim_{t\to\infty} \varphi(x, t)$ satisfies (5.21).

The fact that $\bar{\varphi}(x) \in H^1(\Omega)$ follows from (5.12) and that $\bar{\varphi} = 1$ on S is clear. From the definition of $Z(x)$, (5.18), and from $\varphi(x, t_1) \leqq \varphi(x, t_2)$ for every $t_1 < t_2$ it follows that $\bar{\varphi} \geqq Z$ on Γ, therefore $\bar{\varphi} \in \mathbb{K}$, as was claimed above.

Remark 5.7. It follows from Theorems 5.3 and 5.5 that the solution $\bar{\varphi}$ to the Signorini problem (5.21) is in $C^{1+\alpha}(\bar{\Omega})$ for any $\alpha \in (0, 1)$.

6. A weak solution when $\sigma(x) \equiv 0$. Most of the bounds that were found in § 4 depend on the fact that $0 < \sigma_* \leqq \sigma(x)$. Nevertheless using convex geometry and a monotonicity argument we are able to prove the existence of a weak solution to the degenerate case when $\sigma(x) = 0$ on Γ; i.e. the case that was considered in [1] and [2]. To this end we invert the geometry and assume that S is the inner boundary of Ω and Γ is the outer boundary, both in $C^{1+\alpha}$, any $\alpha \in (0, 1)$ and that Γ is *convex*. For simplicity we take $g(s) = s$, $G(\varphi, h) = \varphi/h$ which satisfies (2.8) and (4.20) for $0 < \varphi \leqq 1$ and $\sigma_* \leqq h \leqq 1 + \varepsilon^{-1}$. In addition we take $\sigma(x) = \sigma$ where $\sigma \geqq 0$ is a constant. Thus we consider the problem

$$\Delta\varphi = 0 \quad \text{in } \Omega, \quad 0 \leqq t \leqq T, \tag{6.1}$$

$$\varphi = 1 \quad \text{on } S, \quad 0 \leqq t \leqq T, \tag{6.2}$$

$$h\varphi_n = \varphi \quad \text{on } \Gamma, \quad 0 \leqq t \leqq T, \tag{6.3}$$

$$h = \sigma + \int_0^t (\varphi_n - \varepsilon)^+ \, d\tau \quad \text{on } \Gamma, \quad 0 \leqq t \leqq T. \tag{6.4}$$

We refer to (6.1)-(6.4) as problem (P_σ). Our interest is in (P_0), which is the problem in the limit as $\sigma \to 0$. It follows from Theorem 4.4 that for any $\sigma > 0$ the solution to (P_σ) is smooth. On the other hand the weak solution to (P_0) is essentially weaker than in definition 4.1, as the following definition shows.

DEFINITION 6.1. A *weak solution to problem* (P_0) is a pair of functions (φ, h) such that for any $0 < T < \infty$

(i) $\varphi \in L^\infty(0, T; H^1(\Omega)) \cap C^\alpha(\bar{\Omega} \times [0, T])$,

$h \in L^\infty(\Gamma \times (0, T))$,

$h_t \in L^\infty(0, T; L^2(\Gamma))$,

for any $\alpha \in (0, 1)$;

(ii) $\Delta\varphi = 0$ in Ω, $0 \leqq t \leqq T$,

$\varphi = 1$ on S, $0 \leqq t \leqq T$;

(iii) There exists a function $\varphi_n \in L^\infty(0, T; L^2(\Gamma))$ such that

$$\int_\Omega \nabla\zeta \cdot \nabla\varphi = -\int_\Gamma \zeta\varphi_n, \qquad 0 \leqq t \leqq T$$

for any smooth function ζ such that $\zeta = 0$ in a neighborhood of S, thus φ_n can be considered as a generalized (inward) normal derivative of φ;

(iv) $\quad h\varphi_n = \varphi$ a.e. on Γ, $0 \leqq t \leqq T$;

(v) $\quad h = \int_0^t (\varphi_n - \varepsilon)^+ \, d\tau$ a.e. on Γ, $0 \leqq t \leqq T$.

We want to prove the existence of a weak solution to (P_0) as a monotone limit of (smooth) solutions to (P_σ), $\sigma > 0$. To this end we need the following a priori estimates.

LEMMA 6.2. *Let ω be the solution of*

$$(6.5) \qquad \Delta\omega = 0 \quad in\ \Omega, \qquad \omega = 1 \quad on\ S, \qquad \sigma\omega_n = \omega \quad on\ \Gamma.$$

Then there exists a constant $\theta > 0$ that depends only on the geometry but is independent of σ, such that

$$(6.6) \qquad 0 \leqq \omega_n < \theta \quad on\ \Gamma$$

for any $\sigma > 0$.

Proof. From the maximum principle $\omega > 0$ and $\omega_n > 0$ on Γ. Let $x_0 \in \Gamma$ be the point where $\max_\Gamma \omega_n$ is attained; at this point $\max_\Gamma \omega$ is attained too. Let $\Pi(x)$ be the hyperplane that passes through the point $(x_0, \omega(x_0))$, i.e. $\Pi(x_0) = \omega(x_0)$, and has a slope $\theta = 2/d_0$ in the direction of the inner normal, where $d_0 = \text{dist}(\Gamma, S)$. From the assumption on the convexity of Γ we have that $\Pi(x) \geqq \omega(x)$ on Γ and by the construction $\Pi(x) > 1$ on S. Hence $\Pi > \omega$ in Ω and therefore by the strong maximum principle, since $\Pi = \omega$ at x_0, $\theta = \Pi_n > \omega_n$ at x_0, i.e., (6.6) holds.

We use this to show the following.

LEMMA 6.3. *There holds for any $\sigma > 0$*

$$(6.7) \qquad \int_\Gamma |\varphi_n|^3 \leqq c, \qquad 0 \leqq t \leqq T$$

where $C > 0$ depends on Γ, θ and ε but is independent of σ or T.

Proof. Let (φ, h) be the solution to (P_σ). Then $\varphi_t \in C^\alpha(\bar{\Omega} \times [0, T])$ by Theorem 4.4 and satisfies

$$(6.8) \qquad \Delta\varphi_t = 0 \qquad \text{in } \Omega, \quad 0 \leqq t \leqq T,$$

$$(6.9) \qquad \varphi_t = 0 \qquad \text{on } S, \quad 0 \leqq t \leqq T,$$

$$(6.10) \qquad h\varphi_{tn} = \varphi_t - h_t\varphi_n \qquad \text{on } \Gamma, \quad 0 \leqq t \leqq T.$$

We multiply (6.8) by φ_t, integrate over Ω, integrate by parts and use (6.9) to obtain (for any $t \in [0, T]$)

$$(6.11) \qquad 0 \leqq \int_\Omega |\nabla\varphi_t|^2 = -\int_\Gamma \varphi_t\varphi_{tn}.$$

Now we multiply (6.10) by φ_{tn} and integrate over Γ. Using (6.11) and the equation $h_t = (\varphi_n - \varepsilon)^+$ we get

$$(6.12) \qquad \int_\Gamma (\varphi_n - \varepsilon)\varphi_n\varphi_{nt} \leqq \int_\Gamma (\varphi_n - \varepsilon)^+\varphi_n\varphi_{nt} \leqq \int_\Gamma \varphi_t\varphi_{tn} \leqq 0$$

where the left inequality follows from the observation that if $\varphi_n \leqq \varepsilon$ then $h_t = 0$. Hence from (6.10) it follows that $\varphi_{tn} > 0$. We rewrite (6.12) as

$$\int_\Gamma (\varphi_n - \varepsilon)^2 \varphi_{nt} + \varepsilon \int_\Gamma (\varphi_n - \varepsilon)\varphi_{nt} \leqq 0,$$

i.e.,

$$\frac{d}{dt}\int_\Gamma \left\{\frac{1}{3}(\varphi_n - \varepsilon)^3 + \frac{\varepsilon}{2}(\varphi_n - \varepsilon)^2\right\} \leqq 0$$

or

$$\begin{aligned}(6.13)\qquad &\int_\Gamma \left\{\frac{1}{3}(\varphi_n(x,t) - \varepsilon)^3 + \frac{\varepsilon}{2}(\varphi_n(x,t) - \varepsilon)^2\right\} dx \\ &\qquad \leqq \int_\Gamma \left\{\frac{1}{3}(\varphi_n(x,0) - \varepsilon)^3 + \frac{\varepsilon}{2}(\varphi_n(x,0) - \varepsilon)^2\right\} dx.\end{aligned}$$

Now clearly $\varphi(x, 0) = \omega(x)$ where ω is the solution of (6.5) and therefore $\varphi_n(x, 0) \leqq \theta$ by (6.6) and so (6.13) implies (6.7), and also

$$(6.14)\qquad \int_\Gamma \varphi_n^2 \leqq C.$$

LEMMA 6.4. *There holds for any* $\sigma > 0$

$$(6.15)\qquad \|\varphi\|_{H^1(\Omega)} \leqq C, \qquad 0 \leqq t \leqq T,$$

where C is independent of σ and of T.

Proof. We integrate $(\varphi - 1)\Delta\varphi$ over Ω and find

$$\int_\Omega |\nabla\varphi|^2 = \int_\Gamma (1-\varphi)\varphi_n \leqq \frac{1}{2}\int_\Gamma (1-\varphi)^2 + \frac{1}{2}\int_\Gamma \varphi_n^2 \leqq C$$

where we used $0 \leqq \varphi \leqq 1$ and (6.14), for any $0 \leqq t \leqq T$.

We denote by $\{\varphi_\sigma, h_\sigma\}$ the smooth solution to (P_σ), $\sigma > 0$. We have that the sequences $\{\varphi_\sigma\}$ and $\{h_\sigma\}$ are monotone increasing with $\sigma > 0$.

LEMMA 6.5. *Let $0 < \sigma_1 < \sigma_2$ and let $\varphi_{\sigma_i} \equiv \varphi_i$, $h_{\sigma_i} \equiv h_i$, $i = 1, 2$ be the smooth solutions to (P_{σ_i}). Then*

$$(6.16)\qquad \begin{aligned}&\varphi_1 \leqq \varphi_2 \quad \textit{in } \Omega \cup \Gamma, \quad 0 \leqq t \leqq T,\\ &h_1 \leqq h_2 \quad \textit{on } \Gamma, \qquad 0 \leqq t \leqq T.\end{aligned}$$

Proof. Define the set $\Sigma = \{t \in [0, T];$ (6.16) holds for $t\}$. We have that $0 \in \Sigma$. Indeed at $t = 0$ we have $h_1 = \sigma_1 < \sigma_2 = h_2$ by assumption. Let $\varphi = \varphi_2 - \varphi_1$, then $\Delta\varphi = 0$ in Ω and $\varphi = 0$ on S and

$$(6.17)\qquad \varphi = \varphi_2 - \varphi_1 = \sigma_2\varphi_{2n} - \sigma_1\varphi_{1n} \quad \text{on } \Gamma.$$

Let $x_0 \in \Gamma$ be a nonpositive minimum point of φ; then by the strong maximum principle $\varphi_n > 0$ at x_0, and so $\varphi_{2n} > \varphi_{1n}$ which is impossible by (6.17). So $\min_\Gamma \varphi > 0$ and hence

$$(6.18)\qquad \begin{aligned}&\varphi_1 < \varphi_2 \quad \text{in } \Omega \cup \Gamma, \quad t = 0,\\ &h_1 < h_2 \quad \text{on } \Gamma, \qquad t = 0.\end{aligned}$$

By continuity there exists $t_* > 0$ such that $[0, t_*) \subset \Sigma$. If $t_* < T$ then there exists $x_0 \in \Gamma$ such that $h_1(x_0, t_*) = h_2(x_0, t_*)$. Assume that $h_1 \not\equiv h_2$ at $t = t_*$. By the proof of (6.18) we have $\varphi_1(x, t_*) < \varphi_2(x, t_*)$ in $\Omega \cup \Gamma$. Then it follows from

$$0 < \varphi = \varphi_2 - \varphi_1 = h_2\varphi_{2n} - h_1\varphi_{1n}, \qquad x_0 \in \Gamma, \quad t = t_*$$

that $\varepsilon < \varphi_{1n} < \varphi_{2n}$; hence $h_{1_t} = (\varphi_{1n} - \varepsilon) < (\varphi_{2n} - \varepsilon) = h_{2t}$ at $x_0 \in \Gamma$, $t = t_*$ but this implies, by continuity, that $h_2(x_0, t) < h_1(x_0, t)$ for $t_* - \delta < t < t_*$, some $\delta > 0$, which is a contradiction to the definition of t_*. If $h_1 \equiv h_2$ at t_*, then if $\varphi_1 \not\equiv \varphi_2$ the argument is as above. If $\varphi_1 \equiv \varphi_2$ at $t = t_*$ then this holds for every $t \geqq t_*$ because of the uniqueness, and (6.16) holds trivially on $[t_*, T]$.

Now we can assert that problem (P_0) has a weak solution (as in Definition 6.1). Thus the problem that was considered in [1] and in [2] has a weak solution.

THEOREM 6.6. *There exists a weak solution* (φ, h) *to problem* (P_0). *Moreover* $0 \leqq h \leqq 1 + \varepsilon^{-1}$.

Proof. Let $\{\varphi_\sigma, h_\sigma\}$ be the solutions to (P_σ) with $\sigma > 0$. It follows from Lemma 3.4 that $\{\varphi_\sigma\}$ is bounded uniformly in $C^\alpha(\bar{\Omega} \times [0, T])$ independently of $\sigma > 0$, and Lemma 6.4 implies that the same holds true in $L^\infty(0, T; H^1(\Omega))$. For all $\sigma > 0$, $\sigma \leqq h_\sigma \leqq 1 + 1/\varepsilon$, from Lemma 4.3 (v) with $K_2 = 1$. Therefore as we let $\sigma \to 0$ we obtain

$$(6.19) \qquad \varphi_\sigma \searrow \varphi \begin{cases} \text{uniformly in } C^\alpha(\bar{\Omega} \times [0, T]), \\ \text{weakly in } H^1(\Omega), 0 \leqq t \leqq T, \\ \text{monotonically on } \bar{\Omega} \times [0, T] \end{cases}$$

and

$$(6.20) \qquad h_\sigma \searrow h \begin{cases} \text{weak* in } L^\infty(\Gamma \times (0, T)), \\ \text{monotonically on } \Gamma \times [0, T]. \end{cases}$$

From the uniform convergence $\varphi_0 \searrow \varphi$ we have $\varphi = 1$ on S, $0 \leqq t \leqq T$ and φ is harmonic in Ω, $0 \leqq t \leqq T$, thus parts of (i) and (ii) of Definition 6.1 are satisfied.

It follows from Lemma 6.3 that $\{\varphi_{\sigma,n}\}$ is bounded in $L^3(\Gamma)$, $0 \leqq t \leqq T$, independently of $\sigma > 0$. Thus $\varphi_{\sigma,n} \to \psi$ weakly in $L^3(\Gamma)$, $0 \leqq t \leqq T$ and therefore, by Sobolev's theorem (see e.g. [8]),

$$(6.21) \qquad \varphi_{\sigma,n} \to \psi \quad \text{strongly in } L^2(\Gamma), \quad 0 \leqq t \leqq T.$$

Let ζ be a smooth function such that $\zeta = 0$ on S, then

$$(6.22) \qquad \int_\Omega \nabla\zeta \cdot \nabla\varphi_\sigma = -\int_\Gamma \zeta\varphi_{\sigma,n}, \qquad 0 \leqq t \leqq T$$

for all $\sigma > 0$. Then (6.19)) and (6.21) give, as $\sigma \to 0$, that

$$(6.23) \qquad \int_\Gamma \nabla\zeta \cdot \nabla\varphi = -\int_\Gamma \zeta\psi, \qquad 0 \leqq t \leqq T.$$

Therefore, since $\Delta\varphi = 0$ in Ω, ψ can be identified with a generalized (inward) normal derivative of φ, so we write $\varphi_n = \psi$ where (6.23) is understood. Thus (iii) is satisfied. Letting $\zeta \in C_0^\infty(\mathbb{R}^n)$, we multiply (6.3) by ζ and integrate over Γ and then let $\sigma \to 0$, then (6.19)-(6.21) give

$$(6.24) \qquad \int_\Gamma \zeta h\varphi_n = \int_\Gamma \zeta\varphi, \qquad 0 \leqq t \leqq T,$$

and so (iv) is satisfied, but moreover $h\varphi_n \in C^\alpha(\Gamma \times [0, T])$. Finally we multiply (6.4) by ζ and integrate over Γ, so

$$\int_\Gamma \zeta h_\sigma = \sigma \int_\Gamma \zeta + \int_\Gamma \int_0^t \zeta(\varphi_{\sigma,n} - \varepsilon)^+ \, d\tau, \qquad 0 \leqq t \leqq T.$$

Letting $\sigma \to 0$ and using (6.20) and (6.21) we get

$$\int_\Gamma \zeta h = \int_\Gamma \int_0^t \zeta(\varphi_n - \varepsilon)^+ \, d\tau, \qquad 0 \leqq t \leqq T, \tag{6.25}$$

and so (v) is satisfied. Moreover by differentiating with respect to t in (6.25),

$$h_t = (\varphi_n - \varepsilon)^+ \quad \text{a.e. on } \Gamma, \quad 0 \leqq t \leqq T,$$

and therefore $h_t \in L^\infty(0, T; L^2(\Gamma))$. This completes (i) and therefore the proof of the theorem.

Acknowledgments. The second author would like to thank the Instituto Argentino de Matemática and the Departmento de Matemática in the Facultad de Ingeniería for their hospitality, with special thanks to J. E. Bouillet, P. M. Jacovkis and E. Lami Dozo.

REFERENCES

[1] J. M. Aitchison, A. A. Lacey and M. Shillor, *A model for an electropaint process*, IMA J. Appl. Math., 33 (1984), pp. 17-31.

[2] L. A. Caffarelli and A. Friedman, *A nonlinear evolution problem associated with an electropaint process*, this Journal, 16 (1985), pp. 955-969.

[3] E. Di Benedetto and A. Friedman, *The ill posed Hele-Shaw model and the Stefan problem for supercooled water*, Trans. Amer. Math. Soc., 282 (1984), pp. 183-204.

[4] C. M. Elliott and J. R. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Pitman, London, 1982.

[5] A. Friedman, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[6] E. B. Hansen and J. A. McGeough, *On electropainting*, SIAM J. Appl. Math., 43 (1983), pp. 627-638.

[7] A. A. Lacey, *Tool design for electrochemical machining in the presence of overpotentials*, preprint. IMA J. Appl. Math., 35 (1986), pp. 357-364.

[8] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[9] V. G. Levich, *Physiochemical Hydrodynamics*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[10] J. A. McGeough, *Principles of Electrochemical Machining*, Chapman Hall, London, 1974.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

017

ASYMPTOTICS OF A RATHER UNUSUAL TYPE IN A FREE BOUNDARY PROBLEM*

C. M. BRAUNER†, W. ECKHAUS‡, M. GARBEY§ AND A. VAN HARTEN‡

Abstract. We consider the singular perturbation problem

$$-\varepsilon\Delta u+\frac{u}{1+u}=f \quad \text{in } \Omega\in\mathbb{R}^2$$

with $u=0$ on $\partial\Omega$. There is a region where u blows up as $\varepsilon\downarrow 0$, whose boundary S is solution of a free boundary problem. Across S there is a sharp transition layer phenomenon. We construct a uniform asymptotic expansion in $\bar{\Omega}$ which exhibits two rather unusual features: the regular expansion in the subdomain inside S has singularities as S is approached; the structure of the local expansion along S involves fractional powers of ε multiplied by powers of $\ln \varepsilon$. We prove the validity of the expansion using barrier-function techniques.

Key words. singular perturbations, free boundary, matched asymptotic expansions, nonlinear elliptic boundary value problems

AMS(MOS) subject classifications. 35B25, 35J75

1. Introduction. In this paper we take up a problem that has been studied by Brauner and Nicolaenko [5], [6], [7] and analyse it by the method of matched asymptotic expansions [1]. Let us first formulate the problem, then summarize some relevant results and show why further analysis is needed and is nontrivial.

Let the function u be a solution of

$$\begin{aligned}&-\varepsilon\Delta u+\frac{u}{1+u}=f(x) \quad \text{in } \Omega\subset\mathbb{R}^2,\\ &u=0 \quad \text{on } \partial\Omega.\end{aligned} \tag{1.1}$$

Ω is a bounded open domain without holes, the boundary $\partial\Omega$ is smooth and connected and Ω lies everywhere on one side of $\partial\Omega$. The function $f(x)$ has the following properties:

$$f\in C^\infty(\bar{\Omega}),\quad f\geqq 0,\quad \max_{x\in\bar{\Omega}} f(x)>1,\quad f<1 \quad \text{on } \partial\Omega. \tag{1.2}$$

From the theory of monotone operators and regularity theory for solutions of elliptic boundary value problems it follows that for a given $\varepsilon>0$ (1.1) with f as in (1.2) has a unique solution $u\in C^\infty(\bar{\Omega})$ (cf. [3], [4]). Furthermore, a priori bounds for this solution can be given as

$$0\leqq u\leqq\frac{K}{\varepsilon}\cdot\max_{\bar{\Omega}} f(x). \tag{1.3}$$

For the lower bound we refer to [5]. The upper bound is valid with $K=|(-\Delta)^{-1}|$. Here $(-\Delta)^{-1}$ denotes the inverse of the Laplacian with Dirichlet boundary condition considered as a bounded operator from $C(\bar{\Omega})$ into $C(\bar{\Omega})$, where $C(\bar{\Omega})$ is endowed with

* Received by the editors May 8, 1985; accepted for publication (in revised form) January 22, 1986.

† Ecole Centrale de Lyon, 69130 Ecully, France.

‡ Department of Mathematics, State University of Utrecht, 3508 TA Utrecht, The Netherlands.

§ Universite de Valenciennes, 59300 Valenciennes, France.

the maximum norm $|\ |_{\max}$ and $\|(-\Delta)^{-1}\|$ denotes its operator norm. The upper bound follows, since

$$u \leqq |u|_{\max} = \varepsilon^{-1}\left|(-\Delta)^{-1}\left(f-\frac{u}{1+u}\right)\right|_{\max} \leqq \varepsilon^{-1}\|(-\Delta)^{-1}\|\left|f-\frac{u}{1+u}\right|_{\max}$$

and, using (1.2), $|f-u/(1+u)| \leqq \max_{\bar{\Omega}} f$.

We aim to study and construct asymptotic approximations for u, valid for small positive values of ε.

From Brauner and Nicolaenko [6], [7] we have the following results. Consider the rescaled function $w=\varepsilon u$. w converges for $\varepsilon \downarrow 0$ to a function $w_0 \in W^{2,p}(\Omega) \cap C^1(\bar{\Omega})$, $p \geqq 2$, where the convergence takes place strongly in $H_0^1(\Omega)$ and weakly in $W^{2,p}(\Omega)$. Note that $w_0 \geqq 0$. We introduce

$$\begin{aligned} &\Omega_+ \underset{\text{def}}{=} \{x \in \Omega \mid w_0(x) > 0\}, \qquad S = \partial\Omega_+, \\ &\Omega_0 = \operatorname{int}\{x \in \Omega \mid w_0(x) = 0\}. \end{aligned} \tag{1.4}$$

In this paper we consider the situation where the open set Ω_+ has no holes, its boundary S is smooth and connected and Ω_+ lies everywhere on one side of S. S is called the free boundary and we suppose that S has a positive distance to $\partial\Omega$, i.e., $\bar{\Omega}_+ \subset \Omega$. We mention that regularity results on the free surface S in 2 dimensions have been established once S is a Jordan curve (cf. [8] or [20], [21]). (See Fig. 1.)

In this situation w_0 restricted to $\bar{\Omega}_+$ is in $C^\infty(\bar{\Omega}_+)$, and w_0 solves the obstacle problem

$$\begin{aligned} &-\Delta w_0 = f-1 \quad \text{in } \Omega_+, \\ &w_0 = 0 \quad \text{on } S, \\ &\frac{\partial w_0}{\partial n} = 0 \quad \text{on } S. \end{aligned} \tag{1.5}$$

Furthermore, in Ω_0 we have $f<1$ (cf. [8]) and $u(x)$ converges pointwise as $\varepsilon \downarrow 0$ to the function $U_0(x)$, which is the unique solution of the formal limit of the equation (1.1), i.e.,

$$\frac{U_0}{1+U_0} = f. \tag{1.6}$$

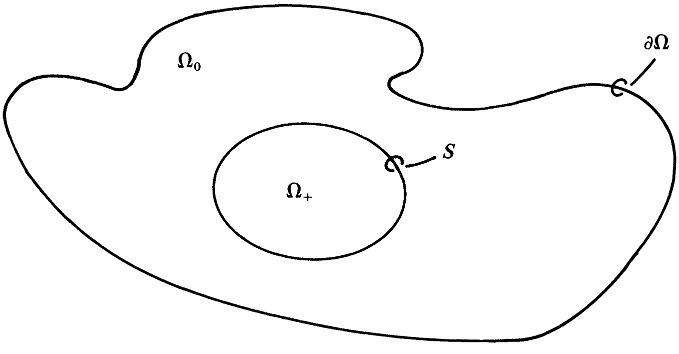


FIG. 1

Finally we can show [5], [10] that

$$\max_{x\in S} f(x) < 1. \tag{1.7}$$

Hence $U_0(x)$ is well defined in $\bar{\Omega}_0$ and $U_0 \in C^\infty(\bar{\Omega}_0)$.

From the point of view of asymptotic approximations we can summarize these results by stating that in Ω_+ the solution u is approximated by $(1/\varepsilon)w_0$, while in Ω_0 the approximation is given by U_0. A natural question arises now: can one construct asymptotic expansions in these domains and global expansion of u valid in the whole domain $\bar{\Omega}$?

Let us show that the question is highly nontrivial by attempting in an elementary way to construct an expansion in Ω_+. We write

$$w = \varepsilon u \tag{1.8}$$

and obtain for w the problem

$$-\Delta w + \frac{\varepsilon}{\varepsilon + w} = f - 1 \quad \text{in } \Omega_+. \tag{1.9}$$

This suggests an expansion

$$w = w_0 + \varepsilon w_1 + \varepsilon^2 w_2 + \cdots. \tag{1.10}$$

With w_0 as defined by (1.5) one gets for w_1 the problem

$$-\Delta w_1 = -\frac{1}{w_0} \quad \text{in } \Omega_+. \tag{1.11}$$

Clearly the right-hand side is singular on the boundary S of Ω_+. For higher terms of the expansion (1.10) one gets equations with stronger singularities.

One must realize at this stage that in a singular perturbation problem (1.1) one should expect along the "free boundary" S the occurrence of a transition layer in which the solution varies rapidly. Regular expansions of the type (1.10) can be expected to be valid only in compact subsets of Ω_+. Relations between regular expansions and expansions in the layer are established by matching. Also, one should realize that the structure of the expansions is not known a priori and that terms other than those depicted in (1.10) can (and will) occur.

In this paper we construct asymptotic expansions for the solution u of (1.1) up to an arbitrary order of accuracy and prove their validity. Our motivation in performing the analysis was not only to provide a full answer to the problem of asymptotic expansions for the specific equation (1.1). We also found that the problem at hand was a true challenge even to experienced practitioners of asymptotic analysis. We therefore hope that the techniques and reasoning that we have developed may be useful in other problems of a similar nature.

Let us now sketch the contents of the following sections: Section 2: The regular expansion in Ω_0 and the layer at $\partial\Omega$; Section 3: The principal terms in the internal layer at S and in the expansion in Ω_+; Section 4: Higher order terms near S and in Ω_+; Section 5: Matching relations of the internal layer at S and the expansions in Ω_0, Ω_+; Section 6: Composition of a global approximation Z_N and estimation of the error $u - Z_N$; Section 7: Discussion of some generalisations.

From the point of view of asymptotic expansion, § 2 will be rather straightforward. However, in the following sections the construction process contains some surprising elements. For example, the structure of the asymptotic expansions near the free surface S and in Ω_+ turns out to be unusual; not only do fractional orders of ε appear as order functions, but so do fractional orders of ε multiplied with powers of $\ln \varepsilon$. Partial results in this sense were also announced by Frank and Wendt, cf. [9]. One of the nontrivial aspects of the construction will be to determine which order functions are needed in the local expansions near S and in Ω_+. For precise information on these order functions, see §§ 3 and 4. Another interesting phenomenon in the construction is the unavoidable occurrence of singularities in the higher order terms of the expansion in Ω_+ when the variables approach the free surface. This requires a rather delicate analysis separating the singularities from the regular parts of the terms. We then show that these singularities match the layer at S and that the layer changes the singularities coming from Ω_+ into regular terms (see §§ 4 and 5). To demonstrate the validity of the constructed global approximation we use an estimation result based on barrier-function techniques (see § 6). Using the results in that section we can improve the convergence statements made earlier in this introduction. For example, if K is an ε-independent compact subset of Ω_0, then

$$\max_{x\in K} |u-U_0| = O(\varepsilon) \quad \text{for } \varepsilon \downarrow 0. \tag{1.12}$$

Instead of an estimate $\|w-w_0\|_{H_0^1(\Omega)} = O(\sqrt{\varepsilon})$ and weak convergence in $W^{2,p}(\Omega)$ we obtain here

$$\|w-w_0\|_{C^1(\bar{\Omega})} = O(\sqrt{\varepsilon}) \quad \text{for } \varepsilon \downarrow 0, \tag{1.13}$$

$$\|w-w_0\|_{W^{2,p}(\Omega)} = O(\varepsilon^q) \quad \text{for } \varepsilon \downarrow 0 \text{ with } q = 1/(2p). \tag{1.14}$$

Finally, in the discussion of generalisations in § 7 we consider the effects of domains of dimension > 2, of more general second order elliptic operators and of more general nonlinearities. To conclude this introduction we remark that the work of M. Garbey's thesis lies at the basis of this report (cf. [10]) and that some of the results can be found summarised in [11].

2. The regular expansion in Ω_0 and the layer at $\partial\Omega$. From the point of view of techniques of singular perturbations, this section is very straightforward. We simply need the results in the sequel.

In Ω_0 we construct a local formal approximation of the solution u as

$$u \simeq U^M \underset{\text{def}}{=} \sum_{k=0}^{M} \varepsilon^k U_k(x). \tag{2.1}$$

Substitution in (1.1) provides us with a recursive system of equations:

$$\frac{U_0}{(1+U_0)} = f, \quad \text{i.e. } U_0 = \frac{f}{(1-f)} \tag{2.2}$$

and

$$U_{n+1} = (1+U_0)^2 \cdot \{\Delta U_n + F_{n+1}(U_0, \cdots, U_n)\} \tag{2.3}$$

with

$$F_{n+1} = \left[\frac{\partial^{n+1}}{\partial \varepsilon^{n+1}}\left(1+U_0+\sum_{k=1}^{n} \varepsilon^k U_k\right)^{-1}\right]_{\varepsilon=0}$$

$$= \sum_{m=2}^{n} \frac{(-1)^m}{(1+U_0)^{m+1}} \sum_{\substack{\vec{k}\in\mathbb{N} \\ |\vec{k}|=n}} \prod_{i=1}^{m} U_{k_i}.$$

The system for the U_n's can be solved uniquely and each U_n is in $C^\infty(\bar{\Omega}_0)$ because of (1.2) and (1.7).

In general U_0 and also the other U_n's will not satisfy the Dirichlet boundary condition on $\partial\Omega$. As a consequence we need a layer along $\partial\Omega$ to correct this.

As usual (cf. [1], [12]) we put

$$u \simeq Z_0^M \underset{\text{def}}{=} U^M + G^M H\left(\frac{d(x)-d_0}{d_0}\right) \quad \text{with } G^M = \sum_{k=0}^{2m+1} (\sqrt{\varepsilon})^k G_k(\zeta, \omega). \tag{2.4}$$

Here ζ is the layer variable

$$\zeta = \frac{d(x)}{\sqrt{\varepsilon}} \tag{2.5}$$

where $d(x)$ denotes the distance of a point x to $\partial\Omega$.

For $x \in \Omega$ we denote by $\tilde{x}$ a point on $\partial\Omega$ such that $|x-\tilde{x}| = d(x)$. Note that $\tilde{x}$ is uniquely determined by x for $x \in \hat{D} \rightleftharpoons \{x \in \bar{\Omega} | d(x) \leqq \hat{d}\}$ with a certain $\hat{d} > 0$ depending on the curvature of $\partial\Omega$. For $x \in \hat{D}$ we define $\omega(x)$ as the distance from $\tilde{x}$ along $\partial\Omega$ to a certain reference point 0 on $\partial\Omega$ measured say in clockwise orientation, $0 \leq \omega < L$ with $L = \text{length } (\partial\Omega)$. (See Fig. 2.)

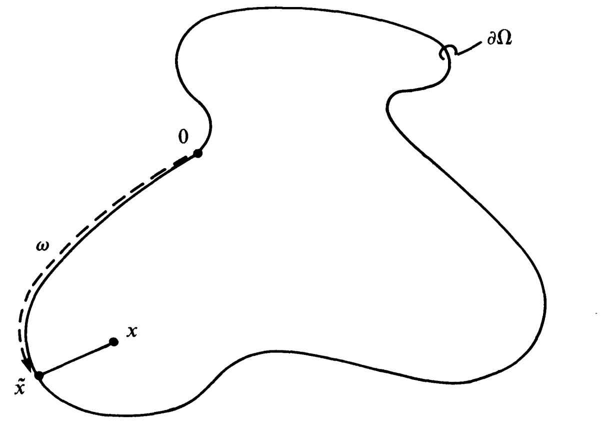


FIG. 2. $x = \tilde{x}(\omega) + d\nu(\omega)$ *with* ν *the inward normal on* $\partial\Omega$, $\langle \nu, \tilde{x}' \rangle = 0$, $\langle \tilde{x}', \tilde{x}' \rangle = \langle \nu, \nu \rangle = 1$.

The function H is introduced in order to avoid singularities in the transformation $x \to (d, \omega)$. H is a C^∞ cut-off function

$$H(s) = \begin{cases} 1 & \text{for } s \leqq \frac{1}{2} \\ 0 & \text{for } s \geqq 1 \end{cases} \quad \text{and } H'(s) \leqq 0. \tag{2.6}$$

The constant d_0 is chosen sufficiently small (say $\frac{1}{3}\hat{d}$), so that $x \to (d, e^{i2\pi\omega/L})$ is a C^∞-diffeomorphism from $\{x \in \bar{\Omega} | d(x) \leqq d_0\}$ onto $[0, d_0] \times T$ with $T = \{z \in \mathbb{C} | |z| = 1\}$.

Substitution in (1.1) leads us to:

$$\frac{\partial^2 G_0}{\partial \zeta^2} = \frac{1}{1+\gamma} - \frac{1}{1+\gamma+G_0} \tag{2.7}$$

with $\gamma(\omega) = U_0|_{\partial\Omega}(\omega) \geqq 0$. Further, we want G_0 to satisfy the following boundary conditions:

$$G_0|_{\zeta=0} = -\gamma, \qquad G_0 \to 0 \quad \text{for } \zeta \to \infty. \tag{2.8}$$

In this way G_0 is uniquely determined. As a function of ζ for a fixed ω it is given by

$$\int_{-\gamma/(1+\gamma)}^{G_0/(1+\gamma)} \frac{dg}{\sqrt{2[g-\ln(1+g)]}} = \frac{\zeta}{1+\gamma} \quad \text{if } \gamma(\omega)>0 \tag{2.9}$$

and

$$G_0 \equiv 0 \quad \text{if } \gamma(\omega)=0. \tag{2.10}$$

If $\gamma(\omega)<0$ then $G_0(\zeta,\omega)$ is strictly increasing on $[0,\infty)$. The behaviour for $\zeta\to\infty$ can be described more explicitly if we note that

$$G_0 = C\exp\left(-\frac{\zeta}{1+\gamma}\right)\cdot\exp\left(-\int_{G_0/(1+\gamma)}^{0} W(g)\,dg\right) \tag{2.11}$$

with $C=\gamma\exp(-\int_{-\gamma/(1+\gamma)}^{0} W(g)\,dg)$.

Using the implicit function theorem we see that G_0 is exponentially decreasing for $\zeta\to\infty$. Moreover, G_0 depends smoothly on ω, ζ and all derivatives $\partial^{k+l}G_0/\partial\zeta^k\partial\omega^l$ vanish exponentially for $\zeta\to\infty$.

For the higher order terms G_n we find problems of the following type:

$$\begin{gathered}\frac{\partial^2 G_{n+1}}{\partial\zeta^2}-\frac{G_{n+1}}{(1+\gamma+G_0)^2}=F_{n+1}(G_0,\cdots,G_n),\\ G_{n+1}|_{\zeta=0}=0 \text{ if } n+1 \text{ odd}, =-U_p|_{\partial\Omega} \text{ if } n+1 \text{ even, with } p=\tfrac{1}{2}(n+1)\end{gathered} \tag{2.12}$$

and

$$F_{n+1}=\frac{1}{(n+1)!}\left[\frac{\partial}{\partial\delta^{n+1}}\bar{F}^{(n)}\right]_{\delta=0},$$

$$\begin{aligned}\bar{F}^{(n)}=1-f(\delta\zeta,\omega)-\delta^2\sum_{k=0}^{M}\delta^{2k}(\Delta U_k)(\delta\zeta,\omega)\\ -\left\{1+\sum_{k=0}^{M}\delta^{2k}U_k(\delta\zeta,\omega)+\sum_{k=0}^{n}\delta^k G_k(\zeta,\omega)\right\}^{-1}\\ -\left\{\delta^2h^{-1}\frac{\partial^2}{\partial\omega^2}+\delta h^{-1}d_1\frac{\partial}{\partial\zeta}+\delta^2h^{-2}d_2\frac{\partial}{\partial\omega}\right\}\sum_{k=n}^{n}\delta^k G_k(\zeta,\omega),\end{aligned}$$

where δ is a shorthand notation for $\sqrt{\varepsilon}$.

Note that by construction of the U_k's

$$1-f-\varepsilon\Delta U^M-(1+U^M)^{-1}=O(\delta^{2M+2})=O(\varepsilon^{M+1}) \tag{2.13}$$

uniformly on $\bar{\Omega}_0$. Using this and induction with respect to n it is not difficult to show that (2.12) has a unique solution, which vanishes exponentially for $\zeta\to\infty$:

$$\begin{aligned}G_{n+1}(\zeta,\omega)=K(\zeta,\omega)\cdot\int_0^\infty K(\zeta_1,\omega)^{-2}\int_{\zeta_1}^\infty K(\zeta_2,\omega)F_{k+1}(\zeta_2)\,d\zeta_2\,d\zeta_1\\ +G_{n+1}(0,\omega)\cdot K(\zeta,\omega).\end{aligned} \tag{2.14}$$

In this expression $K(\zeta,\omega)$ is a suitably chosen solution of the homogeneous equation with $K(0,\omega)=1$:

$$K(\zeta,\omega)=\bar{C}(\omega)\frac{\partial G_0}{\partial\zeta}(\zeta,\omega),$$

$$\bar{C}(\omega)=[2(-\check{\gamma}-\ln(1-\check{\gamma}))]^{1/2} \quad \text{with } \check{\gamma}=\gamma/(1+\gamma).$$

Using (2.11) one can check that K is smoothly dependent on ω, even when $\gamma(\omega)$ has a zero. Moreover, it is not difficult to see that not only G_{n+1} but also derivatives like $\partial^{k+l}G_{n+l}/\partial\zeta^k\partial\omega^l$ vanish exponentially for $\zeta\to\infty$.

To conclude this section we remark that the approximation constructed up to now, i.e., Z_0^M as given in the right-hand side of (2.4), is such that

$$-\varepsilon\Delta Z_0^M+\frac{Z_0^M}{1+Z_0^M}=f-r \tag{2.15}$$

where the error term r is $O(\varepsilon^{M+1})$ uniformly on $\bar\Omega_0$. That is, for $d(x)\geqq\frac{1}{2}d_0$ r is given by the right-hand side of (2.13), apart from some additional exponentially small terms when $\frac{1}{2}d_0\leqq d(x)\leqq d_0$. In the region $0\leqq d(x)\leqq\frac{1}{2}\delta_0$ the error is equal to

$$\bar F^{(2M+2)}-\sum_{n=0}^{2M+1}\frac{\delta^n}{n!}\frac{\partial\bar F^{(2M+2)}}{\partial\delta^n}\bigg|_{\delta=0}.$$

Using (2.12) and the exponential decay of the G_n's we find that this expression is indeed $O(\delta^{2M+2})$.

It is now clear that this formal approximation in the subdomain Ω_0 is of the order $O(\varepsilon^{M+1})$, where M can be taken arbitrarily large. In §§ 3, 4 and 5 we shall work toward such a result in $\Omega\backslash\Omega_0$.

3. The principal terms in the internal layer at S and in the expansion in Ω_+. In this section we commence the study of the transition layer along S and the expansion in Ω_+. The main objective is to determine the set of order functions occurring in various expansions. The difficulties due to singular behaviour of the expansion in Ω_+ do not occur yet. The next term of the expansion of u in Ω_+ turns out to be $\ln\varepsilon\cdot w_1$, where w_1 is the solution of a well-defined Dirichlet problem for the Laplace's equation in Ω_+.

In order to describe the local approximation of the solution in the layer at the free surface S (see Fig. 3) we use the following coordinates. The ρ coordinate is

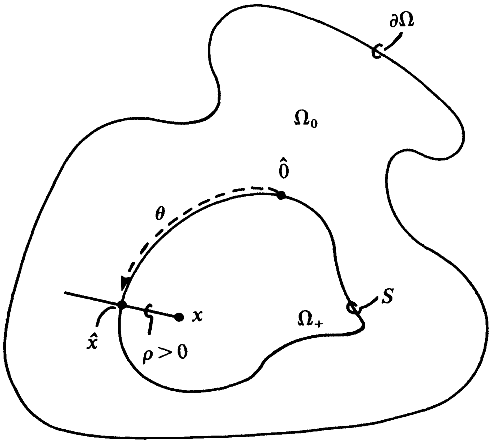


FIG. 3. $x=\hat x(\theta)+\rho n(\theta)$ *with* n *the normal on* S *in the direction of* Ω_+; $\langle n,\hat x'\rangle=0$, $\langle\hat x',\hat x'\rangle=\langle n,n\rangle=1$. *Note that* $\rho>0$ *in* Ω_+, $\rho<0$ *in* Ω_0; *for points sufficiently close to* S $|\rho|=$ *distance to* S, $0\leqq\theta<\theta_0$ *where* $\theta_0=$ *length* (S).

stretched in a significant way:

$$\xi=\rho/\sqrt{\varepsilon}. \tag{3.1}$$

In these coordinates the operator $\varepsilon\Delta$ is given by

$$\varepsilon\Delta=\frac{\partial^2}{\partial\xi^2}+\varepsilon g^{-1}\frac{\partial^2}{\partial\theta^2}+\sqrt{\varepsilon}\cdot g^{-1}e_1\frac{\partial}{\partial\xi}-\varepsilon\cdot g^{-2}\cdot e_2\frac{\partial}{\partial\theta} \tag{3.2}$$

with $g=1+2\sqrt{\varepsilon}\,\xi a+\varepsilon\xi^2 b$, $e_1=\sqrt{\varepsilon}\,\xi b+a$, $e_2=\sqrt{\varepsilon}\,\xi a'+\frac{1}{2}\varepsilon\xi^2 b'$ and $a=\{n',\hat{x}'\rangle$, $b=\langle n',n'\rangle$, where $'$ denotes differentiation with respect to θ. For the local approximation of the solution near S we put

$$u\simeq\psi_0(\xi,\theta)+\delta_1\psi_1(\xi,\theta)+\delta_2\psi_2(\xi,\theta)+\cdots. \tag{3.3}$$

In this expansion the magnitudes of the higher order terms $\delta_1(\varepsilon)$, $\delta_2(\varepsilon)$, etc., are unknown at the start of the construction. They will be determined during the construction process.

If we recall (1.5), it is clear that the local approximation in Ω_+ will be of the following type:

$$u\simeq\varepsilon^{-1}w_0+\bar{\delta}_1\bar{w}_1+\cdots. \tag{3.4}$$

Again the magnitude $\bar{\delta}_1(\varepsilon)$ has to be found during the construction. Let us now first deduce the principal term in the layer.

The equation for ψ_0 is obtained from the $O(1)$ terms in (1.1) by substituting (3.2) for $\varepsilon\Delta$. It is the following O.D.E.:

$$\frac{\partial^2\psi_0}{\partial\xi^2}=1-\hat{f}-\frac{1}{1+\psi_0}. \tag{3.5}$$

The variable θ acts only as a parameter. Here $\hat{f}$ denotes $f|_S$. This function depends only on θ, it is smooth and periodic and $0<\hat{f}<1$ (see (1.2) and (1.7)). In addition to (3.5), matching provides us with "boundary conditions" for $\xi\to-\infty$ and $\xi\to+\infty$. Expanding the regular expansion in Ω_0 in the ξ, θ variables we obtain:

$$\lim_{\xi\to-\infty}\psi_0=\hat{\gamma} \tag{3.6}$$

with $\hat{\gamma}=_{\text{def}}U_0|_S=\hat{f}/(1-\hat{f})$. From (1.5) it follows that $w_0=\frac{1}{2}(1-\hat{f})\rho^2+O(\rho^3)$ for $\rho\downarrow 0$. As a consequence, expansion of (3.4) in ξ, θ coordinate leads us to

$$\psi_0=\tfrac{1}{2}(1-\hat{f})\xi^2+o(\xi^2)\quad\text{for }\xi\to\infty. \tag{3.7}$$

The solutions of (3.5)-(3.7) can easily be found. Namely, (3.6), (3.5) imply that $\partial^2\psi_0/\partial\xi^2$ and also $\partial\psi_0/\partial\xi$ vanish for $\xi\to-\infty$. Multiplication of (3.5) with $\partial\psi_0/\partial\xi$ and integration yields

$$\begin{gathered}\psi_0=\hat{\gamma}+(1+\hat{\gamma})v\left(\frac{\xi}{1+\hat{\gamma}}\right),\\ \int_{v_0}^{v}\frac{dz}{\sqrt{2[z-\ln(1+z)]}}=\eta\quad\text{with }v_0>0.\end{gathered} \tag{3.8}$$

Here the lower endpoint $v_0(\theta)$ plays the role of a free constant, possibly depending on θ. Eventually $v_0(\theta)$ will be determined by a matching argument. The function $v(\eta,\theta)$ is strictly increasing on $(-\infty,\infty)$ from 0 to $+\infty$.

Let us have a closer look at the asymptotic behaviour of ψ_0 for both $\xi \to -\infty$ and $\xi \to +\infty$. Using a procedure as in (2.11) we find that

$$v = c\, e^{\eta}(1+O(e^{\eta})) \quad \text{for } \eta \to -\infty \tag{3.9}$$

with $c = v_0 \exp\left(\int_0^{v_0} W(g)\, dg\right) > 0$. Since

$$\frac{\partial \psi_0}{\partial \xi}(\xi, \theta) = \sqrt{2[v - \ln(1+v)]}\left(\frac{\xi}{1+\hat{\gamma}}\right), \tag{3.10}$$

it is also clear that

$$\frac{\partial \psi_0}{\partial \xi} = c\, e^{\xi/1+\hat{\gamma}}(1+O(e^{\xi/1+\hat{\gamma}})) \quad \text{for } \xi \to -\infty. \tag{3.11}$$

For the analysis of the behaviour for $\xi \to -\infty$ we introduce

$$I(z) = (2z)^{-1/2} \cdot \{(1 - z^{-1} \ln(1+z))^{-1/2} - 1\}. \tag{3.12}$$

Note that $I(z) = O(z^{-3/2} \ln z)$ for $z \to \infty$; hence, $I(z)$ is integrable at ∞.

Since

$$\frac{1}{\sqrt{2[z - \ln(1+z)]}} = \frac{1}{\sqrt{2z}} + I(z), \tag{3.13}$$

the implicit formula for v in (3.8) can be written as

$$v = \frac{1}{2}\left\{\eta + \left(\sqrt{2v_0} - \int_{v_0}^{\infty} I(z)\, dz\right) + \int_{v}^{\infty} I(z)\, dz\right\}^2, \tag{3.14}$$

i.e.,

$$v(\eta, \theta) = \frac{1}{2}\eta^2 \cdot \left\{1 + 2\eta^{-1}\left(\sqrt{2v_0} - \int_{v_0}^{\infty} I(z)\, dz\right) + o(\eta^{-1})\right\} \quad \text{for } \eta \to \infty. \tag{3.15}$$

For ψ_0 this means

$$\psi_0(\xi, \theta) = \frac{1}{2}(1 - \hat{f})\xi^2 + \left\{\sqrt{2v_0} - \int_{v_0}^{\infty} I(z)\, dz\right\}\xi + o(\xi) \quad \text{for } \xi \to \infty. \tag{3.16}$$

Next we observe that the linear term in ξ in ψ_0 gives rise to an $O(\varepsilon^{-1/2})$ term when reexpanded in the ρ, θ variables in Ω_+. Now let us take in (3.4)

$$\bar{\delta}_1 = \varepsilon^{-1/2}. \tag{3.17}$$

Then $\bar{w}_1$ has to satisfy the equation

$$\Delta \bar{w}_1 = 0. \tag{3.18}$$

Matching with the layer at the free surface, which has no $O(\varepsilon^{-1/2})$ term, makes it necessary that

$$\bar{w}_1 = 0 \quad \text{on } S. \tag{3.19}$$

The conclusion is

$$\bar{w}_1 \equiv 0 \quad \text{in } \bar{\Omega}_+. \tag{3.20}$$

Another consequence of the matching is that the linear term in ξ in (3.16) equals $\partial \bar{w}_1 / \partial n|_S$ and hence vanishes, i.e.,

$$\sqrt{2v_0} = \int_{v_0}^{\infty} I(z)\, dz. \tag{3.21}$$

Thus the value of v_0 follows indeed by a matching argument. Somewhat surprisingly $v_0 > 0$ does not depend on θ.

The leading term ψ_0 in the layer is now completely known. It will be important to have a precise description of the asymptotics of ψ_0 for $\xi \to \infty$.

LEMMA 1. *Let us denote*

$$\eta_1 = \frac{\ln \xi}{\xi^2}, \qquad \eta_2 = \frac{1}{\xi^2}. \tag{3.22}$$

Then there exists a $\xi_0 > 0$ such that for ξ sufficiently large, $\xi \geqq \xi_0$

$$\psi_0 = \frac{1}{2}(1-\hat{f})\xi^2 + \frac{2}{1-\hat{f}} \ln \xi + T(\theta, \eta_1, \eta_2) \tag{3.23}$$

where for η_1, η_2 sufficiently small: $|\eta_1| < \nu$, $|\eta_2| < \nu$ and $\theta \in [0, \theta_0)$.

(i) *T is smooth and T is periodic in θ,*

(ii) *T has a convergent power series in η_1 and η_2,*

$$T = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} T_{kl} \eta_1^k \eta_2^l.$$

Proof of Lemma 1. The function v satisfies the equation

$$\sqrt{2v} = \eta \cdot \left\{ 1 + \eta^{-1} \int_v^{\infty} I(z)\, dz \right\}. \tag{3.24}$$

Now put

$$v = \tfrac{1}{2}\eta^2(1+\hat{v})^2. \tag{3.25}$$

After a change of the integration variable $z = \eta^2 s$ we obtain

$$\hat{v} = \int_{\frac{1}{2}(1+\hat{v})^2}^{\infty} (2s)^{-1/2}\{[1 - s^{-1}(2\hat{\eta}_1 + \hat{\eta}_2 \ln s + \hat{\eta}_2 \ln(1+\hat{\eta}_2 s^{-1}))]^{-1/2} - 1\}\, ds \tag{3.26}$$

with

$$\hat{\eta}_1 = \frac{\ln \eta}{\eta^2}, \qquad \hat{\eta}_2 = \frac{1}{\eta^2}.$$

For a moment we consider $\hat{\eta}_1$ and $\hat{\eta}_2$ as independent variables. An application of the implicit function theorem (cf. [13]) shows that for $\hat{\eta}_1$, $\hat{\eta}_2$ sufficiently small $\hat{v}$ is analytic in $\hat{\eta}_1$ and $\hat{\eta}_2$ with

$$\hat{v}(0,0) = 0, \qquad \frac{\partial \hat{v}}{\partial \hat{\eta}_1}(0,0) = 2. \tag{3.27}$$

Using (3.8), (3.25) and expressing $\hat{\eta}_1$, $\hat{\eta}_2$ in terms of η_1, η_2 and $\hat{\gamma}$, we find a transcription of this result for ψ_0:

$$\psi_0 = \hat{\gamma} + \tfrac{1}{2}(1+\hat{\gamma})^{-1}\xi^2[1 + \hat{v}((1+\hat{\gamma})^2(\eta_1 - \eta_2 \ln(1+\hat{\gamma})), (1+\hat{\gamma})^2\eta_2)]^2. \tag{3.28}$$

Herewith the growing terms for $\xi \to \infty$ in (3.23) are easily checked. Further, we deduce from (3.28) that

$$\psi_0 = \xi^2 S(\theta, \eta_1, \eta_2) \tag{3.29}$$

where S has the properties (i) and (ii) as indicated in the lemma. However, (3.29) still differs from (3.23), but a little trick is helpful.

Substitution of (3.28) in the O.D.E. (3.5) yields

$$\frac{\partial^2\psi_0}{\partial\xi^2}=1-\hat{f}-\frac{2(1+\hat{\gamma})}{\xi^2}+\frac{1}{\xi^2}\hat{S}(\theta,\eta_1,\eta_2) \tag{3.30}$$

where $\hat{S}$ has the properties (i) and (ii) and $\hat{S}(\theta,0,0)=0$. Integrating twice in (3.30), meanwhile using the fact that ψ_0 does not contain a linear term for $\xi\to\infty$, we find (3.23). □

As a consequence of the behaviour of ψ_0 for $\xi\to\infty$ we need in the expansion in Ω_+ a term of magnitude ln (ε), i.e.,

$$u=\varepsilon^{-1}w_0+\ln\varepsilon\cdot w_1+\cdots \tag{3.31}$$

where w_1 is the solution of

$$\Delta w_1=0, \tag{3.32}$$

$$w_1=2(1-\hat{f})^{-1}\quad\text{on } S. \tag{3.33}$$

But this reflects back on the layer at S. It is easy to see that matching requires that the next order term in the layer expansion correspond to

$$\delta_1=\sqrt{\varepsilon}\cdot\ln\varepsilon, \tag{3.34}$$

where ψ_1 is a solution of the homogeneous, linearized layer equation

$$\frac{\partial^2\psi_1}{\partial\xi^2}-\frac{1}{(1+\psi_0)^2}\psi_1=0, \tag{3.35}$$

which vanishes for $\xi\to-\infty$, i.e., ψ_1 is proportional to the derivative of ψ_0

$$\psi_1=K(\theta)\frac{\partial\psi_0}{\partial\xi}. \tag{3.36}$$

Now an obvious question is how the other order functions in the asymptotic sequence in the layer are generated. "Minimal" requirements for this sequence of order functions $\mathcal{S}=\{\delta_n\mid n\in\mathbb{N}\cup\{0\}\}$ are

1° $1,\sqrt{\varepsilon}\ln\varepsilon$ and $(\sqrt{\varepsilon})^k$, $k\in\mathbb{N}$ are in $\mathcal{S}$. The first magnitudes are those corresponding to ψ_0,ψ_1; the others are generated by the Taylor series of f in the ξ, θ coordinates.

2° Stability under multiplication: $\delta_n,\delta_m\in\mathcal{S}\Rightarrow\delta_n\cdot\delta_m\in\mathcal{S}$, especially $\delta_n\in\mathcal{S}\Rightarrow\delta_n\sqrt{\varepsilon}\in\mathcal{S}$. The reason is the structure of $\varepsilon\Delta$ in (3.2) and the fact that such products appear automatically in the Taylor series of the nonlinearity $u/(1+u)$.

These conditions 1° and 2° naturally lead us to (at least)

$$\mathcal{S}=\{\delta_1^k\delta_2^l\mid k\in\mathbb{N}\cup\{0\}\text{ and }l\in\mathbb{N}\cup\{0\}\} \tag{3.37}$$

with

$$\delta_1=\sqrt{\varepsilon}\ln\varepsilon,\qquad \delta_2=\sqrt{\varepsilon}.$$

In the sequel we shall demonstrate that this sequence $\mathcal{S}$ is indeed sufficient.

On the other hand, if the layer term $\delta_1^k\delta_2^l\psi_{k,l}(\xi,\theta)$ contains a constant term $\sim\delta_1^k\delta_2^l$ for $\xi\to\infty$, we expect that all these order functions $\delta_1^k\delta_2^l$ will also be present in the expansion in the region Ω_+. In combination with (3.31) this leads us to the following sequence $\mathcal{S}_+$ of order functions needed in Ω_+:

$$\mathcal{S}_+=\{\varepsilon^{-1},\ln\varepsilon\}\cup\mathcal{S}. \tag{3.38}$$

This sequence contains all the order functions generated by $\psi_0+\delta_1\psi_1$ in (ρ, θ) coordinates. It is also closed under Taylor series expansion of the nonlinearity. Below we shall show that the sequence $\mathcal{S}_+$ is sufficient to construct the expansion in Ω_+. One could have the impression that $\mathscr{S}_+$ contains superfluous order functions. It cannot be denied that $\mathscr{S}_+$ contains a few redundant order functions, for example, δ_1. However, the redundancy in $\mathscr{S}_+$ is not large, as we shall see.

4. Higher order terms near S and in Ω_+. In this section we confront the problem of the singular behaviour of the expansion in Ω_+. Our analysis is an interplay of formal construction and matching arguments. We aim to show that singularities of the expansion in Ω_+ near S are counterbalanced by corresponding terms of the layer expansion. The computational complexity comes mainly from the fact that we consider expansions to arbitrary order of accuracy. We have to do so in order to show that the construction does not break up at some stage. Let us denote the expansion in the free surface layer by

$$u \simeq \psi^N \underset{\text{def}}{=} \sum_{k=0}^{k+l\leqq N} \sum_{l=0} \delta_1^k \delta_2^l \psi_{k,l}(\xi, \theta), \tag{4.1}$$

with $\delta_1 = \sqrt{\varepsilon} \cdot \ln \varepsilon$, $\delta_2 = \sqrt{\varepsilon}$ and $\xi = \rho/\sqrt{\varepsilon}$, as before. The notation for the expansion in Ω_+ will be

$$u \cong \Phi^N \underset{\text{def}}{=} \varepsilon^{-1} w_0(x) + \ln \varepsilon \cdot w_1(x) + \sum_{k=0}^{k+l\leqq N} \sum_{l=0} \delta_1^k \delta_2^l \phi_{k,l}(x). \tag{4.2}$$

In this section we present an iterative scheme by which the $\psi_{k,l}$'s and $\phi_{k,l}$'s can be determined in a unique way. To start, we discuss in § 4.1 the construction of the $\psi_{k,l}$'s while in each $\psi_{k,l}$ an additive term $A_{kl}(\partial\psi_0/\partial\xi)$, a solution of the homogeneous equation, is still free. Furthermore, we analyze some of the properties of the $\psi_{k,l}$'s and we show that some of the A_{kl} have to be zero due to a simple matching argument.

In § 4.2 the $\phi_{k,l}$'s are constructed, while in each $\phi_{k,l}$ a linear degree of freedom, corresponding to a choice of a boundary value function $g_{k,l}$ on S, is built in. The $\phi_{k,l}$'s will contain a singular part for $\rho \downarrow 0$. The freedom of $g_{k,l}$ will arise in the boundary values on S of the regular part of $\phi_{k,l}$.

Next, in § 4.3 we describe the scheme by which all free functions $A_{k,l}$ and $g_{k,l}$ can be uniquely determined. As a matter of fact this scheme will be based on a partial matching relation. The full matching will be considered in the next section.

4.1. The $\psi_{k,l}$'s with free $A_{k,l}$'s. Let us first consider the layer along S. Collecting the $O(\sigma_1^k \delta_2^l)$ terms in (1.1) with $\varepsilon\Delta$ as in (3.2) we find the following inhomogeneous, linear O.D.E. for $\psi_{k,l}$, $k+l>0$:

$$\frac{\partial^2 \psi_{k,l}}{\partial \xi^2} - \frac{1}{(1+\psi_0)^2} \psi_{k,l} = \hat{f}_{k,l} \tag{4.1.1}$$

with

$$\hat{f}_{k,l} = (k!\,l!)^{-1} \left[\frac{\partial^{k+l}}{\partial \delta_1^k \partial \delta_2^l} f_{k,l} \right]_{\delta_1=0,\delta_2=0},$$

where

$$f_{k,l} = -f(\delta_2 \xi, \theta) - \delta_2 B \overset{v}{\sum} \delta_1^r \delta_2^s \psi_{r,s} - \left(1 + \overset{v}{\sum} \delta_1^r \delta_2^s \psi_{r,s}\right)^{-1},$$

$$\delta_2 B \underset{\text{def}}{=} \delta_2^2 g^{-1} \cdot \frac{\partial^2}{\partial \theta^2} + \delta_2 \cdot g^{-1} e_1 \frac{\partial}{\partial \xi} - \delta_2^2 g^{-2} e_2 \frac{\partial}{\partial \theta}; \qquad \text{see (3.2).}$$

$\overset{v}{\sum}$ denotes summation over all indices r, s with

$0 \leqq r \leqq k$, $\quad 0 \leqq s \leqq l$, $\quad r+s<k+l$.

In addition, matching with the expansion in Ω_0 provides the following conditions for $\xi \to -\infty$:

$$\begin{aligned} &\psi_{0,l} \text{ is bounded by a polynomial for } \xi \to -\infty, \\ &\psi_{k,l} \to 0 \text{ for } \xi \to -\infty \text{ for } k > 0. \end{aligned} \tag{4.1.2}$$

If $\hat{f}_{k,l}$ is polynomially bounded for $\xi \to -\infty$, then the solutions of (4.1.1) that do not grow exponentially for $\xi \to -\infty$ are given by:

$$\begin{aligned} \psi_{k,l}(\xi, \theta) = \frac{\partial \psi_0}{\partial \xi}(\xi, \theta) \cdot \int_{\xi_0}^{\xi} \left[\frac{\partial \psi_0}{\partial \xi}(\eta, \theta)\right]^{-2} \int_{-\infty}^{\eta} \frac{\partial \psi_0}{\partial \xi}(\zeta, \theta) \hat{f}_{k,l}(\zeta, \theta)\, d\zeta\, d\eta \\ + A_{k,l}(\theta) \cdot \frac{\partial \psi_0}{\partial \xi}(\xi, \theta). \end{aligned} \tag{4.1.3}$$

Here we shall take ξ_0 to be a fixed, sufficiently large number. The function $A_{k,l}(\theta)$ is not determined by (4.1.2) because of the exponential decay of $\partial \psi_0 / \partial \xi$ for $\xi \to -\infty$. This simply means that at the moment (4.1.3) contains an amount of freedom.

Note that, except for this freedom, (4.1.3) allows us to calculate the $\psi_{k,l}$'s recursively. This can be done in several ways, for example:

a. calculate the $\psi_{k,l}$'s with $k + l = 1$, next those with $k + l = 2$, etc., or

b. calculate all $\psi_{0,k}$ by increasing k, next all $\psi_{l,k}$ with $l = 1$, etc.

Though the $\psi_{k,l}$'s are not uniquely determined we can already derive some of their properties. When the $A_{k,l}$'s are chosen as smooth periodic functions in θ, then all $\psi_{k,l}$'s are smooth functions of ξ and θ in $\mathbb{R} \times [0, \theta_0)$, periodic in θ. As for the behaviour of $\psi_{k,l}$ for $\xi \to -\infty$ we can derive the following results:

LEMMA 2.

$$\psi_{0,l} = P_l + \textit{exponentially small terms for } \xi \to -\infty. \tag{4.1.4}$$

Here P_l is a polynomial in ξ of degree $\leqq l$ with coefficients depending smoothly and periodically on θ. Moreover P_l has the same parity as l.

Further, for $k > 0$ there is an $m \in \mathbb{N}$ such that

$$\psi_{k,l} = O\left(\xi^m \exp\left(\frac{\xi}{1+\hat{\gamma}}\right)\right) \quad \textit{for } \xi \to -\infty. \tag{4.1.5}$$

Moreover, the derivatives $\partial^{r+s}\psi_{k,l}/\partial\xi^r \partial\theta^s$ behave in an analogous way.

The derivation of (4.1.4) and (4.1.5) is an interesting, rather easy exercise in induction using recursion as in b and using (3.10)-(3.11); further details are left to the reader.

Note that the conditions in (4.1.2) are indeed fulfilled, regardless of the choice of the $A_{k,l}$'s.

Let us now consider the asymptotic behaviour of $\psi_{k,l}$ for $\xi \to \infty$. It will be convenient to introduce the following concept:

A function χ of $(\xi, \theta) \in \mathbb{R} \times [0, \theta_0)$ is said to be of <u>type p</u> with $p \in \mathbb{Z} \leftrightarrow$

(i) *χ is smooth in ξ and θ and periodic in θ,*

(ii) *for ξ sufficiently large, $\xi \geqq \xi_0$, χ can be represented in the following way:*

$$\chi = \xi^p[X(\eta_1, \theta) + \eta_2 Y(\eta_1, \eta_2, \theta)] \tag{4.1.6}$$

with

$$\eta_1 = \frac{\ln \xi}{\xi^2}, \qquad \eta_2 = \frac{1}{\xi^2}$$

and

- X is polynomial of degree $\leqq \frac{1}{2}p$ in η_1 with coefficients depending smoothly and periodically on θ. For $p<0$ this means that $X \equiv 0$.

- Y and all its derivatives with respect to θ are analytic in η_1, η_2 for $|\eta_1|<\nu$, $|\eta_2|<\nu$ with coefficients of the power series in η_1, η_2 depending smoothly and periodically on θ.

In this definition ξ_0 and ν are positive numbers which are fixed in accordance with their values in Lemma 1. It is not difficult to verify the following extension of Lemma 1:

(4.1.7) $$\psi_0 \text{ is of type } 2, \qquad \frac{\partial\psi_0}{\partial\xi} \text{ is of type } 1.$$

This demonstrates already the relevance of the type p concept. Further, the other solution of the homogeneous version of (4.1.1) is

(4.1.8) $$\Phi \underset{\text{def}}{=} \frac{\partial\psi_0}{\partial\xi} \int_\xi^\infty \left[\frac{\partial\psi_0}{\partial\xi}(\eta,\theta)\right]^{-2} d\eta.$$

Now a little calculation shows that

(4.1.9) $$\xi^2 \left[\frac{\partial\psi_0}{\partial\xi}\right]^{-2} \text{ is of type } 0$$

and

(4.1.10) $$\Phi \text{ is of type } 0.$$

The type p concept is nicely compatible with algebraic and analytic operations and notions. We mention a few useful rules, which can be checked in an elementary way.

(4.1.11) χ of type $p \Rightarrow \xi^r\chi$ is of type $p+r$ for $r \in \mathbb{N}\cup\{0\}$, $\partial\chi/\partial\theta$ is of type p, $\partial\chi/\partial\xi$ is type $p-1$, $\partial^{r+s}\chi/\partial\xi^r\partial\theta^s$ is of type $p-r$.

(4.1.12) χ_1 of type p_1 and χ_2 of type $p_2 \Rightarrow \chi_1\chi_2$ is of type p_1+p_2, χ_1 of type p_1 and χ_2 of type p_2 with $p_2 \leqq p_1$, p_1-p_2 even $\Rightarrow \chi_1+\chi_2$ is of type p_1.

(4.1.13) A polynomial in ξ of degree p of the same parity as p with smooth, periodic θ-dependent coefficients is of type p.

For a function χ of type p there exists a uniquely defined primitive $I(\chi)$ with respect to ξ-integration such that

(4.1.14) $I(\chi)$ is a function of type $p+1$ without a constant term in its expansion for $\xi \to \infty$.

Note that

(4.1.15) $$I(\chi) = \int_{\xi_0}^{\xi} \chi(\eta,\theta)\, d\eta + I_0(\chi)(\theta)$$

with a smooth, periodic function I_0, which in general is $\neq 0$.

Next we shall describe the behaviour of the $\psi_{k,l}$'s for $\xi \to \infty$ and at the same time fix some of the $A_{k,l}$'s. From now on we shall use the shorthand notation

(4.1.16) $$\chi = \text{type}(p_1) + \text{type}(p_2)$$

for $\chi = \chi_1 + \chi_2$ with χ_1 of type p_1 and χ_2 of type p_2.

LEMMA 3.

a. $k=0$:

$$\begin{aligned} &\psi_{0,0}=\text{type}\,(2), \\ &\psi_{0,l}=\text{type}\,(l+2)+\text{type}\,(l-1) \quad \textit{for } l\geqq 1. \end{aligned} \tag{4.1.17}$$

b. $k=1$:

$$\begin{aligned} &\psi_{1,0}=\text{type}\,(1), \\ &A_{1,1}\equiv 0 \quad \textit{and} \quad \psi_{1,1}=\text{type}\,(2)+\text{type}\,(-1), \\ &\psi_{1,l}=\text{type}\,(l+1)+\text{type}\,(l-2) \quad \textit{for } l\geqq 2. \end{aligned} \tag{4.1.18}$$

c. $k\geqq 2$:

$$\begin{aligned} &A_{k,0}=0, \\ &\psi_{k,l}=\text{type}\,(l)+\text{type}\,(l-1) \quad \textit{for } l\geqq 0. \end{aligned} \tag{4.1.19}$$

In order to prove this lemma it is clear that we need information about the solutions of

$$\frac{\partial^2\psi}{\partial\xi^2}-\frac{1}{(1+\psi_0)^2}\psi=F \tag{4.1.20}$$

where F is a function of type p, while on top of that, F and all its derivatives are polynomially bounded for $\xi\to-\infty$.

LEMMA 4. *Under these conditions* (4.1.20) *has a particular solution* ψ_{part}, *such that* ψ_{part} *and all its derivatives are polynomially bounded for* $\xi\to-\infty$ *and*

$$\psi_{\text{part}}=\text{type}\,(p+2)+\alpha_0\Phi \tag{4.1.21}$$

with Φ *as in* (4.1.8)-(4.1.10) *of* type 0 *and*

$$\alpha_0=\int_{-\infty}^{\xi_0}\frac{\partial\psi_0}{\partial\xi}(\eta,\theta)F(\eta,\theta)\,d\eta,$$

i.e., α_0 *is uniquely determined by* F *and in general* $\alpha_0\neq 0$. *All other nonexponentially growing solutions are of the form*

$$\psi=\psi_{\text{part}}+A(\theta)\frac{\partial\psi_0}{\partial\xi}.$$

Proof of Lemma 4. The particular solution can be given explicitly as

$$\begin{aligned} \psi_{\text{part}}&=\frac{\partial\psi_0}{\partial\xi}I\left[\left(\frac{\partial\psi_0}{\partial\xi}\right)^{-2}\int_{-\infty}^{\xi}\frac{\partial\psi_0}{\partial\xi}F\,d\zeta\right] \\ &=\frac{\partial\psi_0}{\partial\xi}I\left[\left(\frac{\partial\psi_0}{\partial\xi}\right)^{-2}I\left(\frac{\partial\psi_0}{\partial\xi}F\right)\right]+\alpha_0\Phi. \end{aligned} \tag{4.1.22}$$

It is easy to check that ψ_{part} is smooth and periodic in θ and that ψ_{part} and its derivatives are polynomially bounded for $\xi\to-\infty$. Now it remains to show that ψ_{part} has the correct structure for $\xi\geqq\xi_0$. This requires more accuracy than simply applying the rules given in (4.1.7)-(4.1.14). It is done by the following steps, the details of which are left to the reader:

$$\frac{\partial\psi_0}{\partial\xi}F=\xi\cdot\text{type}\,(p),\qquad \xi\geqq\xi_0,$$

$$I\left(\frac{\partial\psi_0}{\partial\xi}F\right)=\xi^2\,\text{type}\,(p)+C_1\xi^{p+2}\eta_1^{q+1},\qquad \xi\geqq\xi_0.$$

The smooth periodic function C_1 can only be nonzero if $p \geqq 0$ and p even. Then we denote $q = \frac{1}{2}p$.

The same remark holds for C_2, etc., here and below.

$$\left(\frac{\partial\psi_0}{\partial\xi}\right)^{-2} I\left(\frac{\partial\psi_0}{\partial\xi}F\right) = \text{type}\,(p) + C_2\xi^p\eta_1^{q+1}, \qquad \xi \geqq \xi_0,$$

$$I\left[\left(\frac{\partial\psi_0}{\partial\xi}\right)^{-2} I\left(\frac{\partial\psi_0}{\partial\xi}F\right)\right] = \text{type}\,(p+1) + C_3\xi^{p+1}\eta_1^{q+1}, \qquad \xi \geqq \xi_0,$$

$$\frac{\partial\psi_0}{\partial\xi} I\left[\left(\frac{\partial\psi_0}{\partial\xi}\right)^{-2} I\left(\frac{\partial\psi_0}{\partial\xi}F\right)\right] = \text{type}\,(p+2) + C_4\xi^{p+2}\eta_1^{q+1}, \qquad \xi \geqq \xi_0.$$

Since the $\xi^{p+2}\eta_1^{q+1}$ term can be absorbed in type $(p+2)$ our demonstration of (4.1.21) is complete. □

In addition to Lemma 4 we remark that in certain cases the constant α_0 can be calculated explicitly. For example, when $F = \psi_0^{-(1/2)p}$, then F is of type $-(p-2)$ for $p \geqq 3$ and F is bounded. Here we find $\alpha_0 = (1-\frac{1}{2}p)^{-1}\psi_0(-\infty)^{-(1/2)p+1} \neq 0$. In general, the condition $\alpha_0 = 0$ is equivalent to $\int_{-\infty}^{\infty}(\partial\psi_0/\partial\xi)F\,d\xi = 0$, when F is polynomially bounded for $\xi \to -\infty$ and F is of type $p < 0$.

Let us now give the proof of Lemma 3.

Proof of Lemma 3. The inhomogeneous term in (4.1.1) for $\psi_{k,l}$ is of the following form:

$$\hat{f}_{k,l} = \text{(i)} + \text{(ii)} + \text{(iii)} \tag{4.1.23}$$

with

$$\text{(i)} = -\frac{1}{l!} \cdot \frac{\partial^l f}{\partial\xi^l}(0, \theta) \cdot \xi^l \cdot \delta_{k,0},$$

$$\text{(ii)} = \sum_{\substack{0<r+s\leqq 2\\ 1\leqq j\leqq l\\ j\geqq 2-r}} b_j^{r,s} \cdot \xi^{j-2+r}\frac{\partial^{r+s}\psi_{k,l-j}}{\partial\xi^r\partial\theta^s},$$

$$\text{(iii)} = \sum_{2\leqq m\leqq k+l} \frac{1}{(1+\psi_0)^{m+1}} \sum_{\substack{\vec{k},\vec{l}\in(\mathbb{N}\cup\{0\})^m\\ |k|=k,|\vec{l}|=l}} C^m_{\vec{k},\vec{l}} \prod_{j=1}^m \psi_{k_j,l_j}$$

with

$\delta_{k,0} = 1$ if $k = 0$ and $\delta_{k,0} = 0$ otherwise, coefficients $b_j^{r,s}$ smooth and periodic in θ (following easily from $\delta_2 B$, for example $b_j^{2,0} = 0$, $b_2^{0,2} = 1$, etc.) and certain constants $C^m_{\vec{k},\vec{l}}$.

After these preparations the proof lies mainly in an induction argument.

a. *The case* $k = 0$. The statement holds for $l = 0$. Suppose that (4.1.17) has been verified for $0 \leqq l \leqq l_0$. Then, because of Lemma 2 and (4.1.23), $\hat{f}_{0,l_0+1}$ is smooth and $\hat{f}_{0,l_0+1}$ and all its derivatives are polynomially bounded for $\xi \to -\infty$. Moreover,

$$\hat{f}_{0,l_0+1} = \text{type}\,(l_0+1) + \text{type}\,(l_0-2). \tag{4.1.24}$$

For the first two parts of $\hat{f}_{0,l_0+1}$ (i) and (ii) the verification is straightforward. As for (iii), we note that

$$\prod_{j=1}^m \psi_{0,l_j} = \text{type}\,(|\vec{l}|+2m) + \text{type}\,(|\vec{l}|+2m-3),$$

while

$$(1+\psi_0)^{-m-1}=\text{type}\,(-2m)$$

takes care of the correct compensation. Now (4.1.17) with $l=l_0+1$ follows almost immediately when we apply Lemma 4.

b. *The case* $k=1$. For $\psi_{1,0}$ we refer to (3.36). Note that neither $\psi_{1,0}$ nor $\psi_{0,1}$ has a constant term for $\xi\to\infty$. By matching this implies that the corresponding term, $\phi_{1,0}$ in Ω_+ has to vanish on S. We obtain

$$\Delta\phi_{1,0}=0, \qquad \phi_{1,0}=0 \text{ on } S \tag{4.1.25}$$

and the conclusion is that $\phi_{1,0}\equiv 0$. In turn this implies by matching, combined with our knowledge of $\psi_{0,2}$, that $\psi_{1,1}$ cannot have a linear term in its expansion for $\xi\to\infty$. Next an easy calculation shows that

$$\hat{f}_{1,1}=\text{type}\,(0)+\text{type}\,(-3).$$

Consequently, an application of Lemma 4 leads us to (4.1.18) with $l=1$. For higher l's in (4.1.18) we proceed with induction w.r.t. l in a completely analogous fashion.

c. *The case* $k\geqq 2$.

c1. $k\geqq 2$ *and* $l=0$. We observe that $\hat{f}_{k,0}$ consists only of terms coming from (iii). Suppose that (4.1.19) holds for $k\leqq k_0$. Then we find

$$\hat{f}_{k_0+1,0}=\text{type}\,(-2)+\text{type}\,(-3).$$

Using our knowledge of $\psi_{k_0,0}$ and using the fact that there is no order $\delta_1^{k_0+1}\delta_2^{-1}$ term in the expansion in Ω_+, we find that $\psi_{k_0+1,0}$ cannot have a linear term for $\xi\to\infty$, i.e., $A_{k_0+1,0}=0$. Therefore, an application of Lemma 4 provides us with (4.1.19) for $l=0$.

c2. $k\geqq 2$ *and* $l\geqq 1$. This case is again dealt with by induction w.r.t. l with k fixed and thus the proof is complete. □

An important remark is that the contents of Lemma 3 are fully consistent with our assumption (4.2) on the structure of the expansion in Ω_+, i.e., that re-expansion in ρ, θ coordinates of $\delta_1^k\delta_2^l\psi_{k,l}$ with

$$\xi=\rho\delta_2^{-1}, \quad \eta_2=\delta_2^2\rho^{-2}, \quad \eta_1=\delta_2^2\rho^{-2}\ln\rho-\tfrac{1}{2}\delta_1\delta_2\rho^{-2} \tag{4.1.26}$$

gives rise only to order functions as in (3.38)

$$\delta_2^{-2}, \quad \delta_1\delta_2^{-1}, \quad \delta_1^k\delta_2^l, \quad k\geqq 0, \quad l\geqq 0. \tag{4.1.27}$$

However, in view of the matching with the expansion in Ω_+ it will be useful to have somewhat more precise information on the behaviour of the functions, $\psi_{k,l}$ for large values of ζ. Let us introduce the notation

(4.1.28) $E_{k,l}^{\rho,\theta}=_{\text{def}}$ the operator that gives the coefficient corresponding to the $\delta_1^k\delta_2^l$ term in an expansion written in (ρ,θ) coordinates.

The following result will be very useful later on:

LEMMA 5. *Let* k, l *be given as in* (4.1.27), *i.e.* $(k,l)\in\{(0,-2),(1,-1)\}\cup(\mathbb{N}\cup\{0\})^2$. *Set* $N'=k+l$ *and suppose* $K\geqq N'$. *Then*

$$E_{k,l}^{\rho,\theta}\Bigg(\sum_{r+s\leqq K}\delta_1^r\delta_2^s\psi_{r,s}\Bigg)=\Lambda_{k,l}^{\text{sing}}+\Lambda_{k,l}^{\text{const}}+\Lambda_{k,l}^{0,K}. \tag{4.1.29}$$

Here these parts denote the singular, the constant and the vanishing terms for $\rho\downarrow 0$, *respectively. The form of these respective parts is as follows*:

$$\Lambda_{k,l}^{\text{sing}}=\sum_{\substack{0\leqq i\leqq 1+l/2\\ 0\leqq j\leqq l\\ i+j>0}} C_{i,-j}^{k,l}(\theta)\cdot(\ln\rho)^i\rho^{-j}. \tag{4.1.30}$$

Contributions to the $(\ln\rho)^i\rho^{-j}$ *term in* $\Lambda_{k,l}^{\text{sing}}$ *can only come from values* r, s *for which* $r+s=N'-j$ *and* $0\leqq r\leqq k$.

(4.1.31) $$\Lambda_{k,l}^{\text{const}}=C_{0,0}^{k,l}(\theta).$$

Contributions to this term can only come from values r, s *with*

$$r+s=N',\, l\leqq s\leqq 2l+2, \qquad 0\leqq r\leqq k.$$

Finally

(4.1.32) $$\Lambda_{k,l}^{0,K}=\sum_{\substack{0\leqq i\leqq 1+l/2\\ 1\leqq j\leqq K-N'}} C_{i,j}^{k,l}(\theta)(\ln\rho)^i\rho^j.$$

Contributions to the $(\ln\rho)^i\rho^j$ *term in* $\Lambda_{k,l}^{0,K}$ *can only come from values* r, s *with* $r+s=N'+j$, $0\leqq r\leqq k$. *All coefficients* $C_{i,j}^{k,l}$ *are smooth and periodic in* θ.

Proof of Lemma 5. Actually, all we have to do is a little bit of "combinatorics." First, a singular term can only be produced by a term $\sim\delta_1^r\delta_2^s((\ln\xi)^m/\xi^j)$ for $\xi\to\infty$, which after re-expansion gives rise to terms $\sim\delta_1^{r+n}\delta_2^{s+j-n}((\ln\rho)^{m-n}/\rho^j)$ with $0\leqq n\leqq m$. Such a term contributes to $\Lambda_{k,l}^{\text{sing}}$ precisely if $r+n=k$, $s+j-n=l$ and $i=m-n$. Because of the structure of $\psi_{r,s}$ given in Lemma 3 we also know that $j\geqq 2m-s-2$, where equality only holds if $r=0$ and $j=0$, $m>0$. This implies $2l+2=2s+2i+2j+2-2m\geqq 2i+j+s\geqq 2i+j$ and $2i=2m-2n\leqq j+s+2-2n=l+2-n\leqq 1+2$. Thus (4.1.30) is found. A constant contribution can only come from a term for $\xi\to\infty$ proportional to $\delta_1^r\delta_2^s(\ln\xi)^q$, where because of Lemma 3, $q\leqq\frac{1}{2}(s+2)$. After re-expansion this leads to a constant term $\sim\delta_1^{r+1}\delta_2^{s-q}$, which contributes to $\Lambda_{k,l}^{\text{const}}$ if $k=r+q$, $l=s-q$. Hence the contributions to $\Lambda_{k,l}^{\text{const}}$ in (4.1.31) indeed come from pairs (r, s) with $r+s=k+l$, $s\geqq l$, $2l\geqq 2s-(s+2)=s-2$, $0\leqq r\leqq k$. The verification of (4.1.32) is left to the reader as an exercise. □

As we shall see in § 4.3, the following facts will play an important role in the actual constructions: (i) once all $\psi_{r,s}$ with $r+s\leqq k+l$ are known, the constant term $\Lambda_{k,l}^{\text{const}}$ is known; (ii) the linear term in $\Lambda_{k,l}^{0,K}$ comes precisely from those $\psi_{r,s}$ for which $0\leqq r\leqq k$, $r+s=k+l+1$. In anticipation of the full matching later on in § 5, we shall now derive a concrete estimate for the asymptotic behaviour of the $\psi_{k,l}$'s for large ξ.

(4.1.33) $$\Psi^N \underset{\text{def}}{=} \sum_{k+l\leqq N}\delta_1^k\delta_2^l\psi_{k,l},$$

as in (4.1),

(4.1.34) $$\Lambda_{k,l}^K \underset{\text{def}}{=} \Lambda_{k,l}^{\text{sing}}+\Lambda_{k,l}^{\text{const}}+\Lambda_{k,l}^{0,K},\ K\geqq N, \qquad \Lambda^{N,K} \underset{\text{def}}{=} \sum_{\substack{(k,l)=(0,-2)\text{ or}\\ (1,-1)\text{ or}\\ k\geqq 0, l\geqq 0, k+l\leqq N}} \Lambda_{k,l}^K.$$

It is important to have an estimate for the difference between Λ^N and $\Lambda_1^{N,K}$.

LEMMA 6. *There are constants* $C>0$ *such that for* $\xi\geqq\xi_0$

$$\text{(a)}\ \ |\Psi^N(\xi,\theta)-\Lambda^{N,N}(\delta_2\xi,\theta)|\leqq C\left(\delta_1+\delta_2\sqrt{\ln\xi}+\frac{\sqrt{\ln\xi}}{\xi}\right)^{N+1},$$

(4.1.35) $$\text{(b)}\ \ |\Lambda^{N,K}(\delta_2\xi,\theta)-\Lambda^{N,N}(\delta_2\xi,\theta)|\leqq C\xi^2((\delta_1+\delta_2)^{N+1}+(\delta_2\xi)^{K+1}), \qquad K>N,$$

$$\text{(c)}\ \ |\Lambda^{N_1,K}(\delta_2\xi,\theta)-\Lambda^{N,K}(\delta_2\xi,\theta)|\leqq C\ln(\rho)\cdot\left(\delta_1+\frac{\sqrt{\ln\xi}}{\xi}\right)^{N+1}(1+\delta_2\xi)^K,$$

$$k\geqq N_1>N.$$

For the derivatives of these functions r *times w.r.t.* ξ *and* s *times w.r.t.* θ, $r+s\leqq 2$ *analogous estimates hold, but with an extra factor* ξ^{-r} *in the right-hand side.*

Proof of Lemma 6. Here we shall take advantage of the fact that, because of Lemma 3, we can represent $\Psi^N - \Lambda^{N,N}$ by a convergent power series for $\xi \geqq \xi_0$

$$\psi^N - \Lambda^{N,N} = \sum_{k+l\leqq N} \delta_1^k \delta_2^l \left\{ \xi^l \sum{}' a_{rs}^{kl} \left(\frac{\ln \xi}{\xi^2}\right)^r \cdot \left(\frac{1}{\xi}\right)^s \right\}. \tag{4.1.36}$$

In $\sum'$ only these indices r and s are present, which after re-expansion in $\rho = \xi/\delta_2$ yields an order $\delta_1^p \delta_2^q$ with $p+q > N$. Since $p = k+r-m$, $q = r+m+s$ with some m in $0 \leqq m \leqq r$, then

$$k+2r+s \geqq N+1. \tag{4.1.37}$$

We can therefore estimate

$$\begin{aligned} |\Psi^N - \Lambda^{N,N}| &\leqq C \sum_{k+l\leqq N} \delta_1^k \delta_2^l \xi^l \sum_{2r+s=N+1-k} \left(\frac{1}{\xi}\right)^s \\ &\leqq C \sum_{k+l\leqq N+1} \delta_1^k \delta_2^l (\sqrt{\ln \xi})^l \left(\frac{\sqrt{\ln \xi}}{\xi}\right)^{N+1-k-l} \end{aligned} \tag{4.1.38}$$

and this leads us immediately to the first part of (4.1.35). The second part follows from the representation

$$\begin{aligned} \Lambda^{N,K} - \Lambda^{N,N} &= \sum_{N+1\leqq k+l\leqq K} \delta_1^k \delta_2^l \xi^{l+2} \sum_{\text{finite sum}}{}'' a_{rs}^{kl} \left(\frac{\ln \xi}{\xi^2}\right)^r \left(\frac{1}{\xi}\right)^s, \\ \Lambda^{N_1,K} - \Lambda^{N,K} &= \sum_{N+1\leqq k+l\leqq N_1} \delta_1^k \delta_2^l \sum_{\substack{0\leqq i\leqq 1+l/2 \\ -l\leqq j\leqq K-(k+l)}} C_{i,j}^{k,l} (\ln \rho)^i \rho^j. \end{aligned} \tag{4.1.39}$$

It is left to the reader to fill in further details. □

To conclude this section we shall derive an estimate for the error up to which Ψ^N satisfies the equation (1.1).

LEMMA 7. *There is an ε-independent constant $\rho_1 > 0$ such that for $|\xi| \leqq \rho_1/\delta_2$*

$$\left| -\varepsilon \Delta \Psi^N + \frac{\Psi^N}{1+\Psi^N} - f \right| \leqq M(\delta_1 + \delta_2 |\xi|)^{N+1} \tag{4.1.40}$$

with an ε-independent constant M.

Proof of Lemma 7. The remainder $r_N = -\delta_2^2 \Delta \Psi^N + \Psi^N/(1+\Psi^N) - f$ satisfies, by construction of the $\Psi_{k,l}$'s,

$$T_N r_N = 0, \quad \text{i.e.,}\ r_N = r_N - T_N r_N \tag{4.1.41}$$

where T_N denotes the Taylor series expansion w.r.t. δ_1 and δ_2 up to orders $\delta_1^k \delta_2^l$ with $k+l \leqq N$.

Hence

$$r_N = -(I - T_N) f(\delta_2 \xi, \theta) - (I - T_N) \delta_2^2 \Delta \Psi^N - (I - T_N)(1+\psi_N)^{-1}. \tag{4.1.42}$$

As for the first term the statement in (4.1.41) holds trivially.

For the second term we note that

$$-\delta_2^2 \Delta = \sum_{0<r+s\leqq 2} \delta_2^{2-r} B^{r,s}(\delta_2 \xi, \theta) \frac{\partial^{r+s}}{\partial \xi^r \partial \theta^s}, \tag{4.1.43}$$

where the $B^{r,s}$ can easily be identified using (3.2). Consequently,

$$(I - T_N) \delta_2^2 \Delta \Psi^N = \sum_{\substack{0<r+s\leqq 2 \\ k+l\leqq N}} \sum_{j=0}^{l-2+r} (I - T_j) B^{r,s} \cdot \delta_1^k \delta_2^{l-j} \frac{\partial^{r+s} \psi_{k,l-j-2+r}}{\partial \xi^r \partial \theta^s}. \tag{4.1.44}$$

Now

$$|(I-T_j)B^{r,s}| \leqq C(\delta_2\xi)^{j+1} \quad \text{and} \quad \left|\frac{\partial^{r+s}\psi_{k,l-j-2+r}}{\partial\xi^r\partial\theta^s}\right| \leqq C|\xi|^{l-j},$$

because of Lemmas 2 and 3, and there is an ε-independent constant $C>0$ such that

$$(4.1.45) \qquad |(I-T_N)\delta_2^2\Delta\Psi^N)| \leqq C \sum_{k+l\leqq N} \delta_1^k(\delta_2|\xi|)^{l+1}.$$

This implies that the second term is as required in (4.1.40).

For the third term we proceed as follows:

$$(4.1.46) \qquad \begin{aligned} |(I-T_N)(1+\psi_N)^{-1}| &= \left|\frac{1}{N!}\int_0^1 (1-t)^N \frac{\partial^{N+1}}{\partial t^{N+1}}\left(1+\sum_{k+l\leqq N} t^{k+l}\delta_1^k\delta_2^l\psi_{k,l}\right)^{-1} \cdot dt\right. \\ &\leqq C \sum_{k+l=N+1} \delta_1^k\delta_2^l \sum_{m=2}^{N+1}\left(1+\frac{1}{2}\psi_0\right)^{-m-1} \prod_{\substack{i=1 \\ k_i+l_i\geqq 1 \\ \Sigma k_i=k, \Sigma l_i=l}}^{m} |\psi_{k_i,l_i}| \end{aligned}$$

with some suitably chosen ε-independent constant C. Here we used Lemmas 2 and 3 to estimate $1+\psi_N \geq 1+\frac{1}{2}\psi_0$ for $|\xi| \leqq \rho_1/\delta_2$. Now, for $\xi \leqq \xi_0$ we estimate $1+\frac{1}{2}\psi_0 \geqq 1$ and $|\prod_{i=1}^m \psi_{k_i,l_i}| \leqq C(1+|\xi|^1)$ using Lemma 2. For $\xi \geqq \xi_0$ we estimate $1+\frac{1}{2}\psi_0 \geqq C\xi^2$ and $\prod_{i=1}^m |\psi_{k_i,l_i}| \leqq C\xi^{l+2m}$ and again the estimate as required for (4.1.40) is easily found. □

Note that Lemma 7 shows that we have constructed a formal approximation of arbitrary order in the layer along the free surface.

4.2. The $\phi_{k,l}$'s with free $g_{k,l}$'s. Let us now come back to the construction of a formal approximation in Ω_+. For $\phi_{k,l}$ we find by substituting (4.2) in (1.1) and collecting the terms of order $\varepsilon\delta_1^k\delta_2^l$ an equation of the following form:

$$(4.2.1) \qquad \Delta\phi_{k,l} = -\bar{F}_{k,l}$$

with

$$\bar{F}_{k,l} = (k!\,l!)^{-1}\frac{\partial^{k+l}}{\partial\delta_1^k\partial\delta_2^l}\{w_0+\delta_1\delta_2 w_1+\delta_2^2(1+\Sigma'\delta_1^r\delta_2^s\phi_{r,s})\}^{-1}$$

where $'$ denotes summation over all indices r, s such that $0 \leqq r \leqq k$, $0 \leqq s \leqq l-2$, $r+s < k+l-2$. For example, the equation for $\phi_{0,0}$ becomes

$$(4.2.2) \qquad \Delta\phi_{0,0} = -\frac{1}{w_0}.$$

Since w_0 behaves as ρ^2 for $\rho\downarrow 0$ (see (1.5)) the right-hand side of this equation is singular for $\rho\downarrow 0$. Analogously, singularities can be expected in the other $\bar{F}_{k,l}$'s. Consequently, it is logical to split $\phi_{k,l}$ into a singular part and a regular part. Now, it will be crucial to the construction that the singular part of $\phi_{k,l}$ can be determined in *a unique way* from the previous $\phi_{\bar{k},\bar{l}}$ with $\bar{k} \leqq k$, $\bar{l} \leqq l-2$, $\bar{k}+\bar{l} < k+l-2$ by an iterative procedure *using only* the equation given in (4.2.1). The regular part of $\phi_{k,l}$ will have freedom in the form of its boundary values on S. Moreover, the following lemma provides us with more information on the structure of the singularities near the free surface.

LEMMA 8. *The equations given in* (4.1.1) *have a set of solutions $\phi_{k,l}$ with the following properties*:

$$(4.2.3) \qquad \phi_{k,l} = \phi_{k,l}^{s,M}(\rho,\theta)\cdot H\left(\frac{\rho-\rho_1}{\rho_1}\right)+\phi_{k,l}^{r,M},$$

where $\phi_{k,l}^{s,M}$ will consist of certain singular terms and $\phi_{k,l}^{r,M}$ is sufficiently regular:

$$(4.2.4)\qquad \phi_{k,l}^{s,M}=\sum_{\substack{-l\leqq j\leqq M+2\\ 0\leqq i\leqq 1+l/2\\ i>0 \text{ or } j<0}} \Phi_{i,j}^{k,l}(\theta)\cdot(\ln\rho)^i\rho^j,\qquad \Phi_{i,j}^{k,l} \textit{ smooth and periodic in } \theta,$$

$$(4.2.5)\qquad \phi_{k,l}^{r,M}\in C^{M+2+\alpha}(\bar{\Omega}_+),\qquad \alpha\in(0,1).$$

In this decomposition $M\in\mathbb{N}\cup\{0\}$ is arbitrary. H is the cut-off function given in (2.6) *and $\rho_1>0$ is ε-independent, but sufficiently small. The construction of the $\phi_{k,l}$'s goes iteratively, namely, once all $\phi_{k,l}$ with $k+l\leqq N$ are known with properties as in* (4.2.3)–(4.2.5), *then for $k+l=N+1$:*

(i) *$\phi_{k,l}^{s,M}$ follows uniquely within the specified class by requiring:*

$$(4.2.6)\qquad \Delta\phi_{k,l}^{s,m}+\bar{F}_{k,l}\in C^{M+\alpha}(\{(\rho,\theta)\,|\,0\leqq\rho\leqq\rho_1\}),$$

(ii) *solution of*

$$(4.2.7)\qquad \Delta\phi_{k,l}^{r,M}=-\bar{F}_{k,l}-\Delta(\phi_{k,l}^{s,M}H),$$

$$(4.2.8)\qquad \phi_{k,l}^{r,M}=g_{k,l},$$

where $g_{k,l}\in C^\infty(S)$ is free.

Proof of Lemma 8. Working out the differentiations in the definition of $\bar{F}_{k,l}$ leads us to

$$(4.2.9)\qquad \bar{F}_{k,l}=\sum_{\substack{|\vec{s}|\leqq l-2m-2r-p\\ |\vec{r}|=k-p\\ m\geqq 1,0\leqq n\leqq m}} J_{\vec{r},\vec{s}}^{k,l,p,m,n} w_0^{-m-1} w_1^p \prod_{\substack{n\leqq m \text{ terms}\\ n=0 \text{ is allowed}}} \phi_{r_i,s_i}$$

with certain constants $J_{\vec{r},\vec{s}}^{k,l,p,m,n}$. If we assume that the lemma is true for $k+l\leqq N$, a simple calculation shows that for a pair (k,l) with $k+l=N+1$

$$(4.2.10)\qquad \bar{F}_{k,l}=\bar{F}_{k,l}^{s,M}H+\bar{F}_{k,l}^{r,m}$$

with

$$\bar{F}_{k,l}=\sum_{\substack{-(l+2)\leqq j\leqq M\\ 0\leqq i\leqq 1/2l\\ j<0 \text{ or } i>0}} \bar{J}_{i,j}^{k,l}(\ln\rho)^i\rho^j,\qquad \bar{J}_{i,j}^{k,l} \text{ smooth and periodic in } \theta,$$

and

$$\bar{F}_{k,l}^{r,M}\in C^{M+\alpha}(\bar{\Omega}_+),\qquad \alpha\in(0,1),$$

where $M\in\mathbb{N}\cup\{0\}$ is arbitrary. Using (4.2.2) we see that (4.2.10) also holds in the case $k=0$, $l=0$, where the iteration is starting.

Our next step is to solve the equation

$$(4.2.11)\qquad \Delta\phi_{k,l}^{s,M}=-\bar{F}_{k,l}^{s,M}+O((\ln\rho)^{1+\frac{1}{2}l}\cdot\rho^{M+1}).$$

This really involves no more than some linear algebra to determine the coefficients $\Phi_{i,j}^{k,l}$. In a straightforward way one can compute that these coefficients have to satisfy a system of linear equations of the following type:

$$(4.2.12)\qquad E_{ij}\Phi_{ij}^{kl}=\sum_{(i',j')>_s(i,j)}\sum_{s=0}^{2} D_{i',j'}^{k,l,s}\frac{\partial^s\Phi_{i',j'}^{k,l}}{\partial\theta^s}-\bar{J}_{i-\delta(j),j-2}^{k,l}$$

with

$$E_{ij}=\begin{cases} j(j-1) & \text{if } j<0 \text{ or } j\geqq 2,\\ (-1)^{j+1}i & \text{if } j=0 \text{ or } j=1,\end{cases}$$

$$\delta(j)=\begin{cases} 1 & \text{if } j=0 \text{ or } j=1,\\ 0 & \text{else,}\end{cases}$$

$$(i',j')\underset{s}{>}(i,j)\quad \text{if } |(\ln\rho)^i\rho^j|=o(|(\ln\rho)^{i'}\rho^{j'}|)\quad \text{for } \rho\downarrow 0,$$

i.e., if $[j'<j]$ or $[j=j'$ and $i'>i]$.

Note that $>_s$ is a linear ordering on the pairs (i,j) and that the system of equations in (4.2.12) can be solved recursively starting with $(i,j)=(1+l/2,-l)$, corresponding to the principal term w.r.t. this ordering and then going down in accordance with the ordering. In this way $\phi_{k,l}^{s,M}$ is uniquely determined and (4.2.6) holds. Next, we conclude that the right side of (4.2.7) is in $C^{M+\alpha}(\bar{\Omega}_+)$. Consequently, (4.2.5) holds (see [14], [15]). Thus the solutions of (4.2.1) defined in (4.2.3) have indeed the properties specified in Lemma 3. □

Let us now define

$$\text{(4.2.13)}\qquad \Phi^N=\varepsilon^{-1}w_0+(\ln\varepsilon)\cdot w_1+\sum_{k+l\leqq N}\delta_1^k\delta_2^l\phi_{k,l} \text{ as in (4.2).}$$

We conclude this section with the following error estimate indicating how well Φ^N satisfies (4.1.1).

LEMMA 9. *On the domain* $\{x\in\Omega_+\,|\,$*distance* $(x,S)\geqq\delta_1\}$ *we have*

$$\text{(4.2.14)}\qquad \left|-\varepsilon\Delta\Phi^N+\frac{\Phi^N}{1+\Phi^N}-f\right|\leqq K\cdot\frac{\varepsilon}{\rho^2}\cdot\left(\delta_1+\frac{\delta_2}{\rho}\right)^{N+1}$$

with an ε*-independent constant* $K>0$ *and* $\rho=d(x,S)$.

Proof of Lemma 9. We note that the error is given by

$$\left|\varepsilon\cdot\frac{1}{N!}\int_0^1(1-t)^N\cdot\frac{\partial^{N+1}}{\partial t^{N+1}}\left\{w_0+\delta_1\delta_2t^2w_1+\delta_2^2t^2\left(1+\sum_{r+s\leqq N}\delta_1^r\delta_2^s\phi_{r,s}\right)\right\}^{-1}\cdot dt\right|,$$

which for $d(x,S)\geqq\delta_1$ can be estimated as

$$\leqq K_1\varepsilon\sum_{\substack{k+l=N+1\\ |\vec{s}|\leqq l-2m-2r-p\\ |\vec{r}|=k-p}}\delta_1^k\delta_2^l\left|w_0^{-m-1}w_1^p\prod_{\substack{n\geqq m\\ \text{terms}}}\phi_{r_i,s_i}\right|$$

$$\leqq K_2\varepsilon\sum_{\substack{k+l=N+1\\ |\vec{s}|\leqq l-2m}}\delta_1^k\delta_2^l(\rho^2)^{-m-1}\cdot\left(\frac{1}{\rho}\right)^{|\vec{s}|}$$

and then (4.2.14) follows easily. □

This lemma demonstrates that the construction yields a formal approximation of arbitrary order in Ω_0.

4.3. The scheme leading to uniquely defined $\psi_{k,l}$'s and $\phi_{k,l}$'s with determined $A_{k,l}$'s and $g_{k,l}$'s. Of course we still have to determine a bunch of free "constants" in the expansion in the layer along S and in the expansion in Ω_+. This is done by a suitable matching argument. In § 5 we shall show that this leads us to a layer expansion along S and an expansion in Ω_+ with an overlapping domain of validity.

In order to fix the free $A_{k,l}$'s, we notice that matching of the layer expansion and the expansion in Ω_+ requires that

- $g_{k,l}$ is the constant term of order $\delta_1^k\delta_2^l$ found by re-expansion of the layer in (ρ, θ) coordinates,

- the linear term in $\phi_{k,l}$ near S coincides with the linear term found in the $O(\delta_1^k\delta_2^l)$ term obtained by re-expanding the layer in (ρ, θ) coordinates. Note that the term $\sim\rho$ of order $\delta_1^k\delta_2^l$ in the layer expansion has a coefficient

$$(1-\hat{f})A_{k,l+1}+B_{k,l+1}$$

where $B_{k,l+1}$ is completely determined by the functions $\psi_{r,s}$ with $0 \leqq r < k$, $r+s=k+l+1$ or with $r+s \leqq k+l$ (see (4.1.32) and Lemma 4). Hence $A_{k,l+1}$ has to satisfy the following equation:

$$A_{k,l+1}=(1+\hat{\gamma})\{\Phi_{0,1}^{k,l}-B_{k,l+1}\} \tag{4.3.1}$$

with $\Phi_{0,1}^{k,l}$ as in (4.2.4).

Consequently, the scheme that provides us with the $\psi_{k,l}$'s and $\phi_{k,l}$'s in a unique way runs as follows. We start with $N=0$ and then successively

$$\begin{tikzcd}
\boxed{\text{the } \psi_{k,l}\text{'s with } k+l=N \text{ known}} \arrow[r] & \Lambda_{k,l}^{\text{const}} \text{ is known (see (4.1.31))} \arrow[d] \\
 & g_{k,l} \underset{\text{def}}{=} \Lambda_{k,l}^{\text{const}} \arrow[d] \\
\begin{array}{l}\text{all } \Phi_{0,1}^{k,l}\text{'s with } k+l=N \\ \quad \text{are known (see (4.2.4))}\end{array} \arrow[d] & \boxed{\text{the } \phi_{k,l}\text{'s with } k+l=N \text{ known}} \arrow[l] \\
A_{0,N+1} \text{ is known (see (4.3.1))} \arrow[d] & \\
A_{1,N} \text{ is known (see (4.3.1))} & \\
\vdots \arrow[d] & \\
A_{N,1} \text{ is known (see (4.3.1))} & \\
A_{N+1,0}=0 \text{ (see Lemma 3)} \arrow[d] & \\
\boxed{\text{the } \psi_{k,l}\text{'s with } k+l=N+1 \text{ known}} \arrow[r] & \text{etc.}
\end{tikzcd}$$

This scheme leads to $g_{k,l}$ and $A_{k,l}$, which are smooth and periodic in θ. It will be clear that only a very partial matching between the layer expansion and the expansion in Ω_+ has been built into this scheme. In the next section we consider the matching relations in more detail.

5. Matching relations of the internal layer at S and the expansions in Ω_0, Ω_+.

This section is a preparation for the construction of a global uniformly valid expansion in $\bar{\Omega}$ and the demonstration of its validity. We show here that various expansions constructed in preceding sections have overlap domains in which their differences are small. First we shall show that the internal layer expansion Ψ^{N+2} and the regular expansion U^N up to terms of the order δ_2^{2N} have an overlap domain, and we shall estimate their difference there.

LEMMA 10. *There are ε-independent constants ρ_1, ξ_1, M and $\mu>0$ such that for* $-\rho_1 \leqq \rho \leqq -\delta_2\xi_1$

$$|U^N-\Psi^{N+2}| \leqq M\{(\delta_2+\rho)^{N+2}+\exp(-\mu\rho/\delta_2)\} \tag{5.1}$$

and

$$\left|\rho^r \frac{\partial^{r+s}}{\partial\rho^r\partial\theta^s}(U^N - \Psi^{N+2})\right| \leqq M\{(\delta_2+\rho)^{N+2} + \exp(-\mu\rho/\delta_2)\}, \qquad r+s \leqq 2.$$

Note that overlap up to a certain order takes place in any region $-\hat{\rho}_1(\varepsilon) \leqq \rho \leqq -\hat{\rho}_2(\varepsilon)$ with $\hat{\rho}_1 = o(1)$ and $\delta_2 = o(\hat{\rho}_2)$. The difference is almost as small as possible in a region $-\hat{A}\delta_1 \leqq \rho \leqq -\hat{B}\delta_1$ with suitably chosen ε-independent constants $\hat{A}$, $\hat{B} > 0$.

Proof of Lemma 10. This proof is not very difficult. It uses the observation that there is exactly one function $P^{\bar{N}}$, $\bar{N} =_{\text{def}} N+1$ of the form

$$(5.2) \qquad P^{\bar{N}} = \sum_{l+j \leqq N} p_{l,j}(\theta)\delta_2^l \rho^j$$

with smooth, periodic coefficients $p_{l,j}$, such that

$$(5.3) \qquad \left|-\delta_2^2 \Delta P^{\bar{N}} + \frac{P^{\bar{N}}}{1+P^{\bar{N}}} - f\right| = o(\delta_2 + |\rho|)^{\bar{N}} \quad \text{for } \delta_2 \downarrow 0, \rho \uparrow 0.$$

On the other hand, both $T_{\bar{N}} U^N$, the Taylor expansion w.r.t. δ_2 and ρ up to terms $\delta_2^l \rho^j$ with $l+j \leqq \bar{N}$ of U_N, and $Q^{\bar{N}} = \sum_{l=0}^{\bar{N}} \delta_2^l P_l(\rho/\delta_2, \theta)$ with P_l the polynomial found in the expansion of $\psi_{0,1}$ for $\xi \to -\infty$ (see Lemma 2), satisfy (5.2) and (5.3) (see (2.13) and Lemma 7). Hence $T_{\bar{N}} U^N = Q^{\bar{N}} = P^{\bar{N}}$. Further, $|U_N - T_{\bar{N}} U_N| \leqq M_1(\delta_2+\rho)^{\bar{N}+1}$, $|\Psi^{\bar{N}} - \Phi^{\bar{N}}| \leqq M_2 \exp(-\mu\rho/\delta_2)$ and $|\Psi^{N+2} - \Psi^N| \leqq M_1(\delta_2+\rho)^{\bar{N}+1}$ (see Lemma 2). Thus (5.1) has been derived. The estimates for the derivatives follow in an analogous way. □

Next we shall demonstrate that Ψ^{N+2} and Φ^N, the expansion in Ω_+, have an overlapping domain of validity. It is no surprise that in this case we have to work harder to get an appropriate estimate for the difference.

LEMMA 11. *There are ε-independent constants ρ_1, ξ_1, $M > 0$ such that for $\xi_1 \leqq \xi \leqq \rho_1/\delta_2$*

$$(5.4) \quad |\Psi^{N+2} - \Phi^N| \leqq M\left\{\left(\delta_1 + \delta_2\sqrt{\ln \xi} + \frac{\sqrt{\ln \xi}}{\xi}\right)^{N+1} \cdot \ln(\rho) + \xi^2 \cdot (\delta_1 + \delta_2 \xi)^{N+3}\right\}$$

and such an estimate also holds true for

$$\left|(\delta_2\xi)^r \frac{\partial^{r+s}}{\partial\rho^r\partial\theta^s}(\Psi^{N+2} - \Phi^N)\right|, \qquad r+s \leqq 2.$$

Proof of Lemma 11. We split the difference as

$$(5.5) \qquad \Psi^{N+2} - \Phi^N = (\Psi^{N+2} - \Lambda^{N+2,K}) + (\Lambda^{N,K} - \Phi^N) + (\Lambda^{N+2,K} - \Lambda^{N,K})$$

with $K > N$. The first and the last part have already been estimated in a way required by (5.4) (see Lemma 6).

An estimate of the second part is found by applying the following result:

for each (k, l) and for each $K_1 > 0$ there exists a $K > 0$ such that

$$(5.6) \qquad \begin{aligned} &|\Delta(\Lambda_{k,l}^K - \phi_{k,l})| = O(\rho^{K_1}) \quad \textit{for } \rho \downarrow 0, \\ &|(\Lambda_{k,l}^K - \phi_{k,l})| = O(\rho^{K_1+2}) \quad \textit{for } \rho \downarrow 0. \end{aligned}$$

Once (5.6) has been demonstrated, it is clear that the second part is such that there is an $\tilde{M} > 0$ independent of ε such that

$$(5.7) \qquad |\Lambda^{N,K} - \Phi^N| \leqq M\rho^{N+5}$$

by taking K sufficiently large. Therefore the estimate in (5.4) is complete.

Let us now derive (5.6). This is done by induction w.r.t. $N_0 =_{\text{def}} k+1$. We start with:

a. $N_0=-2$, i.e. $k=0$, $l=-2$. Using Lemma 7 and the first part of Lemma 6 (with N replaced by K) we find that

$$\Delta(\Lambda^K_{0,-2}) = 1 - f + O(\rho^{K+1}). \tag{5.8}$$

As a consequence of Lemma 5 and (4.1.17) we obtain

$$\Lambda^K_{0,-2}\Big|_{\rho=0} = 0, \frac{\partial \Lambda^K_{0,-2}}{\partial\rho}\Big|_{\rho=0} = 0. \tag{5.9}$$

Together with the problem for w_0 in (1.5), this implies that

$$\begin{gathered}\Delta(w_0 - \Lambda^K_{0,-2}) = O(\rho^{K+1}),\\ (w_0-\Lambda^K_{0,-2})|_{\rho=0}=0, \qquad \frac{\partial}{\partial\rho}(w_0-\Lambda^K_{0,-2})|_{\rho=0}=0.\end{gathered} \tag{5.10}$$

Moreover, the r.h.s. of the equation is $C^{K+1+\alpha}$ for $\rho \geqq 0$, ρ sufficiently small, because of Lemma 5 and Lemma 8. Using a Cauchy-Kowalewsky type of construction (cf. [15]) we find that

$$|w_0 - \Lambda^K_{0,-2}| = O(\rho^{K+3}) \quad \text{for } \rho \downarrow 0. \tag{5.11}$$

b. Suppose (5.6) has been proved for $k+l \leqq N_0$. Consider, then, values such that $k+l=N_0+1$, $k+l=0$ if $N_0=-2$. Given $K_1>0$, choose K so large that

$$|\Lambda^K_{r,s} - \phi_{r,s}| = O(\rho^{N_0+1+K_1})$$

for all r, s with $r+s \leqq N_0$. From the structure of $\bar{F}_{k,l}$ as given in (4.2.1) it follows that

$$\Delta\phi_{k,l} = -\bar{F}_{k,l} = -\bar{F}^*_{k,l} + O(\rho^{K_1}) \tag{5.12}$$

where $\bar{F}^*_{k,l}$ denotes $\bar{F}_{k,l}$ with w_0, w_1, $\phi_{r,s}$, $r+s \leqq N_0$ replaced by their approximation $\Lambda^K_{r,s}$, $r+s\leqq N_0$. Again using Lemma 7 and the first part of Lemma 6 with $N \to K$, we obtain

$$\Delta\Lambda^K_{k,l} = -\bar{F}^*_{k,l} + O(\rho^{K_1}). \tag{5.13}$$

Because of the uniqueness of the singular terms $(\ln\rho)^i\rho^j$ with $j<0$ or $i>0$ in the solution of $\Delta\phi = -\bar{F}^*_{k,l}$ (see (4.2.12)), we obtain

$$\begin{gathered}\phi_{k,l} - \Lambda^K_{k,l} \text{ is } C^{K_1+2+\alpha} \quad \text{for } \rho \geqq 0, \rho \text{ sufficiently small},\\ \Delta(\phi_{k,l} - \Lambda^K_{k,l}) = O(\rho^{K_1}).\end{gathered} \tag{5.14}$$

Due to the partial matching relations imposed in § 4.3 we have as boundary conditions on S

$$(\phi_{k,l} - \Lambda^K_{k,l})|_{\rho=0} = 0, \qquad \frac{\partial}{\partial\rho}(\phi_{k,l}-\Lambda^K_{k,l})|_{\rho=0} = 0. \tag{5.15}$$

Again a construction of Cauchy-Kowalewsky type shows that

$$|\phi_{k,l} - \Lambda^K_{k,l}| = O(\rho^{K_1+2}) \quad \text{for } \rho\downarrow 0. \tag{5.16}$$

The estimates on the derivatives can now also be verified in an elementary way; details are left to the reader. □

We conclude this section with the remark that for $N \geqq 0$ the overlap region is nonempty and that the difference is small in the region

$$A\sqrt{\delta_2} \leqq \rho \leqq B\sqrt{\delta_2}, \quad \text{i.e., } A\delta_2^{-1/2} \leqq \xi \leqq B\delta_2^{-1/2} \tag{5.17}$$

with suitably chosen ε-independent A and $B>0$. This region is special, since exactly in this region Ψ^{N+2} and Φ^N give rise to the same order of error in the equation;

$$\text{error} \sim (\delta_2\xi)^{N+3} = O_{\text{sharp}}\left(\frac{\delta_2}{\rho}\right)^{N+3} = O_{\text{sharp}}(\varepsilon^{(N+3)/4}) \tag{5.18}$$

in that region (see Lemma 7 and Lemma 9). In fact it is this error, which is worst in the global formal approximation, that we discuss in the next section.

6. Composition of a global approximation Z_N and estimation of the error $u-Z_N$. Now we are in a position to put all the previously constructed local approximations together into a global approximation Z_N. We define

$$Z_N = \Psi^{N+2} H\left(\frac{\rho^2}{\sqrt{\varepsilon}}\right) + (Z_0^N + \Phi^N)\cdot\left[1 - H\left(\frac{\rho^2}{\sqrt{\varepsilon}}\right)\right] \tag{6.1}$$

with

Ψ^{N+2} the expansion in the layer along with the free surface S, see (4.1.33),

Z_0^N the regular expansion in Ω_0 corrected with the layer expansion along $\partial\Omega$, see (2.4),

Φ^N the expansion in Ω_+ (see (4.2.13)),

$H\left(\frac{\rho^2}{\sqrt{\varepsilon}}\right)$ is a smooth cut-off function, which is $\equiv 1$ for $|\rho| \leqq (\frac{1}{2}\sqrt{\varepsilon})^{1/2}$ and which vanishes for $|\rho| \geqq \varepsilon^{1/4}$ (see (2.6)).

THEOREM 1. *The construction process has been successful in the sense that Z_N is a global, formal approximation*:

$$\max_{x\in\bar{\Omega}}\left|-\varepsilon\Delta Z_N + \frac{Z_N}{1+Z_N} - f\right| \leqq R\cdot\sqrt{\varepsilon^{1/2}\ln\left(\frac{1}{\varepsilon}\right)}^{(N+3)}, \tag{6.2}$$

where R is an ε-independent constant >0.

Proof of Theorem 1. The estimate in (6.2) can be derived by making a subdivision of $\bar{\Omega}$ into

$$\Omega_l = \{(\rho,\theta) \mid \rho^2 \leqq \tfrac{1}{2}\sqrt{\varepsilon}\}, \quad \Omega_{l,+} = \{(\rho,\theta) \mid \rho>0, \tfrac{1}{2}\sqrt{\varepsilon} \leqq \rho^2 \leqq \sqrt{\varepsilon}\},$$

$$\Omega_{l,0} = \{(\rho,\theta) \mid \rho<0, \tfrac{1}{2}\sqrt{\varepsilon} \leqq \rho^2 \leqq \sqrt{\varepsilon}\}, \quad \Omega_+' = \Omega_+\backslash\{\Omega_l \cup \Omega_{l,+}\}, \quad \Omega_0' = \bar{\Omega}_0\backslash\{\Omega_l \cup \Omega_{l,0}\}.$$

Partial results in the direction of (6.2) were already found. In Ω_0' we use (2.15) and obtain an $O(\varepsilon^{N+3})$ error. The larger contributions come from $\Omega_l, \Omega_{l,+}$ and Ω_+'. We apply Lemma 7 and Lemma 9 in Ω_l and $\Omega_{l,+}$ respectively. In both regions we find an $O(\varepsilon^{1/4(N+3)})$ error, because of the special choice of the cut-off at an $O(\varepsilon^{1/4})$-distance from S (compare (5.18)). In the overlap region $\Omega_{l,+}$ we have

$$Z_N = \Phi^N + H\left(\frac{\rho^2}{\sqrt{\varepsilon}}\right)(\Psi^{N+2} - \Phi^N).$$

Using Lemma 9 and the matching relations in Lemma 11, we can see that here the error is at most $(\sqrt{\delta_2|\ln\varepsilon|})^{N+3}$. Specially, the term

$$\delta_2^2\cdot\left(\partial^2 H\left(\frac{\rho^2}{\sqrt{\varepsilon}}\right)\Big/\partial\rho^2\right)\cdot(\Psi^{N+2} - \Phi^N)$$

gives rise to this order (see Lemma 11). Other terms yield contributions of a smaller order. The region $\Omega_{l,0}$ is dealt with in an analogous way using Lemma 10. □

The next step is to convert the estimate in (6.2) into an estimate for $u-Z_N$. This can be done with the following result:

LEMMA 12. *Suppose that* $Z\in C^\infty(\bar{\Omega})$, $Z\geqq 0$ *satisfies*

$$-\varepsilon\Delta Z+\frac{Z}{1+Z}-f=r,\qquad Z=0 \text{ on } \partial\Omega. \tag{6.3}$$

Then there is an ε-independent constant $\nu>0$ *such that*

$$|u-Z|_{\sup}\leqq\frac{\nu}{\varepsilon}|r|_{\sup}. \tag{6.4}$$

Proof of Lemma 12. We use a technique based on barrier functions and the maximum principle (see [17], [12] and [18]). Suppose that a positive function $v\in C^\infty(\bar{\Omega})$ can be found such that

$$\begin{aligned}&-\varepsilon\Delta(Z+v)+\frac{Z+v}{1+Z+v}-f\geqq 0\quad\text{in }\bar{\Omega},\\ &Z+v\geqq 0\quad\text{on }\partial\Omega.\end{aligned}\tag{6.5}$$

Then $Z+v$ is an upper barrier for the solution u, i.e.,

$$u\leqq Z+v\quad\text{on }\bar{\Omega}. \tag{6.6}$$

To prove (6.6) we use the fact that $Z+v-u=_{\text{def}}w$ satisfies $w\in C^\infty(\bar{\Omega})$ and

$$\begin{aligned}&-\varepsilon\Delta w+\bar{c}w\geqq 0\quad\text{in }\bar{\Omega},\\ &w\geqq 0\quad\text{on }\partial\Omega\end{aligned}$$

with

$$\bar{c}=(1+Z+v)^{-1}(1+u)^{-1},$$

i.e.,

$$\bar{c}\in C^\infty(\bar{\Omega})\quad\text{and}\quad\bar{c}\geqq 0.$$

Implicitly, we used the fact that $u\in C^\infty(\bar{\Omega})$ (cf. [4]). The maximum principle for second order elliptic Dirichlet problems (cf. [19]) implies that $w\geqq 0$.

If a negative function $v\in C^\infty(\bar{\Omega})$ can be found such that (6.5) holds with reversed signs in the region where $Z+v\geqq 0$, then $\max(0,Z+v)$ is a lower barrier for the solution u. This follows from (1.3) in combination with a maximum principle argument.

Now, we make the following choice for the function v:

$$v=b\cos\left(\frac{\pi}{2}\cdot\frac{x_1-x_1^0}{d}\right)\cdot\cos\left(\frac{\pi}{2}\cdot\frac{x_2-x_2^0}{d}\right) \tag{6.7}$$

with a value for $b\geqq 0$, which we shall specify presently. The point x^0 and the number $d>0$ are chosen in such a way that the product of the cosines is $\geqq\frac{1}{2}$ on $\bar{\Omega}$. With this function v we obtain that

$$-\varepsilon\Delta(Z+v)+\frac{Z+v}{1+Z+v}-f\geqq r-\varepsilon\Delta v\geqq r+b\cdot\frac{\varepsilon}{2}\left(\frac{\pi}{2d}\right)^2, \tag{6.8}$$

which is $\geqq 0$, when $b=(\nu/\varepsilon)\min(0,-\min_{\bar{\Omega}} r)$ with $\nu=8(d/\pi)^2$. In this way an upper bound is found. For the lower bound we proceed in an analogous way. Actually we even obtain a somewhat better estimate than we do from (6.3):

$$\max\left(0,Z-\frac{\nu}{\varepsilon}\max(0,\max_{\bar{\Omega}} r)\right)\leqq u\leqq Z+\frac{\nu}{\varepsilon}\min(0,-\min_{\bar{\Omega}} r). \tag{6.9}$$

□

An immediate consequence of Lemma 12 and Theorem 1 is

$$|u-Z_N|_{\sup}=O\left(\varepsilon^{-1}\left(\sqrt{\varepsilon^{1/2}\ln\left(\frac{1}{\varepsilon}\right)}\right)^{N+3}\right)\quad \text{if } N\geqq 0. \tag{6.10}$$

However, this estimate can be improved somewhat. It follows that

$$|Z_N-Z_{N_1}|\sup=O\left(\left(\sqrt{\varepsilon^{1/2}\ln\left(\frac{1}{\varepsilon}\right)}\right)^{N+1}\right)\quad \text{if } N_1\geqq N\geqq 0, \tag{6.11}$$

as one may verify with a straightforward calculation. Since $|u-Z_N|_{\sup}\leqq |u-Z_{N_1}|_{\sup}+|Z_N-Z_{N_1}|_{\sup}$, the following result is found from (6.10)-(6.11).

THEOREM 2. *In the sup norm it follows that the constructed global approximation Z_N differs from the solution at most by an amount*

$$|u-Z_N|_{\sup}=O\left(\sqrt{\varepsilon^{1/2}\ln\left(\frac{1}{\varepsilon}\right)}^{N+1}\right),\qquad N\geqq 0. \tag{6.12}$$

Of course the estimate given in the introduction (1.12) is contained in (6.12) for N sufficiently large. In order to prove (1.13) and (1.14) we need estimates on the derivatives $(\partial/\partial x_i)(u-Z_N)$, $(\partial^2/\partial x_i\partial x_j)(u-Z_N)$. Note that

$$\left[-\varepsilon\Delta+\frac{1}{(1+u)^2}\right]\frac{\partial}{\partial x_i}(u-Z_N)=r_{i,N}^{(1)} \tag{6.13}$$

with

$$r_{i,N}^{(1)}=\frac{\partial r}{\partial x_i}+\frac{\partial Z_N}{\partial x_i}\left\{\frac{1}{(1+Z_N)^2}-\frac{1}{(1+u)^2}\right\},$$

i.e.,

$$\|r_{i,N}^{(1)}|_{\sup}\leqq R_{i,N}^{(1)}\varepsilon^{-1}\left(\sqrt{\varepsilon^{1/2}\ln\left(\frac{1}{\varepsilon}\right)}\right)^{N+1}.$$

Using a technique shown in Lemma 12 we see that

$$\left|\frac{\partial}{\partial x_i}(u-Z_N)\right|_{\sup}=O(\varepsilon^{-1}|r_{i,N}^{(1)}|_{\sup}). \tag{6.14}$$

Now for N sufficiently large (1.13), the explicit expression for Z_N and its derivatives is contained in (6.14), (6.12). An interesting observation is that the main contribution in the estimate (1.13) comes from the layer along $\partial\Omega$. Differentiating (6.13) w.r.t. x_j and proceeding in an analogous way we obtain

$$\left|\frac{\partial^2}{\partial x_i\partial x_j}(u-Z_N)\right|_{\sup}=O(\varepsilon)\quad \text{for } N \text{ sufficiently large.} \tag{6.15}$$

Thus we are led to (1.14). Again the main contribution in (1.14) comes from the layer along $\partial\Omega$.

7. Discussion of some generalisations. Two generalisations of the problem as specified in (1.1) can be dealt with in a rather straightforward way: (i) higher dimensional domains Ω and (ii) general second order, negative, elliptic operators A with smooth coefficients, instead of the Laplacian Δ. The conditions in the Introduction have a formulation, which already applies to this situation. Of course certain minor changes have to be made in the construction. For example, in general one has to work with several local coordinate systems (ρ, θ); to describe a neighbourhood of S and expressions such as (3.2) will be more complicated. However, nothing changes in the structure of the approximation Z_N, and our methods of proving the validity of the approximation based on the maximum principle still work.

A different story arises for a generalisation to more general nonlinearities, i.e., for a problem

$$\begin{aligned} &-\varepsilon Au+\Phi(x, u)=f\geqq 0 \quad \text{on } \Omega, \\ &u=0 \quad \text{on } \partial\Omega, \end{aligned} \tag{7.1}$$

with

$$\Phi\in C^{\infty}(\bar{\Omega}\times\mathbb{R}), \quad \Phi(x, 0)=0, \quad \frac{\partial\Phi}{\partial u}>0, \quad \lim_{u\to\infty} \Phi(x, u)=1.$$

Note that a more general case with $\lim_{u\to\infty} \Phi(x, u)=\omega(x)$ on $\bar{\Omega}$, $\omega\in C^{\infty}(\bar{\Omega})$, $\omega>0$ can be reduced to the one above by dividing by ω. Now the structure of the approximation near S and Ω_+ will depend heavily on how $\Phi(x, u)$ tends to its limit as $u\to\infty$. In cases where Φ has the following asymptotics:

$$\Phi(x, u)\simeq 1-\lambda_1(x)u^{-1}+\sum_{k\geqq 2} \lambda_k(x)u^{-k} \quad \text{for } u\to\infty \tag{7.2}$$

with coefficients $\lambda_k\in C^{\infty}(\bar{\Omega})$, $\lambda_1>0$, the approximation will still have a structure as in this paper and its construction is analogous. In cases where Φ has completely different asymptotics for $u\to\infty$, the structure of the approximation will also be different. For example, for $\Phi=1-e^{-u}$ the structure of the approximation will be easier, since only order functions of the type $\varepsilon^{p/2}$, $p=-2, 0, 1, \cdots$ will then appear (see [10]).

The method of proving validity given in § 6 extends without difficulties to these more general nonlinearities.

REFERENCES

[1] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam, 1979.

[2] M. MICHAELIS AND M. MENTEN, Biochem. Z., 49 (1973).

[3] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non-linéaire*, Dunod, Paris, 1969.

[4] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, Springer, Berlin, 1972.

[5] C. M. BRAUNER AND B. NICOLAENKO, *Singular perturbations and free boundary value problems*, Proc. 4th Internat. Conf. on Computing Methods in Appl. Sci., North-Holland, Amsterdam, 1980.

[6] ———, *Internal layers and free boundary problems*, Proc. Bail 1 Conf., Boole Press, Dublin, 1980.

[7] ———, *A general approximation of some free boundary problems by bounded penalization*, Proc. Sem. Coll. de France, Pitman, to appear.

[8] A. FRIEDMAN, *Variation Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[9] L. S. FRANK AND W. D. WENDT, *Solutions asymptotiques pour une classe de perturbations singulières elliptiques semilinéaires*, C.R. Acad. Sci. Paris, Sér. I Math., 295.

[10] M. GARBEY, *Sur l'étude d'une classe de problèmes de perturbation singulière elliptiques gouvernés par une inéquation variationelle par la methode des développements asymptotiques raccordés*, Ecole Centrale de Lyon, thesis, 1984.

[11] C. M. BRAUNER, W. ECKHAUS, M. GARBEY AND A. VAN HARTEN, *On the transition layer along a free boundary*, Proc. Bail 3 Conf., Boole Press, Dublin, 1984.

[12] A. VAN HARTEN, *Non-linear singular perturbation problems*, J. Math. Anal. Appl., 65 (1978), pp. 126–168.

[13] J. T. SCHWARTZ, *Non-linear Functional Analysis*, Gordon and Breach, New York, 1969.

[14] D. GILBARG AND N. S. TRUDINGER, *Elliptic* P.D.E. *of Second Order*, Springer, Berlin, 1977.

[15] O. A. LADYŽENSKAJA AND N. N. URALTSEVA, *Equations aux derivées partiells de type elliptique*, Dunod, Paris, 1968.

[16] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Interscience, New York, 1962.

[17] A. VAN HARTEN, *Singularly perturbed non-linear 2nd order elliptic boundary value problems*, Ph.D. thesis, Univ. of Utrecht, 1975.

[18] W. ECKHAUS AND E. M. DE JAGER, *Asymptotic solutions of singular perturbation problems for linear differential equations of elliptic type*, Arch. Rational Mech. Anal. (1966).

[19] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[20] H. LEWY AND G. STAMPACCHIA, *On the regularity of the solution of a variation inequality*, Comm. Pure Appl. Math., 22 (1969).

[21] L. A. CAFFARELLI AND N. H. RIVIERE, *Smoothness and analyticity of free boundaries in variational inequalities*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 3 (1976).

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

AN ABSTRACT D'ALEMBERT FORMULA*

JEROME A. GOLDSTEIN† AND JAMES T. SANDEFUR, JR.‡

Abstract. The abstract d'Alembert formula expresses solutions of the factored linear equation $\prod_{j=1}^{n}(d/dt-A_j)u(t)=0$ as $u(t)=\sum_{j=1}^{n}u_j(t)$, where $(d/dt-A_j)u_j(t)=0$ for $j=1,\cdots,n$. Here each A_j generates a (C_0) semigroup on a Banach space X and suitable hypotheses hold. Applications are made to asymptotic theory, specifically to equipartition of energy and to scattering theory.

Key words. d'Alembert formula, wave equations, scattering theory, equipartition of energy, semigroups of operators

AMS(MOS) subject classifications. 47D05, 34G10, 35P25, 35L30

1. Introduction. Let $A_1,\cdots,A_n$ generate mutually commuting strongly continuous semigroups of bounded linear operators on a Banach space X. If A_i-A_j is invertible and has a big enough range for $i\neq j$, then any solution u of the factored nth order equation

$$\prod_{j=1}^{n}\left(\frac{d}{dt}-A_j\right)u(t)=0$$

decomposes as a sum $u(t)=\sum_{j=1}^{n}u_j(t)$ of solutions of $(d/dt-A_j)u_j(t)=0$. The motivating example for this decomposition is the classical d'Alembert formula for the one-(space-) dimensional wave equation, which expresses the solution of

$$\frac{\partial^2 u}{\partial t^2}-c^2\frac{\partial^2 u}{\partial x^2}=0 \qquad \left(=\prod_{j=1}^{2}\left(\frac{d}{dt}-(-1)^j cd/dx\right)u\right)$$

as $u(t,x)=F(x+ct)+G(x-ct)$.

The abstract d'Alembert formula is presented in § 2 and examples are given in § 3. In §§ 4, 5 and 6, we use the abstract d'Alembert formula to unify, extend and simplify equipartition of energy results. Specifically, in § 4, we improve a result of Goldstein and Rosencrans [3] by deriving equipartition of energy for the telegraph equation

$$u_{tt}+2bu_t+u_{xx}=0$$

in the sense that $\|u_t(t)\|^2/\|u_x(t)\|^2\to 1$ as $t\to\infty$. In § 5, we show for the "damped" wave equation

$$u_{tt}+2bu_{tx}+u_{xx}=0,$$

that $\lim_{t\to\infty}\|u_t(t)\|^2/\|u_x(t)\|^2$ exists, that this limit, L, satisfies

$$[(b^2+1)^{1/2}-|b|]^2\leqq L\leqq[(b^2+1)^{1/2}+|b|]^2,$$

and that L depends on the initial data. In § 6 we give equipartition of energy results for equations of higher order (in time), which unify the results of Mochizuki [11] on factored wave equations and of Goldstein and Sandefur [6] on equations of order 2^n. Applications to scattering theory are given in § 7.

* Received by the editors December 11, 1985; accepted for publication April 1, 1986. This research was partially supported by a grant from the National Science Foundation.

† Department of Mathematics, Tulane University, New Orleans, Louisiana 70118.

‡ Department of Mathematics, Georgetown University, Washington, District of Columbia 20057.

2. The abstract d'Alembert formula. Let X be a Banach space. The following assumption will be made throughout this section.

Hypothesis (H1). For $j=1,\cdots,n$, A_j is the infinitesimal generator of a (C_0) semigroup $T_j=\{T_j(t)\colon t\geqq 0\}$ on X. These semigroups commute, that is,

$$T_j(t)T_k(s)=T_k(s)T_j(t)$$

holds for $j,k\in\{1,\cdots,n\}$ and all $t,s\geqq 0$.

Of concern is the nth order differential equation

$$\prod_{j=1}^{n}\left(\frac{d}{dt}-A_j\right)u(t)=0 \qquad (t\geqq 0), \tag{1}$$

to be solved for $u\colon[0,\infty)\to X$. Special cases of (1) arose naturally in our study of equipartition of energy; cf. [6]. In fact, a very special case of the abstract d'Alembert formula is buried in [6], and it was our desire to understand better the calculation in [6] that provided the initial motivation for the present paper.

Our first theorem establishes the well-posdeness of the Cauchy problem for (1). Theorems 2 and 3 give a clean statement of the d'Alembert formula. Theorem 4 gives a messier but a more useful version of the formula.

Following Sandefur [12], we say that the initial value problem for (1) is *well posed* on $\mathbb{R}^+=[0,\infty)$ if

(i) For every set of initial data $\{\phi_1,\cdots,\phi_n\}$ with $\phi_i\in D$, where D is some dense set in X, there exists a unique solution u of (1) such that $\prod_{j=1}^{k}(d/dt-A_j)u\in C^{n-k}(\mathbb{R}^+,X)$ for $k=0,1,\cdots,n-1$ and $u^{(k-1)}(0)=\phi_k$ for $k=1,\cdots,n$;

(ii) If $\{u_m\}$ is a sequence of solutions of (1) (in the sense of (i) above) and if $\prod_{j=1}^{k}(d/dt-A_j)u_m(t)|_{t=0}\to 0$ as $m\to\infty$ for $k=0,1,\cdots,n-1$, then $\prod_{j=1}^{k}(d/dt-A_j)u_m(t)\to 0$ as $m\to\infty$ for $k=0,1,\cdots,n-1$ and uniformly for t in bounded subintervals of $\mathbb{R}^+$.

THEOREM 1. *Assume Hypothesis* (H1). *Then the initial value problem for* (1) *is well posed.*

For the proof we need an extension of a classical ancient result of Gel'fand [1].

LEMMA. *Let Hypothesis* (H1) *hold. Then* $\bigcap_{j=1}^{n}C^\infty(A_j)$ *is dense in* X.

Here $C^\infty(A_j)=\bigcap_{m=1}^{\infty}\operatorname{Dom}(A_j^m)$. For $n=1$ this result is due to Gel'fand; see [7, p. 308]. We shall prove it for $n=2$, assuming it for $n=1$. Two remarks must be made in connection with this: (i) the $n=1$ case actually follows from the proof given below if we specialize A_2 to be zero, so that $T_2(t)\equiv I$; (ii) the proof of the general case is by induction and is in fact the following proof except for some inessential modifications. Thus the details of the general case may safely be omitted.

Let $f\in C^\infty(A_2)$, which by hypothesis is dense in X. Let $\phi\in C_c^\infty((0,\infty),\mathbb{R})$, the subscript c denoting "compactly supported," and set

$$g=\int_0^\infty \phi(t)T_1(t)T_2(t)f\,dt.$$

We *claim* that

$$g\in\operatorname{Dom}(A_1^m)\cap\operatorname{Dom}(A_2^m)$$

for all $m\geqq 1$. It then follows that $g\in\bigcap_{j=1}^{2}C^\infty(A_j)$. Since $g\to f$ as $\phi\to\delta_0$ (for example, take $\phi_\varepsilon\geqq 0$ supported in $[\varepsilon,2\varepsilon]$ with $\int_0^\infty\phi_\varepsilon(t)\,dt=1$ and let $\varepsilon\to 0^+$), we conclude that the lemma follows from the *claim.*

To prove the *claim*, let ϕ be supported in $[\varepsilon, 1/\varepsilon]$ and let $0<h<\varepsilon$. Then

$$T_1(h)g=\int_0^\infty \phi(t)T_1(t+h)T_2(t)f\,dt=\int_0^\infty \phi(t-h)T_1(t)T_2(t-h)f\,dt.$$

Consequently

$$\begin{aligned} h^{-1}[T_1(h)g-g]&=\int_0^\infty h^{-1}[\phi(t-h)-\phi(t)]T_1(t)T_2(t-h)f\,dt\\ &\quad+\int_0^\infty \phi(t)T_1(t)h^{-1}[T_2(t-h)f-T_2(t)]f\,dt \qquad (2)\\ &\to-\int_0^\infty \phi'(t)T_1(t)T_2(t)f\,dt-\int_0^\infty \phi(t)T_1(t)T_2(t)A_2f\,dt.\end{aligned}$$

Hence $g\in \operatorname{Dom}(A_1)$ and A_1g is given by (2). By induction the above argument shows that $g\in\operatorname{Dom}(A_1^m)$ and

$$A_1^mg=\sum_{k=0}^m(-1)^m\binom{m}{k}\int_0^\infty \phi^{(k)}(t)T_1(t)T_2(t)A_2^{m-k}f\,dt;$$

here $\binom{m}{k}=m!/(k!(m-k)!)$ is the usual binomial coefficient. Next, by the commutativity hypothesis (H1), $f\in\operatorname{Dom}(A_2)$ implies $T_1(t)f\in\operatorname{Dom}(A_2)$ and

$$A_2T_1(t)f=T_1(t)A_2f,$$

and similarly for A_2^m in place of A_2. Consequently, since A_2^m is closed and $f\in\operatorname{Dom}(A_2^m)$,

$$\int_0^\infty \phi(t)T_1(t)T_2(t)f\,dt\in\operatorname{Dom}(A_2^m)$$

and

$$A_2^m\left(\int_0^\infty \phi(t)T_1(t)T_2(t)f\,dt\right)=\int_0^\infty \phi(t)T_1(t)T_2(t)A_2^mf\,dt.$$

Since m is arbitrary, the proof is complete. □

Proof of Theorem 1. Let D be any dense subspace of $\bigcap_{j=1}^n C^\infty(A_j)$ that is left invariant by each A_j. For instance, $D=\bigcap_{j=1}^n C^\infty(A_j)$ will do. Theorem 1 now follows directly from the lemma and from a result of Sandefur [12, p. 731]. □

Consider (1) with $n=2$ and $A_1=A_2=A$, where A generates T. Then all solutions of (1) are of the form $T(t)(\phi+t\psi)$ where ϕ, ψ are in Dom (A). If we want solutions of (1) to be of the form $u(t)=\sum_{i=1}^n T_i(t)\psi_i$, then to avoid terms such as $tT_i(t)$, it is necessary that A_i-A_j be injective for $i\neq j$.

Hypothesis (H2). $0\in\rho(A_i-A_j)$ for $i\neq j$. More precisely, zero is in the resolvent set of the closure of A_i-A_j for $i\neq j$.

Write $u\in\mathcal{N}(\prod_{j=1}^k(d/dt-A_j))$ to mean that $\prod_{j=1}^k(d/dt-A_j)u(t)=0$ for $t\geqq 0$. Here $\mathcal{N}$ stands for "null space."

THEOREM 2. *Assume Hypotheses* (H1) *and* (H2). *Then* $w\in\mathcal{N}(\prod_{j=1}^n(d/dt-A_j))$ *implies* $w=\sum_{j=1}^n w_j$ *where* $w_j\in\mathcal{N}(d/dt-A_j)$.

Proof. The proof is by induction on n. (Previously we have assumed that the j's increased from right to left. For simplicity of notation we assume in this proof that the j's decrease from right to left.)

The case $n=1$ is trivially true. Assume the theorem to be true for $n-1$ with $n\geqq 2$. Let

$$\prod_{j=1}^{n}\left(\frac{d}{dt}-A_j\right)w=0$$

and define $u=(d/dt-A_n)w$. Then

$$\prod_{j=1}^{n-1}\left(\frac{d}{dt}-A_j\right)u=0.$$

By the induction hypothesis, $u=\sum_{j=1}^{n-1}u_j$ where $u_j\in\mathcal{N}(d/dt-A_j)$ for $j=1,\cdots,n-1$. Next, $w'-A_nw=u$ implies, by the variation of parameters formula (alias Duhamel's formula),

$$\begin{aligned} w(t)&=T_n(t)w(0)+\int_0^t T_n(t-s)u(s)\,ds\\ &=T_n(t)w(0)+\sum_{j=1}^{n-1}\int_0^t T_n(t-s)T_j(s)u_j(0)\,ds \qquad \text{since } u_j\in\mathcal{N}\left(\frac{d}{dt}-A_j\right)\\ &=T_n(t)w(0)+\sum_{j=1}^{n-1}\int_0^t \frac{d}{ds}[T_n(t-s)T_j(s)(A_j-A_n)^{-1}u_j(0)]\,ds\\ &\qquad\qquad \text{by Hypothesis (H2) and since } 0\in\rho(A_j-A_n)\\ &=T_n(t)\left\{w(0)-\sum_{k=1}^{n-1}(A_k-A_n)^{-1}u_k(0)\right\}\\ &\qquad+\sum_{j=1}^{n-1}T_j(t)\{(A_j-A_n)^{-1}u_j(0)\}\equiv\sum_{k=1}^{n}w_j(t), \end{aligned}$$

where $w_j\in\mathcal{N}(d/dt-A_j)$ for $j=1,\cdots,n$. □

The next theorem is a variant of Theorem 2 to the context of generalized solutions. $u=F(x+t)$ (for $x,t\in\mathbb{R}$) satisfies $u_{tt}=u_{xx}$ in some sense, even if F is not a C^2 function. This is the notion we want to generalize to an abstract context.

According to [12], (1) can be reduced to the system $dw(t)=Aw(t)$ where

$$A=\begin{bmatrix} A_1 & I & \cdots & 0\\ 0 & A_2 & & 0\\ \vdots & & \ddots & \vdots\\ & & & I\\ 0 & & & A_n \end{bmatrix}$$

acts on the product space X^n. Moreover, u is a solution of (1) if and only if u is the first component of a solution w of $w'=Aw$.

Now let A_0 generate a (C_0) semigroup T_0 on a Banach space Y. We call $T_0(\cdot)f$ a *mild solution* of $dv/dt=A_0v$ for every $f\in Y$. A mild solution need not be differentiable, but it is uniquely determined by its initial condition. By a *mild solution* of (1) we mean, using the notation of the above paragraph, the first component of a mild solution w of the first order matrix equation corresponding to (1).

Write $u\in\bar{\mathcal{N}}(\prod_{j=1}^{k}(d/dt-A_j))$ to mean that u is a mild solution of $\prod_{j=1}^{k}(d/dt-A_j)u(t)=0$ for $t\geqq 0$.

THEOREM 3. *Assume Hypotheses* (H1) *and* (H2). *Then* $w \in \bar{\mathcal{N}}(\prod_{j=1}^{n}(d/dt - A_j))$ *implies* $w = \sum_{j=1}^{n} w_j$, *where* $w_j \in \bar{\mathcal{N}}(d/dt - A_j)$.

Proof. Using the Phillips perturbation theorem, we can show (cf. [12]) that u is a mild solution to (1) if and only if

$$u(t) = T_1(t)\psi_1 + \sum_{m=1}^{n-1} \int_0^t \int_0^{t_m} \cdots \int_0^{t_2} T_1(t-t_m)T_2(t_m - t_{m-1}) \cdots T_m(t_2 - t_1)T_{m+1}(t_1)\psi_{m+1}\, dt_1 \cdots dt_m$$

where

$$\psi_m = \phi_m + \sum_{k=1}^{m-1} (-1)^k \sum_{1 \leq i_1 < \cdots < i_k \leq m-1} (A_{i_k} \cdots A_{i_1}\phi_{m-k})$$

and $\phi_m = u^{(m)}(0)$. In particular, when $n = 2$,

$$u(t) = T_1(t)\phi_1 + \int_0^t T_1(t-s)T_2(s)(\phi_2 - A_1\phi_1)\, ds.$$

The proof of the necessity is by induction. Assume the result true for $n-1$ and let u be given by the above formula with suitable $\psi_1, \cdots, \psi_n$. Note that

$$u(t) = u_0(t) + \int_0^t \int_0^{t_{n-1}} \cdots \int_0^{t_2} T_1(t - t_{n-1}) \cdots T_n(t_1)\psi_n\, dt_1 \cdots dt_{n-1}$$

where u_0 is a mild solution of an equation of form (1) but of order $n-1$. By the induction hypothesis, $u_0(t) = \sum_{j=1}^{n-1} v_j(t)$, where $v_j \in \bar{\mathcal{N}}(d/dt - A_j)$. By the argument of the proof of Theorem 2,

$$\int_0^{t_2} T_{n-1}(t_2 - t_1)T_n(t_1)\psi_n\, dt_1 = T_n(t_2)[(A_n - A_{n-1})^{-1}\psi_n] + T_{n-1}(t_2)[(A_{n-1} - A_n)^{-1}\psi_n],$$

$$\int_0^{t_3}\int_0^{t_2} T_{n-2}(t_3 - t_2)T_{n-1}(t_2 - t_1)T_n(t_1)\psi_n\, dt_1\, dt_2 = \sum_{j=n-2}^{n} T_j(t_3)\chi_j,$$

and so on (by induction), whence

$$u(t) = u_0(t) + \sum_{j=1}^{n} T_j(t)\alpha_j = \sum_{j=1}^{n} T_j(t)\beta_j$$

and thus $u = \sum_{i=1}^{n} u_i$ with $u_i \in \bar{\mathcal{N}}(d/dt - A_i)$. □

At first glance it seems "obvious" that the converse of Theorem 3 holds. It certainly seems justified calling $u(t) = \sum_{i=1}^{n} T_i(t)\beta_i$ a *weak solution* of (1), but the point is that this u is not necessarily a mild solution of (1). Let $n = 2$ and

$$\begin{aligned} u(t) &= T_1(t)[(A_1 - A_2)^{-1}\phi_2 - A_2(A_1 - A_2)^{-1}\phi_1] \\ &\quad + T_2(t)[(A_2 - A_1)^{-1}\phi_2 - A_1(A_2 - A_1)^{-1}\phi_1] \\ &\equiv T_1(t)\beta_1 + T_2(t)\beta_2. \end{aligned}$$

This u is a "weak solution" in the above sense, but u is in general not a mild solution of

$$\left(\frac{d}{dt} - A_1\right)\left(\frac{d}{dt} - A_2\right)u = 0$$

unless $\phi_1 \in \operatorname{Dom}(A_1)$. Similarly, u is a mild solution of

$$\left(\frac{d}{dt} - A_2\right)\left(\frac{d}{dt} - A_1\right)u = 0$$

if $\phi_1 \in \operatorname{Dom}(A_2)$. In particular, despite the commutativity, the order of the factors in (1) matters as far as the notion of a mild solution is concerned.

It is also of interest to note that u can formally be a solution to the Cauchy problem even if $\phi_1 \notin \operatorname{Dom}(A_1)$ and $\phi_2 \notin \operatorname{Dom}(A_2)$; but in this case u need not be a mild solution.

THEOREM 4. *Let Hypothesis* (H1) *hold. Let D be a dense subspace of $\bigcap_{j=1}^{n} C^\infty(A_j)$ such that the problem*

$$\prod_{j=1}^{n}\left(\frac{d}{dt}-A_j\right)w(t)=0 \qquad (t \geqq 0),$$

$$w^{(k)}(0)=f_k, \qquad k=0,\ 1,\cdots,n-1$$

has a unique solution whenever $f_0, f_1, \cdots, f_{n-1}$ are given in D and are such that $A_1^{i_1}\cdots A_n^{i_n} f_j \in D$ holds for $0 \leqq j \leqq n-1$ and $i_k \geqq 0$, $\sum_{k=1}^{n} i_k \leqq n-1$. Suppose further that $D \subset \operatorname{Ran}(A_i - A_j)$ for $i \neq j$. Then any $w \in \mathcal{N}(\prod_{j=1}^{n}(d/dt - A_j))$ with initial date in D satisfies $w=\sum_{j=1}^{n} w_j$ where $w_j \in \mathcal{N}(d/dt - A_j)$.

Proof. For $n=1$, this is trivial. For $n=2$, the proof of Theorem 2 gives, for $u=(d/dt-A_2)w$, $w(t)=T_2(t)\{w(0)-(A_1-A_2)^{-1}u(0)\}+T_1(t)\{(A_1-A_2)^{-1}u(0)\}$. By hypothesis,

$$u(0)=w'(0)-A_2 w(0) \in D \subset \operatorname{Ran}(A_1-A_2).$$

Thus, in the case of $n=2$, the proof of Theorem 3 does not really require the condition $0 \in \rho(A_1-A_2)$; only the assumptions of Theorem 4 are needed.

For the case of general n, the induction argument of Theorem 3 gives $w=\sum_{j=1}^{n} w_j$, where $w_j \in \mathcal{N}(d/dt - A_j)$, proved we can show that each of the vectors $(A_j-A_n)^{-1}u_j(0)$ (for $j \leqq n-1$) is in the subspace D. But this follows from the proof of Theorem 3 together with the conditions $A_1^{i_1}\cdots A_n^{i_n} w_j(0) \in D$ for $j \in \{0,\cdots,n-1\}$, $i_k \geqq 0$, and $\sum_{k=1}^{n} i_k \leqq n-1$. The details of the induction proof can be safely omitted. □

3. Four simple examples.

Example 1. Let A generate a (C_0) semigroup on X and let $0<\beta_1<\beta_2\cdots<\beta_n$. Set $A_j=\beta_j A$ for $j=1,\cdots,n$. Then Theorem 4 applies provided

$$R^\infty(A) \cap C^\infty(A)$$

is dense in X, where $R^\infty(A)=\bigcap_{n=1}^{\infty}\operatorname{Ran}(A^n)$.

Example 2. This provides an illustration of Example 1.

In Example 1 take X to be a complex Hilbert space and take $A=iH$ where H is a spectrally absolutely continuous self-adjoint operator on X. (Cf., e.g., Kato [8].) Then, by the spectral theorem (a unitarily equivalent representation of) A acts as multiplication by the identity function on $\bigoplus_{\alpha\in J} L^2(\mathbb{R}, \nu_\alpha(x)\,dx)$, where $0 \leqq \nu_\alpha \in L^1_{\text{loc}}(\mathbb{R})$ for each $\alpha \in J$, i.e.

$$A\left(\bigoplus_\alpha f_\alpha(x_\alpha)\right)=\bigoplus_\alpha x_\alpha f_\alpha(x_\alpha) \qquad (x_\alpha \in \mathbb{R},\ \alpha \in J).$$

The set

$$D=\left\{\bigoplus_\alpha f_\alpha : f_\alpha \in C_c^\infty(\mathbb{R}\backslash\{0\}) \text{ for each } \alpha \in J\right\}$$

is a subspace of $C^\infty(A) \cap R^\infty(A)$ and is dense for X. Thus by Example 1, Theorem 4 applies in this case.

A special case is the wave equation in free space: $u_{tt}=\Delta u$ where Δ acts on $L^2(\mathbb{R}^n)$. When $n=1$, we can show that Theorem 4 applies to the wave equation $u_{tt}=u_{xx}$ in $L^p(\mathbb{R})$, $1\leqq p<\infty$ and in $BUC(\mathbb{R})$, the bounded uniformly continuous functions in the supremum norm.

Example 3. The equation

$$\left(\frac{d^2}{dt^2}-\alpha^2\Delta\right)\left(\frac{d^2}{dt^2}-\beta^2\Delta\right)u(t)=0$$

in $X=L^2(\mathbb{R}^n)$ is satisfied by the shear and pressure waves in linear elasticity in free space. This is a fourth order equation covered by the analysis of Example 2, provided that α and β are distinct positive constants. The equation has the form $\prod_{j=1}^4 (d/dt-ic_jA)u(t)=0$, where $A=(-\Delta)^{1/2}$, $c_1=-c_2=\alpha$, $c_3=-c_4=\beta$.

Example 4. In the set up of Example 3 introduce a nonnegative potential $V\in L^p(\mathbb{R}^n)+L^\infty(\mathbb{R}^n)$, where $p\geqq 2$, $p\geqq n/2$. Then the fourth order equation

$$\left(\frac{d^2}{dt^2}-\alpha(\Delta-V(x))\right)\left(\frac{d^2}{dt^2}-\beta(\Delta-V(x))\right)u(t,x)=0$$

is covered by Theorem 4.

4. The abstract telegraph equation. Consider the dissipative wave equation or abstract telegraph equation

$$\frac{d^2u}{dt^2}+2b\frac{du}{dt}+H^2u=0, \tag{3}$$

where H is a spectrally absolutely continuous self-adjoint operator on a complex Hilbert space X. Define the kinetic and potential energies of a solution u at time t to be

$$K(t)=\|u'(t)\|^2, \qquad P(t)=\|Hu(t)\|^2.$$

When $b=0$, total energy is conserved $(E(t)=K(t)+P(t)=E(0))$ and energy is equipartitioned $(K(t), P(t)\to E(0)/2$ as $t\to\infty)$. When $b>0$, $E(t)$ decays to zero, but nevertheless a weak form of equipartition of energy was established by Goldstein and Rosencrans [3], who showed that $0<\liminf_{t\to\infty} K(t)/P(t)\leqq \limsup_{t\to\infty} K(t)/P(t)<\infty$, provided that $0\in\rho(H)$ and b is sufficiently small. With the aid of the abstract d'Alembert formula this result will be sharpened substantially.

THEOREM 5. *Let H be a spectrally absolutely continuous self-adjoint operator on a complex Hilbert space X and suppose $H^2\geqq\alpha^2I$. Let $0<b<\alpha$. Then every nonzero solution of* (3) *admits sharp equipartition of energy in the sense that*

$$\lim_{t\to\infty} K(t)/P(t)=1.$$

Proof. The operator $B=(H^2-b^2I)^{1/2}$ is specially absolutely continuous, and $C=iB$ generates a (C_0) unitary group which we denote by $\{T(t)=e^{tC}: t\in\mathbb{R}\}$. Equation (3) can be rewritten as

$$\left(\frac{d}{dt}-A_+\right)\left(\frac{d}{dt}-A_-\right)u=0,$$

where

$$A_\pm=-bI\pm C.$$

Thus by Theorem 2, every solution of (3) is of the form

$$u(t)=e^{-bt}(T(t)\phi+T(-t)\psi),$$

where $\phi, \psi \in \operatorname{Dom}(H)=\operatorname{Dom}(C)$. The kinetic and potential energies become

$$K(t)=e^{-2bt}\|A_+T(t)\phi+A_-T(-t)\psi\|^2,$$

$$P(t)=e^{-2bt}\|H(T(t)\phi+T(-t)\psi)\|^2.$$

Consequently, by the law of cosines, the unitarity of $T(\pm t)$ and the Riemann-Lebesgue lemma, $\lim_{t\to\infty} K(t)/P(t)$ exists and equals

$$[\|A_+\phi\|^2+\|A_-\psi\|^2][\|H\phi\|^2+\|H\psi\|^2]^{-1}.$$

We let $x=H\phi$; then $y=H\psi$ yields

$$\lim_{t\to\infty} K(t)/P(t)=[\|A_+H^{-1}x\|^2+\|A_-H^{-1}y\|^2][\|x\|^2+\|y\|^2]^{-1}.$$

But for each $z\in X$,

$$\|A_\pm H^{-1}z\|^2=\langle(-bI\pm C)H^{-1}z, (-bI\pm C)H^{-1}z\rangle$$

$$=\langle(b^2I-C^2)H^{-2}z, z\rangle$$

since $H^*=H$, $C^*=-C$, and $[H, C]=0$

$$=\langle(b^2I-(b^2I-H^2))H^{-2}z, z\rangle=\|z\|^2,$$

whence

$$\lim_{t\to\infty} K(t)/P(t)=1. \qquad \square$$

Remarks. Many cases of equipartition of energy from a finite time onward are known. By this we mean that, for $\{T(t)\colon t\in\mathbb{R}\}$ a (C_0) unitary group and for D a suitable dense subset of $X, x, y\in D$ implies that there exists a $\tau=\tau(x,y)>0$ such that $\operatorname{Re}\langle T(t)x, y\rangle=0$ for $|t|>\tau$. When $\{e^{tC}\}$ satisfies this condition we get

$$K(t)=P(t)=\tfrac{1}{2}e^{-2bt}$$

for $t>\tau=\tau_1(\phi,\psi)$ for suitable data ϕ and ψ.

Let the notation and assumptions of Theorem 5 hold. The function $v(t)=e^{bt}u(t)$ satisfies

$$v''+(H^2-b^2I)v=0.$$

If we let $K_v(t)=\|v'(t)\|^2$, $P_v(t)=\|(H^2-b^2I)^{1/2}v(t)\|^2$, then $K_v(t)/P_v(t)\to 1$ as $t\to\infty$ for each v corresponding to nonzero initial data. By combining calculations involving K_v, P_v with those involving K and P, we can conclude that (for H absolutely continuous and $H\geqq\alpha^2I$, $\alpha>b>0$)

$$\lim_{t\to\infty}\frac{\operatorname{Re}\langle u'(t), u(t)\rangle+b\|u(t)\|^2}{\|Hu(t)\|^2}=0.$$

In particular,

$$\limsup_{t\to\infty}\frac{\operatorname{Re}\langle u'(t), u(t)\rangle}{\|u'(t)\|^2}\leqq 0$$

for all nonzero solutions u of (3).

5. A "damped" wave equation. Let H be a spectrally absolutely continuous self-adjoint operator on a Hilbert space X. Of concern is the "damped" wave equation

$$\frac{d^2u}{dt^2}+2ibH\frac{du}{dt}+H^2u=0, \tag{4}$$

which differs from (3) by the presence of iH in the "friction" term. The $\sqrt{-1}$ factor makes the first order term simply a perturbation and not really a friction term. We let b be any real number and we make as initial data

$$u(0)=u_0\in \mathrm{Dom}\,(H), \qquad u'(0)=u_1\in X. \tag{5}$$

Let $c=(b^2+1)^{1/2}$.

THEOREM 6. *The limit* $L=\lim_{t\to\infty} K(t)/P(t)$ *exists and satisfies*

$$(c-|b|)^2\leqq L\leqq(c+|b|)^2.$$

Furthermore $L\to 1$ *as* $b\to 0$, *and the bounds for* L *are best possible.*

Proof. Let $A_\pm=(-b\pm c)iH$ and let $\{T_\pm(t)=\exp(tA_\pm): t\in\mathbb{R}\}$ be the corresponding (C_0) unitary groups. Then (4) factors as

$$\left(\frac{d}{dt}-A_+\right)\left(\frac{d}{dt}-A_-\right)u=0.$$

By Example 2, the solution to (4), (5) is

$$u(t)=T_+(t)\phi_++T_-(t)\phi_-,$$

where $\phi_\pm\in\mathrm{Dom}(H)$ and

$$H\phi_\pm=2^{-1}[Hu_0\pm c^{-1}(-iu_1+bHu_0)].$$

The kinetic and potential energies are

$$K(t)=\|u'(t)\|^2=\|(-b+c)T_+(t)H\phi_++(-b-c)T_-(t)H\phi_-\|^2,$$
$$P(t)=\|Hu(t)\|^2=\|T_+(t)H\phi_++T_-(t)H\phi_-\|^2.$$

As usual we expand by the law of cosines and employ the unitarity of $T_\pm(t)$ and the Riemann-Lebesgue lemma. The conclusion is that the limit L of $K(t)/P(t)$ as $t\to\infty$ exists and equals

$$L=[(c-b)^2\|H\phi_+\|^2+(b+c)^2\|H\phi_-\|^2][\|H\phi_+\|^2+\|H\phi_-\|^2]^{-1}.$$

Letting $x=\|H\phi_+\|^2$ and $y=\|H\phi_-\|^2$ it follows that L is constant on the lines $y=ax$ in the first quadrant, $0\leqq a\leqq\infty$. (Here $a=\infty$ refers to the line $x=0$.) Think of $L=L(a)$ as a function of $a\in[0,\infty]$. Then $L(a)=[(c-b)^2+(b+c)^2a^2][1+a^2]^{-1}$, $L(0)=(c-b)^2$, $L(\infty)=(b+c)^2$. Since $dL/da=8abc[1+a^2]^{-2}$ it follows that L is increasing in a (resp. decreasing in a) for $b>0$ (resp. $b<0$). Consequently L attains its maximum and minimum on the set $\{0,\infty\}$. That is,

$$(c-|b|)^2\leqq L\leqq(c+|b|)^2. \tag{6}$$

Note that $(c\pm|b|)^2\to 1$ as $b\to 0$. Taking $u_1=-iH(c+b)u_0$ we get $H\phi_+=0$ and $L=(b+c)^2$, while taking $u_1=i(c-b)Hu_0$ gives $H\phi_-=0$ and $L=(b-c)^2$. Thus (6) is sharp. □

The results of this section and the previous section can be unified by considering the equation

$$u''+2Bu'+H^2u=0,$$

where H is a spectrally absolutely continuous self-adjoint operator on a Hilbert space X and $B=b_1I+ib_2H$ where b_1, b_2 are real constants. We omit the details.

The change of variables $v(t)=e^{ibtH}u(t)$ converts (4) into

$$\frac{d^2v}{dt^2}+(1+b^2)H^2v=0$$

which admits equipartition of energy using different notions of potential and kinetic energies than those used in Theorem 6.

6. Equipartition of energy for *n*th order equations. In [6] we characterized equipartition of energy for a large class of abstract Cauchy problems of order 2^m. The theorem and methods of [6] appeared to subsume all of the literature on equipartition of energy for higher order abstract Cauchy problems that we knew of except for the interesting result of Mochizuki [11], who treated a special equation of order $2m$ and got sharp results. In this section we show how Mochizuki's result and more general results for arbitrary order equations follow from Theorem 4.

We then compare his results with ours for fourth order equations of which

$$\left(\frac{d^2}{dt^2}-\alpha^2\Delta\right)\left(\frac{d^2}{dt^2}-\beta^2\Delta\right)u=0,$$

where $\alpha\neq\beta$, is an example. This particular equation arises in the study of linear elasticity. (See Example 3 in § 3.)

Let H be a spectrally absolutely continuous self-adjoint operator acting on a complex Hilbert space X. Consider the nth order equation

$$\prod_{j=1}^{n}\left(\frac{d}{dt}-i\beta_jH\right)u(t)=0 \quad (t\in\mathbb{R}), \tag{7}$$

where $\{\beta_1,\cdots,\beta_n\}$ is a set of n distinct real numbers. Let $T_j=\{T_j(t)=\exp(it\beta_jH)\colon t\in\mathbb{R}\}$ be the (C_0) unitary group generated by $A_j=+i\beta_jH$. Then by Example 2, the initial value problem for (7) is well posed and every solution (in a dense set of solutions) of (7) has the form

$$u(t)=\sum_{j=1}^{n}T_j(t)\phi_j.$$

Assume $\phi_j\in\operatorname{Dom}(H^{n-1})$ for $j=1,\cdots,n$. Then

$$\lim_{t\to\pm\infty}\|H^{n-k-1}u^{(k)}(t)\|^2=\lim_{t\to\pm\infty}\left\|\sum_{j=1}^{n}(+i\beta_j)^kH^{n-1}T_j(t)\phi_j\right\|^2$$

$$=\sum_{j=1}^{n}\beta_j^{2k}\|H^{n-1}\phi_k\|^2$$

by the law of cosines and the Riemann-Lebesgue lemma. (Cf. the argument in [2] or [5].) If we define the jth partial energy to be

$$E_j(t)=\|H^{n-j-1}u^{(j)}(t)\|^2$$

and the total energy to be

$$E(t)=\sum_{j=0}^{n-1}E_j(t),$$

then $E_j(t)$ and $E(t)$ have limits as $t\to\pm\infty$, but $E(t)$ is not constant in general. (E is constant if $n=2$ and $\beta_2=-\beta_1$; this is the classical abstract wave equation.)

Now consider the abstract $2m$th order wave equation

$$\prod_{k=1}^{m}\left(\frac{d^2}{dt^2}+\beta_k^2H^2\right)u(t)=0 \qquad (t\in\mathbb{R}) \tag{8}$$

treated by Mochizuki [11]. Following Mochizuki we assume H is a spectrally absolutely continuous self-adjoint operator on X and that $0<\beta_1<\beta_2<\cdots<\beta_n$. Letting T_j be as above, we have (using Example 2) that every solution of (8) is of the form

$$u(t)=\sum_{j=1}^{m}(T_j(t)\phi_j+T_j(-t)\psi_j), \tag{9}$$

where $\phi_1,\cdots,\phi_m,\psi_1,\cdots,\psi_m\in\operatorname{Dom}(H^{2m-1})$. Define the $(k+1)$st partial energy of u to be

$$\begin{aligned}E_{k+1}(t)&=\|H^{2m-1-k}u^{(k)}(t)\|^2\\&=\left\|\sum_{j=1}^{m}(i\beta_j)^kH^{2m-1}(T_j(t)\phi_j+(-1)^kT_j(-t)\psi_j)\right\|^2\end{aligned}$$

for $k=0,1,\cdots,2m-1$. Then we deduce Mochizuki's asymptotic result, namely

$$\lim_{t\to\pm\infty}E_{k+1}(t)=\sum_{j=1}^{m}\beta_j^{2k}(\|H^{2m-1}\phi_j\|^2+\|H^{2m-1}\psi_j\|^2)$$

for $k=0,\cdots,2m-1$.

Remark 1. Finding ϕ_j and ψ_k in terms of the initial data is relatively easy linear algebra. In the case of (9) with $m=2$ we have, setting $u_j=u^{(j)}(0)$,

$$u_0=(\phi_1+\psi_1)+(\phi_2+\psi_2),\qquad u_1=iH[\beta_1(\phi_1-\psi_1)+\beta_2(\phi_2-\psi_2)],$$

$$u_2=-H^2[\beta_1^2(\phi_1-\psi_1)+\beta_2^2(\phi_2-\psi_2)],\qquad u_3=-iH^3[\beta_1^3(\phi_1-\psi_1)+\beta_2^3(\phi_2-\psi_2)].$$

Consequently

$$\phi_1=\chi_{1+},\quad \psi_1=\chi_{1-},\quad \phi_2=\chi_{2+},\quad \psi_2=\chi_{2-}$$

where

$$\chi_{1\pm}=\frac{[\beta_1x\pm y]}{[2\beta_1(\beta_1^2-\beta_2^2)]},\qquad \chi_{2\pm}=\frac{[\beta_2x\pm y]}{[2\beta_2(\beta_1^2-\beta_2^2)]},$$

$$x=u_0+H^{-2}u_2,\qquad y=-iH^{-1}u_1-iH^{-3}u_3.$$

By using the parallelogram law, we get the convergence of the above energies in terms of the initial data as

$$E_1(t)\to[2(\beta_2^2-\beta_1^2)]^{-1}\{(\beta_1^{-2}+\beta_2^{-2})\|H^3y\|^2+2\|H^3x\|^2\},$$

$$E_2(t)\to[2(\beta_2^2-\beta_1^2)]^{-1}\{2\|H^3y\|^2+(\beta_1^2+\beta_2^2)\|H^3x\|^2\},$$

$$E_3(t)\to[2(\beta_2^2-\beta_1^2)]^{-1}\{(\beta_1^2+\beta_2^2)\|H^3y\|^2+(\beta_1^4+\beta_2^4)\|H^3x\|^2\},$$

$$E_4(t)\to[2(\beta_2^2-\beta_1^2)]^{-1}\{(\beta_1^4+\beta_2^4)\|H^3y\|^2+(\beta_1^6+\beta_2^6)\|H^3x\|^2\}.$$

Remark 2. Let u be given by (9). Then $u(t)=\sum_{k=1}^{m}u_k(t)$, where u_k satisfies

$$\left(\frac{d^2}{dt^2}+\beta_k^2H^2\right)u_k=0.$$

Thus u is partitioned into m additive components, $u_k, k=1,\cdots,m$. We can define kinetic and potential energies as $K_k(t)=\|u_k'(t)\|^2$ and $P_k(t)=\|\beta_k H u_k(t)\|^2$; and then define partial energies as $\tilde{E}_k(t)\equiv K_k(t)+P_k(t)$. Let $\tilde{E}(t)=\sum_{k=1}^m \tilde{E}_k(t)$. Observe that we not only have conservation of energy, but that we have conservation of each partial energy since $\|u_k'(t)\|^2+\|\beta_k H u_k(t)\|^2\equiv \tilde{E}_k(t)=\tilde{E}_k(0)$. In addition, we have equipartition of the partial energies, that is,

$$\lim_{t\to\pm\infty} K_k(t)=\lim_{t\to\pm\infty} P_k(t)=2^{-1}\tilde{E}_k(0).$$

When $m=2$, an example of which is the equation of linear elasticity, this result implies that the solutions u is the sum of two waves (the shear wave and the pressure wave in the example), each of which conserves energy and is equipartitioned between kinetic energy and potential energy as $t\to\pm\infty$.

In Goldstein and Sandefur [6] we showed how (7) could be reduced to a first order system on X^n when $n=2^m$. This first order system is controlled by a unitary semigroup $\mathcal{T}(t)$. In particular, in the case of (7) with $m=2$, this first order system is

$$U'(t)=\mathcal{A}U(t), \tag{10}$$

where

$$\mathcal{A}=\begin{pmatrix} 0 & \alpha_1 & 0 & \alpha_2 \\ \alpha_1 & 0 & \alpha_2 & 0 \\ 0 & \alpha_2 & 0 & \alpha_1 \\ \alpha_2 & 0 & \alpha_1 & 0 \end{pmatrix} iH, \qquad U(t)=\begin{pmatrix} w_1(t) \\ w_2(t) \\ w_3(t) \\ w_4(t) \end{pmatrix},$$

$\alpha_1=2^{-1}(\beta_1+\beta_2)$ and $\alpha_2=2^{-1}(\beta_1-\beta_2)$.

By defining the energies as $\mathcal{E}(t)=\sum_{j=1}^n \|w_j(t)\|^2$, we showed that the total energy is conserved and that energy is equipartitioned in the sense that $\lim_{t\to\pm\infty}\|w_j(t)\|^2=2^{-n}\mathcal{E}(0)$, for $j=1,\cdots,n$.

In the case of (10), a straightforward calculation shows that

$$U(t)=\begin{pmatrix} w_1(t) \\ w_2(t) \\ w_3(t) \\ w_4(t) \end{pmatrix}=\begin{pmatrix} \alpha_2 iH[u''(t)-\beta_1\beta_2H^2u] \\ \alpha_1 iH[u''(t)-\beta_1\beta_2H^2u] \\ -2\alpha_1\alpha_2H^2u'(t) \\ u'''(t)+2^{-1}(\beta_1^2+\beta_2^2)H^2u'(t) \end{pmatrix} \tag{11}$$

is a solution to (10). By our d'Alembert formula, we know that $u=u_1+u_2$, where u_j satisfies $u_j''+\beta_j^2H^2u_j=0$, $j=1,2$. Making this substitution into (11) we get

$$U(t)=-2H^2\alpha_1\alpha_2\begin{pmatrix} i\beta_1Hu_1+i\beta_2Hu_2 \\ i\beta_1Hu_1-i\beta_2Hu_2 \\ u_1'+u_2' \\ u_1'-u_2' \end{pmatrix}.$$

Since $\|i\beta_jHu_j\|^2=P_j(t)$ and $\|u_j'\|^2=K_j(t), j=1,2$, we see, from Remark 2, polarization, and the Riemann-Lebesgue lemma, why the components, $\|w_j\|^2$, are equipartitioned for $j=1,\cdots,4$.

In addition, observe that the two waves, u_1 and u_2, are such that $\langle u_1, u_2\rangle\to 0$ as $t\to\pm\infty$.

In the general case of $n=2^m$ in (7), using d'Alembert's formula to write the solution of (7) as

$$u(t)=\sum_{j=1}^{n} T_j(t)\phi_k,$$

where the ϕ_k depend on the initial data, we can show (after tedius calculations) that each component, w_k, of the first order system is of the form

$$w_k(t) = \sum_{j=1}^{n} a_{jk} T_j(t)\psi_j,$$

where $a_{jk} = 0$ or 1, and $\{\psi_j\}$ depends on $\{\phi_j\}$ but not on k. In the case of (11), knowing that $u_j = T_j(t)\phi_j + T_j(-t)\psi_j$, $j = 1, 2$, we have that

$$U(t) = \begin{pmatrix} T_1(t)\phi_1^* + T_1(-t)\psi_1^* + T_2(t)\phi_2^* + T_2(-t)\psi_2^* \\ T_1(t)\phi_1^* + T_1(-t)\psi_1^* - T_2(t)\phi_2^* - T_2(-t)\psi_2^* \\ T_1(t)\phi_1^* - T_1(-t)\psi_1^* + T_2(t)\phi_2^* - T_2(-t)\psi_2^* \\ T_1(t)\phi_1^* - T_1(-t)\psi_1^* - T_2(t)\phi_2^* + T_2(-t)\psi_2^* \end{pmatrix},$$

where $\phi_j^* = -2i\beta_j H^3\phi_j$ and $\psi_j^* = -2i\beta_j H^3\psi_j$, $j = 1, 2$.

Using polarization and the Riemann-Lebesgue lemma, it can be seen in the general case that $\lim_{t\to\infty} \|w_k(t)\|^2 = c$, c being independent of k, $k = 1, \cdots, n = 2^m$. By conservation of energy, $c = n^{-1}\mathscr{E}(0)$, and we again see that the 2^m system exhibits equipartition of energy.

7. Applications to scattering theory. Let A_0, A_1 generate (C_0) semigroups on X. We define the *subspace of asymptotic equivalence* X_{ae} for the ordered pair (A_0, A_1) to consist of all $f \in X$ for which there is a vector $f_+ \in X$ such that $\|T_0(t)f_+ - T_1(t)f\| \to 0$ as $t \to \infty$, where T_j is the semigroup generated by A_j.

If A_0, A_1 generate (C_0) groups T_0, T_1 on X, let X_{ae}^+, X_{ae}^- denote, respectively, the subspaces of asymptotic equivalence for the pairs (A_0, A_1) and $(-A_0, -A_1)$. We say that the *scattering problem for* $(A_0\ A_1)$ *admits completeness* provided

(i) $X_{ae}^+ = X_{ae}^-$;

(ii) For each $g \in X$ there are vectors $g_\pm \in X_{ae}^\pm$ such that

$$\|T_1(t)g_+ - T_0(t)g\| + \|T_1(-t)g_- - T_0(-t)g\| \to 0 \quad \text{as } t \to \infty.$$

Suppose that A_0, A_1 are skew-adjoint operators on a Hilbert space X with iA_0 spectrally absolutely continuous. Let $H_j = iA_j$ for $j = 0, 1$. Then the scattering problem for (A_0, A_1) admits completeness if and only if the wave operators for the pair (H_0, H_1) exist and are complete (in the sense described in Kato's book [8]). In this case the scattering operator exists and is unitary, and $X_{ae}^\pm$ can be identified with the absolutely continuous subspace for H_1.

We can identify the semigroup T_j with the first order differential equation $u' = A_j u$ for X-valued functions on $\mathbb{R}$. The definitions given above extend to more general situations; we proceed rather informally. For $j = 0, 1$ let $(\mathrm{DE})_j$ be a linear ordinary differential equation for functions from $\mathbb{R}$ to X. We say that *the scattering problem for the pair* $(\mathrm{DE})_0$, $(\mathrm{DE})_1$ *admits completeness* provided the following two conditions hold:

(I) There is a space X_{ae} of continuous functions from $\mathbb{R}$ to X such that for every solution u_1 of $(\mathrm{DE})_1$ in X_{ae} there is a solution $u_\pm$ of $(\mathrm{DE})_0$ such that $\|u_1(t) - u_\pm(t)\| \to 0$ as $t \to \pm\infty$.

(II) For every solution u_0 of $(\mathrm{DE})_0$ there are solutions $u_\pm$ of $(\mathrm{DE})_1$ in X_{ae} such that

$$\|u_+(t) - u_0(t)\| + \|u_-(-t) - u_0(-t)\| \to 0 \quad \text{as } t \to \infty.$$

If we topologize the solution spaces of well-posed initial value problems by using suitable norm (or graph) topologies on the spaces of initial data, then we can extend the above definition (of completeness) to the context of a dense set of solutions.

Now suppose $(A_1^{(0)}, \cdots, A_n^{(0)})$, $(A_1^{(1)}, \cdots, A_n^{(1)})$ are two commuting families of infinitesimal generators of (C_0) groups on X, that is, let Hypothesis (H1) hold for $\{A_i^{(j)}: i=1, \cdots, n\}$ with $j=0, 1$. Assume that the scattering problem for the pair $(A_i^{(0)}, A_i^{(1)})$ admits completeness for $i=1, \cdots, n$. Finally assume $0 \in \rho(A_i^{(j)} - A_k^{(j)})$ for $i \neq k$ and $j=0, 1$. Then for $j=0, 1$, every solution of $\prod_{i=1}^n (d/dt - A_i^{(j)})u_j(t) = 0$ is of the form

$$u_j(t) = \sum_{i=1}^n \exp\{tA_i^{(j)}\}f_{ij}$$

by Theorem 2. (Similarly we could weaken the hypothesis that $0 \in \rho(A_i^{(j)} - A_k^{(j)})$ and use Theorem 4 instead.) It follows that the scattering problem for the pair of equations $\prod_{i=1}^n (d/dt - A_i^{(0)})u = 0$, $\prod_{i=1}^n (d/dt - A_i^{(1)})u = 0$ admits completeness.

Usually scattering theory for second order or higher order equations is developed by writing the equations as first order systems and using the standard results of scattering theory including the invariance principle and other tools (cf., e.g., Kato [9]). The abstract d'Alembert formula gives an alternate approach to these problems. The d'Alembert formula also enables one to extend in a natural and simple way the notions of scattering theory from the context of isometric groups and Hilbert space to the context of semigroups and Banach spaces.

We close with three simple examples of pairs of equations for which the scattering problem admits completeness. Let q be a real integrable C^∞ function on $\mathbb{R}$. Take $X = L^2(\mathbb{R})$ and

$$A_1^{(0)} = \frac{d}{dx}, \qquad A_2^{(0)} = -A_1^{(0)},$$

$$A_1^{(1)} = \frac{d}{dx} + iq(x), \qquad A_2^{(1)} = -A_1^{(1)}.$$

The corresponding differential equations are

$$(\mathrm{DE})_0 \qquad \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2},$$

$$(\mathrm{DE})_1 \qquad \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + 2iq(x)\frac{\partial u}{\partial x} + (iq'(x) - q^2(x))u.$$

For the second example take $X = L^2(\mathbb{R}^2)$,

$$A_1^{(0)} = -A_2^{(0)} = i(-\Delta)^{1/2},$$
$$A_1^{(1)} = -A_2^{(1)} = i(-\Delta + V(x))^{1/2},$$

where V is a nonnegative function in $L^p(\mathbb{R}^n) + L^\infty(\mathbb{R}^n)$ with $p \geqq 2$, $p > n/2$ and such that $V(x) = O(|x|^{-1-\varepsilon})$ for some $\varepsilon > 0$ as $|x| \to \infty$.

The final example seems not to have been treated in the literature before. Take $X = L^2(\mathbb{R}^2)$,

$$A_1^{(0)} = -A_2^{(0)} = \alpha i(-\Delta)^{1/2}, \qquad A_3^{(0)} = -A_4^{(0)} = \beta i(-\Delta)^{1/2},$$
$$A_1^{(1)} = -A_2^{(1)} = \alpha i(-\Delta + V)^{1/2}, \qquad A_3^{(1)} = -A_4^{(1)} = \beta i(-\Delta + V)^{1/2}.$$

Thus, with the aid of the Birman-Kato invariance principle, the scattering problem for the fourth order elastic wave equation with a potential term $V(x)$ (cf. Example 4 of § 3) reduces to the scattering problem for the Schrödinger equation with the same potential.

REFERENCES

[1] I. M. GEL'FAND, *On one-parameter groups of operators in normed spaces*, Dokl. Akad. SSSR, 25 (1939), pp. 713-718.

[2] J. A. GOLDSTEIN, *An asymptotic property of solutions of wave equations*, Proc. Amer. Math. Soc., 23 (1969), pp. 359-363.

[3] J. A. GOLDSTEIN AND S. I. ROSENCRANS, *Energy decay and partition for dissipative wave equations*, J. Differential Equations, 36 (1980), pp. 66-73 and 43 (1982), p. 156.

[4] J. A. GOLDSTEIN AND J. T. SANDEFUR, JR., *Asymptotic equipartition of energy theorems*, J. Math. Anal. Appl., 67 (1979), pp. 58-74.

[5] ———, *Equipartition of energy*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. III, H. Brezis and J. L. Lions, eds., Pitman, London, 1982, pp. 209-219.

[6] ———, *Equipartition of energy for higher order abstract hyperbolic equations*, Comm. Partial Differential Equations, 7 (1982), pp. 1217-1251.

[7] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, Amer. Math. Soc. Colloq. Publ., 31, Providence, RI, 1957.

[8] T. KATO, *Perturbation Theory for Linear Operators*, Springer, New York, 1966.

[9] ———, *Scattering theory with two Hilbert spaces*, J. Funct. Anal., 1 (1967), pp. 342-369.

[10] W. LITTMAN, *The wave operator and L^p-norms*, J. Math. Mech., 12 (1963), pp. 55-68.

[11] K. MOCHIZUKI, *Asymptotic property of solutions of some higher order hyperbolic equations*, I, II, Proc. Japan Acad. Ser. A Math. Soc., 46 (1970), pp. 262-272.

[12] J. T. SANDEFUR, JR., *Higher order abstract Cauchy problems*, J. Math. Anal. Appl., 60 (1977), pp. 728-742.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

019

GENERATION OF ANALYTIC SEMIGROUPS BY ELLIPTIC OPERATORS WITH UNBOUNDED COEFFICIENTS*

PIERMARCO CANNARSA† AND VINCENZO VESPRI‡

Abstract. Strongly elliptic differential operators with (possibly) unbounded lower order coefficients are shown to generate analytic semigroups of linear operators on $L^2(R^N)$, $L^{2,\mu}(R^N)$, $C(R^N)$ and $C^\alpha(R^N)$. Some of these generation results are applied to parabolic initial value problems.

Key words. analytic semigroups, elliptic operators, unbounded coefficients

AMS(MOS) subject classifications. 47D05, 47505

Introduction. A number of papers has been devoted to studying the generation of analytic semigroups by elliptic operators in unbounded domains $\Omega \subset R^N$ (see, e.g. [9], [22], [21], [28], [29]). Authors, however, have usually concentrated their attention on operators with bounded coefficients.

In this paper we consider strongly elliptic operators of second order, E, with coefficients defined on R^N and allowed to grow at infinity in accordance with suitable structure conditions. Such conditions are exactly the ones we need for the applications in which we are interested.

We prove that E generates an analytic semigroup of linear operators in various Banach spaces, such as $L^2(R^N)$, $L^{2,\mu}(R^N)$, $C(R^N)$ and $C^\alpha(R^N)$ (see § 1 for a detailed description of these results). Furthermore, all these results are obtained by the same method.

It has to be noted that we have confined ourselves to the case of second order operators defined in the whole space for the sake of simplicity and also because this is what we need for the applications we have in mind. Our method may be extended to more general situations (for example, Dirichlet boundary conditions may be treated by adapting the boundary analysis of [14]). In fact, we use $\mathscr{L}^{2,\lambda}$-regularity techniques which also apply to solutions of elliptic systems of equations.

The main applications of our results concern the initial value problem for parabolic equations with unbounded coefficients. Although such equations have been studied for a long time (see, e.g. [23], [24], [3], [4], [8], [7]), interest in this topic has increased since connections with stochastic control and filtering theory were pointed out ([6], [20], [5]). While earlier treatments were mainly based on fundamental solutions, more recent ones have also developed different approaches, using dynamic programming [20], semigroup theory [18] or the Feynman-Kac formula [27].

In § 7 of this paper we obtain the existence, uniqueness and regularity of solutions to the Cauchy problem for second order parabolic operators with unbounded coefficients, as a consequence of the generation theorems proved in §§ 2-6. The results of these sections are applied to a class of stochastic partial differential equations in [16].

Part of the contents of this paper was announced by the authors in [13] and [30].

* Received by the editors December 10, 1984; accepted for publication (in revised form) April 7, 1986. The authors are members of GNAFA (Consiglio Nazionale delle Ricerche).

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, on leave from Gruppo Insegnamento Matematiche, Academia Navale, 57100 Livorno, Italy.

‡ Dipartimento di Matematica, II Universita degli Studi di Roma, Via Orazio Raimondo, 00173 Roma, Italy.

1. Notation and main results. We denote by $\|x\|$ (resp. $\|z\|$) the Euclidean norm of a point $x \in R^N$ (resp. $z \in C^N$). For any $x \in R^N$ and $r > 0$ we define

$$B_r(x) = \{y \in R^N: \|y - x\| < r\}.$$

For any $j = 1, \cdots, N$ and any multi-index $\beta = (\beta_1, \cdots, \beta_N)$ we set

$$D_j = \partial/\partial x_j, \qquad D^\beta = D_1^{\beta_1} \cdots D_N^{\beta_N}.$$

Let Ω be an open domain in R^N. We denote by $C(\Omega)$ the space of uniformly continuous and bounded complex-valued functions defined in Ω. We set $|u|_{0,\infty,\Omega} = \sup_{x \in \Omega} |u(x)|$ for any $u \in C(\Omega)$. Moreover, for any integer $m > 0$, we define $C^m(\Omega) = \{u \in C(\Omega): D^\beta u \in C(\Omega) \text{ for } |\beta| \leqq m\}$ and $|u|_{j,\infty,\Omega} = \sum_{|\beta|=j} |D^\beta u|_{0,\infty,\Omega}$ for $0 \leqq j \leqq m$.

In this paper we consider second order differential operators of the form

$$E = \sum_{i,j=1}^{N} a_{ij}(x) D_i D_j + \sum_{j=1}^{N} b_j(x) D_j - c(x) \tag{1.1}$$

where a_{ij}, b_j, c $(i, j = 1, \cdots, N)$ are complex-valued functions defined in R^N. We will always assume that E is strongly elliptic, i.e.,

$$\operatorname{Re} \sum_{i,j} a_{ij}(x) z_j \bar{z}_i \geqq \nu \|z\|^2 \quad \forall z \in C^N \tag{1.2}$$

for any $x \in R^N$ and some $\nu > 0$. The growth of the coefficients of E will be referred to a given function $V: R^N \to R$ such that for any $x \in R^N$ and any $y \in R^N$ with $\|x - y\| \leqq 1$

$$\begin{array}{ll} \text{(i)} & V(x) \geqq 1, \\ \text{(ii)} & |V(x) - V(y)| \leqq k\|x - y\|[V(x) + V(y)] \quad \text{for some } k > 0. \end{array} \tag{1.3}$$

More precisely, we will impose the following assumptions:

There exist constants $K_1, K_2, V_0 > 0$ *such that*

$$K_1 V(x) - V_0 \leqq \operatorname{Re} c(x) \leqq K_2 V(x) + V_0 \quad \forall x \in R^N. \tag{1.4}$$

There exists a constant $B \in [0, 2]$ *such that*

$$\sum |b_j(x)|^2 \leqq \nu B^2 K_1 V(x) \quad \forall x \in R^N, \tag{1.5}$$

$$a_{ij},\ b_j/V^{1/2},\ c/V \in C(R^N) \quad \text{for } i, j = 1, \cdots, N. \tag{1.6}$$

Since the top order coefficients are assumed to be bounded, this framework is roughly comprable with the one of [6, pp. 108–117]. Condition (1.4) is similar to one of the hypotheses of [23]. However, here we need no differentiability assumptions on the coefficients of E. Moreover, condition (1.5) only affects the asymptotic behavior of coefficients b_j.

Remark 1.1. From assumption (1.3)(ii) it follows that there exists a constant $k > 0$ such that

$$V(x) \leqq k V(y) \quad \text{if } \|x - y\| \leqq 1. \tag{1.7}$$

We begin by analyzing the spectral properties of operator E when realized in $L^2(R^N)$. Let Ω be an open domain in R^N and m a nonnegative integer. We denote by $H^m(\Omega)$ the space of the complex-valued functions u, defined in Ω, which belong to $L^2(\Omega)$ together with all their partial derivatives of orders $|\beta| \leqq m$. We define

$$|u|_{0,\Omega} = \left(\int_\Omega |u(x)|^2 \, dx \right)^{1/2}, \qquad |u|_{j,\Omega} = \left(\sum_{|\alpha|=j} |D^\alpha u|_{0,\Omega}^2 \right)^{1/2}$$

for $1 \leqq j \leqq m$. Moreover, we denote by $H^m_{\mathrm{loc}}(\Omega)$ the space of functions $u \in H^m(D)$ for any domain D with compact closure in Ω.

Since the coefficients of E may grow at infinity, we need weighted spaces rather than the usual Sobolev spaces $H^m(\Omega)$. For $j = 1, 2$ we denote by $H^j(\Omega, V)$ the space of functions $u \in H^j(\Omega)$ such that

$$\|u\|^2_{j,\Omega,V} = \sum_{|\beta| \leqq j} \int_\Omega V^{j-|\beta|} |D^\beta u|^2 \, dx < +\infty.$$

Extending the method [14, § 3], in § 2 of this paper we prove the following result.

THEOREM 1.2. *Assume* (1.2)-(1.6). *There exists a number* $\omega_1 \geqq V_0$ *such that, if* λ *is a complex number with* $\operatorname{Re} \lambda \geqq \omega_1$ *then the equation* $(\lambda - E)u = f \in L^2(R^N)$ *has a unique solution* $u \in H^2(R^N, V)$ *and*

$$\|u\|_{2,R^N,V} + |\lambda - \omega_1|^{1/2} \|u\|_{1,R^N,V} + |\lambda - \omega_1| \, |u|_{0,R^N} \leqq k |f|_{0,R^N} \tag{1.8}$$

where k is a constant independent of λ.

The previous result can also be extended to more general weighted spaces (see § 2 for details).

For $0 \leqq \mu \leqq N$ we denote by $L^{2,\mu}(\Omega)$ the space of functions u $L^2_{\mathrm{loc}}(\Omega)$ such that

$$|u|^2_{L^{2,\mu}(\Omega)} = \sup_{x \in \Omega, 0 < r \leqq 1} r^{-\mu} \int_{\Omega \cap B_r(x)} |u(y)|^2 \, dy < +\infty.$$

Morrey spaces $L^{2,\mu}(\Omega)$ are useful to study the regularity of solutions to elliptic problems (see e.g. [10]).

In § 4 we show that E generates an analytic semigroup on the Banach space $L^2(R^N) \cap L^{2,\mu}(R^N)$, equipped with the norm

$$\|u\|_{0,\mu,R^N} = |u|_{0,R^N} + |u|_{L^{2,\mu}(R^N)}.$$

For $j = 1, 2$ let us denote by $H^{j,\mu}(\Omega, V)$ the space of function $u \in H^j(\Omega, V)$ such that

$$|u|^2_{H^{j,\mu}(\Omega,V)} = \sum_{|\beta| \leqq j} |V^{(j-|\beta|)/2} D^\beta u|^2_{L^{2,\mu}(\Omega)} < +\infty$$

and set

$$\|u\|_{H^{j,\mu}(\Omega,V)} = \|u\|_{j,\Omega,V} + |u|_{H^{j,\mu}(\Omega,V)}.$$

THEOREM 1.3. *Assume* (1.2)-(1.6) *and let* $\lambda \in C$ *with* $\operatorname{Re} \lambda \geqq \omega_1$. *Then the equation*

$$(\lambda - E)u = f \in L^2(R^N) \cap L^{2,\mu}(R^N), \qquad 0 < \mu < N$$

has a unique solution $u \in H^{2,\mu}(R^N, V)$ *and*

$$\|u\|_{H^{2,\mu}(R^N,V)} + |\lambda - \omega_1|^{1/2} \|u\|_{H^{1,\mu}(R^N,V)} + |\lambda - \omega_1| \, \|u\|_{0,\mu,R^N} \leqq k \|f\|_{0,\mu,R^N} \tag{1.9}$$

where k is a constant independent of λ.

See [14, § 5] for a similar result in the case of bounded space domains.

In [14, § 6], generation on Morrey spaces is used to derive the uniform estimate obtained by Stewart [28], [29]. Stewart's method consists in bounding suitable localizations of solutions by classical L^p estimates [2] and applies to elliptic operators of arbitrary order with bounded coefficients. The analogue of Stewart's results for operators with unbounded coefficients is stated below and will be proved in § 5.

THEOREM 1.4. *Assume* (1.2)-(1.6). *There exists* $\omega_2 \geqq \omega_1$ *such that, if* λ *is a complex number with* $\operatorname{Re} \lambda \geqq \omega_2$, *then the equation* $(\lambda - E)u = f\, L^\infty(R^N)$ *has a unique solution* $u \in H^2_{\text{loc}}(R^N) \cap C^1(R^N)$ *satisfying* $|V^{1/2}u|_{0,\infty,R^N} < +\infty$. *Moreover*

$$|\lambda - \omega_2|^{1/2}[|u|_{1,\infty,R^N} + |V^{1/2}u|_{0,\infty,R^N}] + |\lambda - \omega_2| \cdot |u|_{0,\infty,R^N} \leqq k|f|_{0,\infty} \tag{1.10}$$

where k is a constant independent of λ.

Finally, we turn to the last topology we shall consider in this paper, namely Hölder topology. For $0 < \alpha < 1$ we denote by $C^\alpha(\Omega)$ the space of functions $u \in C(\Omega)$ such that

$$[u]_{\alpha,\Omega} = \sup_{x,y\Omega, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|^\alpha} < +\infty.$$

The Hölder norm is defined as follows:

$$\|u\|_{\alpha,\Omega} = |u|_{0,\infty,\Omega} + [u]_{\alpha,\Omega}.$$

Generation of analytic semigroups in Hölder spaces was first obtained by Campanato [12] in bounded space domains (see also [14, § 7] for the case of systems). A different procedure leading to an analogous generation theorem was developed in [25]. This procedure is based on interpolation techniques and on Stewart's result [28], [29] and applies to operators with bounded coefficients. In this paper we treat operators with unbounded coefficients by adapting the method of [12].

For $j = 1, 2$ we denote by $C^j(\Omega, V)$ the space of functions $u \in C^j(\Omega)$ such that

$$\|u\|_{Cj(\Omega,V)} = \sum_{|\beta| \leqq j} |V^{(j-|\beta|)/2} D^\beta u|_{0,\infty,\Omega} < \infty.$$

For $0 < \alpha < 1$ we set

$$C^{j,\alpha}(\Omega, V) = \{u \in C^j(\Omega, V) : |u|_{C^{j,\alpha}(\Omega,V)} = \sum_{|\beta| \leqq j} [V^{(j-|\beta|)/2} D^\beta u]_{\alpha,\Omega} < \infty\}.$$

We equip $C^{j,\alpha}(\Omega, V)$ with the norm

$$\|u\|_{C^{j\alpha}(\Omega,V)} = \|u\|_{C^j(\Omega,V)} + |u|_{C^{j,\alpha}(\alpha,V)}.$$

THEOREM 1.5. *Assume* (1.2)-(1.5) *and suppose that*

$$a_{ij}, b_j / V^{1/2}, c/V \in C^\alpha(R^N) \quad \textit{for some } \alpha \in]0, 1[. \tag{1.11}$$

There exists $\omega_2 \in R$ *such that, if* λ *is a complex number with* $\operatorname{Re} \lambda \geqq \omega_2$, *then the equation* $(\lambda - E)u = f \in C^\alpha(R^N)$ *has a unique solution* $u \in C^{2,\alpha}(R^N, V)$ *and*

$$\|u\|_{C^{2,\alpha}(R^N,V)} + |\lambda - \omega_2|^{1/2} \|u\|_{C^{1,\alpha}(R^N,V)} + |\lambda - \omega_2| \cdot \|u\|_{\alpha,R^N} \leqq k\|f\|_{\alpha,R^N} \tag{1.12}$$

where k is a constant independent of λ

Therefore, E generates an analytic semigroup on $C^\alpha(R^N)$ with domain $C^{2,\alpha}(R^N, V)$. Our semigroup fails to be strongly continuous at 0 because $C^{2,\alpha}(R^N, V)$ is not dense in $C^\alpha(R^N)$. However, the closure $C^{2,\alpha}(R^N, V)$ of the domain of E can be characterized as follows. Let us set

$$C^\alpha_V(R^N) = \{u \in C^\alpha(R^N) : Vu \in C^\alpha(R^N)\},$$

$$C^\alpha_0(R^N) = \{u \in C^\alpha(R^N) : u \to 0 \quad \text{as } \|x\| \to \infty\}.$$

Then $\overline{C^{2,\alpha}(R^N, V)}$ consists of all functions $u \in C^\alpha_V(R^N) \cup C^\alpha_0(R^N)$ such that

$$\lim_{r \to 0} \left[\sup_{0 < \|x-y\| < r} \frac{|u(x) - u(y)|}{\|x - y\|^\alpha} \right] = 0.$$

This space, which is also known as the space of "little-Hölder continuous" functions (see [25]), will be denoted by $h^\alpha(R^N, V)$.

2. Generation in the L^2 topology. In this section we prove Theorem 1.2. As in [14, § 3], the proof consists of two steps. First, we obtain an estimate for divergence form elliptic operators (Lemma 2.1 below). Then, we approximate operator E by operators in divergence form and apply the contraction mapping theorem. Therefore, we have concentrated our exposition on what is new with respect to [14, § 3]. The remainder of the proof is just sketched.

Consider the operator

$$\mathscr{E} = \sum_{i,j=1}^{N} D_j(a_{ij}(x)D_j) + \sum b_j(x)D_j - c(x)$$

with measurable complex-valued coefficients satisfying (1.2), (1.4), (1.5) and

$$a_{ij},\ b_j/V^{1/2},\ c/V \in L^\infty(R^N), \qquad i,j = 1, \cdots, N. \tag{2.1}$$

Here V is a real-valued function satisfying (1.3). For $\lambda \in C$, $u \in H^1(R^N, V)$ and $\phi \in C^\infty(R^N)$ with compact support, let us set

$$a_\lambda(u, \phi) = \lambda \int_{R^N} u\bar{\phi}\,dx + \int_{R^N} \left[\sum_{i,j} a_{ij}D_j\overline{uD_i\phi} - \sum b_jD_ju\bar{\phi} + cu\bar{\phi}\right] dx.$$

Then, a_λ can be extended to a continuous sesquilinear form on $H^1(R^N, V) \times H^1(R^N, V)$. Moreover, α_λ is coercive for $\operatorname{Re}\lambda \geqq V_0$.

$$(Re\,\lambda - V_0)|u|^2_{0,R^N} + k\|u\|^2_{1,R^N,V} \leqq \operatorname{Re} a_\lambda(u, u) \quad \forall u \in H^1(R^N, V)$$

(from now on we denote by k any positive constant independent of λ). Therefore, by standard variational arguments we conclude that for $\operatorname{Re}\lambda \geqq V_0$ and $f \in L^2(R^N)$ the equation

$$a_\lambda(u, \phi) = \int_{R^N} f\bar{\phi}\,dx \quad \forall \phi \in H^1(R^N, V) \tag{2.2}$$

has a unique solution $u \in H^1(R^N, V)$ and

$$\|u\|^2_{1,R^N,V} \leqq k|f|_{0,R^N}|u|_{0,R^N}. \tag{2.3}$$

Moreover, from (2.3) one can easily get

$$|\lambda - V_0| \cdot |u|_{0,R^N} \leqq k|f|_{0,R^N} \tag{2.4}$$

and

$$|\lambda - V_0|^{1/2} \cdot \|u\|_{1,R^N,V} \leqq k|f|_{0,R^N}. \tag{2.5}$$

Now, if the coefficients of $\mathscr{E}$ are more regular, then u solves our equation a.e. as we show below.

LEMMA 2.1. *Let the coefficients of $\mathscr{E}$ be differentiable in R^N and satisfy, in addition to* (1.2), (1.4), (1.5), (2.1), *the following condition*:

$$|D_ha_{ij}|_{0,\infty,R^N} + |D_hb_j/V^{1/2}|_{0,\infty,R^N} + |D_hc/V|_{0,\infty,R^N} = K < +\infty \tag{2.6}$$

for any $h = 1, \cdots, N$. If $\operatorname{Re}\lambda \geqq V_0$, *then the solution u of equation $(\lambda - \mathscr{E})u = f \in L^2(R^N)$ is of class $H^2(R^n, V)$ and*

$$\|u\|_{2,R^N,V} \leqq \left(k_1 + \frac{k_2}{|\lambda - V_0|^{1/2}}\right)|f|_{0,R^N} \tag{2.7}$$

where both k_1 and k_2 are independent of λ and k_1 is independent of K.

Proof. Since $u \in H^2_{\mathrm{loc}}(R^N)$ by classical regularity results, in (2.2) we can choose $\phi = D_h\psi$, where $\psi \in C^\infty(R^N)$ with compact support. A simple integration by parts yields

$$\begin{aligned} a(D_hu, \psi) = &-\int_{R^N} f\overline{D_h\psi}\, dx - \int_{R^N} \sum_{ij} D_ha_{ij}D_ju\overline{D_i\psi}\, dx \\ &-\int_{R^N} \left(D_hcu - \sum_j D_hb_jD_ju \right) \bar{\psi}\, dx. \end{aligned} \tag{2.8}$$

Clearly, (2.8) holds for any $\psi \in H^1(R^N)$ with compact support.

Let θ be a standard cutoff function, i.e. $\theta \in C^\infty(R^N)$ and

$$\begin{aligned} &0 \leqq \theta \leqq 1, \qquad \theta(y) = 1 \quad \text{if } \|y\| \leqq 1, \\ &\theta(y) = 0 \quad \text{if } \|y\| \geqq 2 \end{aligned} \tag{2.9}$$

and set $\theta_r(x) = \theta(x/r)$ for any $r > 0$. By assuming $\psi = \theta_r^2 D_hu$ in (2.8) we obtain

$$a_\lambda(\theta_rD_hu, \theta_rD_hu) = F_r + G_r$$

where

$$F_r = \int_{R^N} \theta_rfD_h\theta_R\overline{D_hu}\, dx - \int_{R^N} \theta_rf\overline{D_h(\theta_rD_hu)}\, dx$$

and the remaining term G_r contains the derivatives of a_{ij}, b_j, c. Then

$$\|\theta_rD_hu\|^2_{1,R^N,V} \leqq k(|F_r| + |G_r|). \tag{2.10}$$

Now, from (1.4), (1.5), (2.1), (2.6) we obtain, for any $\varepsilon > 0$

$$|G_r| \leqq \varepsilon \|\theta_rD_hu\|^2_{1,R^N,V} + k(\varepsilon, K)\|u\|^2_{1,R^N,V}. \tag{2.11}$$

Moreover

$$|F_r| \leqq \varepsilon \|\theta_rD_hu\|^2_{1,R^N,V} + k(\varepsilon)|f|^2_{0,R^N}. \tag{2.12}$$

By choosing ε sufficiently small and letting r go to infinity, from (2.10), (2.11), (2.12) we get

$$\sum_h \|D_hu\|^2_{1,R^N,V} \leqq k|f|^2_{0,R^N} + k(K)\|u\|^2_{1,R^N,V}. \tag{2.13}$$

Therefore, equation $(\lambda - \mathscr{E})u = f$ is satisfied a.e. and

$$(c + V_0)u = f + (V_0 - V)u + \sum_{ij} (a_{ij}D_iD_ju + D_ia_{ij}D_ju) + \sum_j b_jD_ju.$$

Consequently, $u \in H^2(R^N, V)$ and

$$|Vu|_{0,R^N} \leqq |f|_{0,R^N} + |\lambda - V_0|\, |u|_{0,R^N} + k\sum_h \|D_u\|_{1,R^N,V} + k(K)\|u\|_{1,R^N,V}.$$

The desired estimate (2.7) may thus be derived from the last inequality and from (2.13), recalling (2.4) and (2.5).

We now turn to the case of an operator E of type (1.1).

Proof of Theorem 1.2. Suppose assumptions (1.2), (1.4) hold for the coefficients of E with constants ν, K_1, K_2, V_0, B and set

$$\sum_{ij} |a_{ij}|_{0,\infty,R^N} + \sum_j |b_j/V^{1/2}|_{0,\infty,R^N} + |C/V|_{0,\infty,R^N} = M. \tag{2.14}$$

Now, if $(J_\varepsilon)_{\varepsilon>0}$ is a standard family of mollifiers (i.e. $J_\varepsilon(x)=\varepsilon^{-n}J(x/\varepsilon)$, where $J\in C^\infty(R^N)$, $J\geqq 0$, $J(x)=0$ if $\|x\|\geqq 1$ and $\int_{R^N}J(x)\,dx=1$), let us set

$$E^\varepsilon=\sum a_{ij}^\varepsilon(x)D_iD_j+\sum b_j^\varepsilon(x)D_j-c^\varepsilon(x),\qquad x\in R^N,\quad \varepsilon>0.$$

Here, $g^\varepsilon(x)=\int_{R^N}J_\varepsilon(x-y)g(y)dy$ for any $g\in L^1_{\text{loc}}(R^N)$. It is easy to check that E^ε satisfy (1.2), (1.4), (2.15) with the same constants ν, K_1, K_2, V_0, B, M as above, replacing V by V^ε in each inequality. Moreover

$$|D_ha_{ij}^\varepsilon(x)|+|D_hD_sa_{ij}^\varepsilon(x)|+|D_hb_j^\varepsilon(x)/V^\varepsilon(x)^{1/2}|+|D_hc^\varepsilon(x)/V^\varepsilon(x)|\leqq K_\varepsilon<+\infty$$

for any $x\in R^N$, $\varepsilon>0$ and $1\leqq i,j,h,s\leqq N$. Also, E^ε approximates E as $\varepsilon\to 0$ in the sense that

$$(2.15)\quad |a_{ij}-a_{ij}^\varepsilon|_{0,\infty,R^N},\ \left|\frac{b_j-b_j^\varepsilon}{V^{\varepsilon^{1/2}}}\right|_{0,\infty,R^N},\ \left|\frac{c-c^\varepsilon}{V^\varepsilon}\right|_{0,\infty,R^N},\ \left|\frac{V-V^\varepsilon}{V^\varepsilon}\right|_{0,\infty,R^N}\to 0\quad \text{as } \varepsilon\to 0.$$

Therefore, the elliptic problems

$$\begin{cases}u\in H^2(R^N,V)\\(\lambda-E)u=f\end{cases}\quad\text{and}\quad\begin{cases}u\in H^2(R^N,V^\varepsilon)\\(\lambda-E^\varepsilon)u=f+(E-E^\varepsilon)u\end{cases}$$

are equivalent if ε is sufficiently small. Let us set

$$\mathscr{E}^\varepsilon=\sum_{ij}D_i(a_{ij}^\varepsilon D_j)+\sum\left(b_j^\varepsilon-\sum_i D_ia_{ij}^\varepsilon\right)D_j-c^\varepsilon.$$

By Lemma 2.1 we conclude that there exists V_0^ε such that for $\operatorname{Re}\lambda\geqq V_\varepsilon^0$ and $u\in H^2(R^N,V^\varepsilon)$ the equation

$$(\lambda-\mathscr{E}^\varepsilon)U=f+(E-E^\varepsilon)u$$

has a unique solution $U=T_{\lambda,\varepsilon}$, $u\in H^2(R^N,V^\varepsilon)$. Also, from (2.4), (2.5), (2.7), (2.15) we get

$$\begin{aligned}&\|T_{\lambda,\varepsilon}u\|_{2,R^N,V^\varepsilon}+|\lambda-V_0^\varepsilon|^{1/2}\cdot\|T_{\lambda,\varepsilon}u\|_{1,R^NV^\varepsilon}+|\lambda-V_0^\varepsilon|\cdot|T_{\lambda,\varepsilon}u|_{0,R^N}\\(2.16)\qquad&\leqq\left(k_1+\frac{k_2(\varepsilon)}{|\lambda-V_0^\varepsilon|^{1/2}}\right)(|f|_{0,R^N}+|(E-E^\varepsilon)u|_{0,R^N})\\&\leqq\left(k_1+\frac{k_2(\varepsilon)}{|\lambda-V_0^\varepsilon|^{1/2}}\right)(|f|_{0,R^N}+k\sigma(\varepsilon)\|u\|_{2,R^N,V^\varepsilon})\end{aligned}$$

where $\sigma(\varepsilon)\to 0$ as $\varepsilon\to 0$ and k_1 is independent of ε. From (2.16) it follows that we can choose ε so small as to have $k_1k\sigma(\varepsilon)<1$ and then fix $\omega_1\geqq V_0^\varepsilon$ so large as to ensure that $T_{\lambda,\varepsilon}$ is a contraction mapping for $\operatorname{Re}\lambda\geqq\omega_1$. The remainder of the proof is standard.

Remark 2.2. A useful generalization of Theorem 2.1 can be obtained in a very straightforward way by using weighted spaces. Let π be a twice differentiable real-valued function defined on R^N and set $\Pi(x)=\exp(\pi(x)]$. Denote by $L^2_\Pi(R^N)$ resp. $H^j_\Pi(R^N,V)$, $j=1,2$) the space of functions u such that $\Pi u\in L^2(R^N)$ (resp. $\Pi u\in H^j(R^N,V)$, $j=1,2$). Then, for $f\in=L^2_\Pi(R^N)$

$$\begin{cases}u\in H^2_\Pi(R^N,V)\\(\lambda-E)u=f\end{cases}\quad\text{if and only if}\quad\begin{cases}\Pi u\in H^2(R^N,V)\\(\lambda-E_\pi)(\Pi u)=\Pi f\end{cases}$$

where

$$E_\pi=E-\sum_{ij}(a_{ij}+a_{ij})D_i\pi D_j-\sum b_jD_j\pi-\sum a_{ij}(D_iD_j\pi-D_i\pi D_j\pi).$$

As easily seen, operator E_π satisfies the assumptions of Theorem 1.2 provided that for each $\varepsilon > 0$ there exists $k(\varepsilon) \in R$ such that

$$\sum_j |D_j\pi(x)|^2 + \sum_{ij} |D_iD_j\pi(x)| \leqq \varepsilon V(x) + k(\varepsilon) \quad \forall x \in R^N. \tag{2.17}$$

Therefore, by applying Theorem 1.2 to E_π, we conclude that E generates an analytic semigroup on $L^2_{\Pi}(R^N)$.

3. Some useful lemmas. In this section we have collected some miscellaneous results that will be applied often in the sequel.

Let E be a differential operator of type (1.1), satisfying conditions (1.2)-(1.6). Let M be defined as in (2.14). If $x_0 \in R^N$, we set

$$E^0 = \sum_{ij} a_{ij}(x^0)D_iD_j + \sum_j b_j(x^0)D_j - c(x^0). \tag{3.1}$$

In the sequel, we will abbreviate $B_r = B_r(x^0)$ for any $r > 0$ and denote by k any constant depending only on ν, k_1, k_2, B and M. The result below is well known.

LEMMA 3.1. *For any $f \in L^2(B_r)$ and $\lambda \in C$ with* $\operatorname{Re} \lambda \geqq V_0$ *the Dirichlet problem*

$$(\lambda - E^0)u = f \quad \text{in } B_r, \qquad u = 0 \quad \text{on } \partial B_r$$

has a unique solution $u \in H^2(B_r)$ and

$$\|u\|_{2,B_r,V} \leqq k|f|_{0,B_r}. \tag{3.2}$$

Suppose that $u \in H^1(B_r)$ is a (weak) solution of the equation

$$(\lambda - E^0)u = \zeta \in C, \qquad \operatorname{Re} \lambda \geqq V_0. \tag{3.3}$$

Then, the following inequalities can be obtained arguing as in [12, p. 499]: denoting by u_{B_r} the integral mean value of u on B_r, for $0 < s < r$

$$|u|_{1,B_s} \leqq \frac{k}{r-s}\left(\frac{r}{s}\right)^{N/2} |u - u_{B_r}|_{0,B_r} \tag{3.4}$$

and, if $\zeta = 0$,

$$|u|_{1,B_s} \leqq \frac{k}{r-s}|u|_{0,B_r}. \tag{3.5}$$

Estimates (3.4) and (3.5) imply the result below, proved by Campanato in the case of bounded coefficients.

LEMMA 3.2. *If $u \in L^2(B_r)$ is a solution of* (3.3), *then*

$$|u - u_{B_{\tau r}}|^2_{0,B_{\tau r}} \leqq k\tau^{N+2}|u - u_{B_r}|^2_{0,B_r} \tag{3.6}$$

for any $\tau \in [0, 1]$. Moreover, if $\zeta = 0$, then

$$|u|^2_{0,B_{\tau r}} \leqq k\tau^N|u|^2_{0,B_r}. \tag{3.7}$$

We merely need to recall that (3.7) may be derived from (3.5) by the same procedure used in [12, p. 503] to show that u satisfies

$$|u|^2_{1,B_{\tau r}} \leqq k\tau^N|u|^2_{1,B_r}. \tag{3.8}$$

Estimate (3.6), which is trivial if $\frac{1}{2} \leqq \tau \leqq 1$, will then follow from (3.8) and (3.4) for $\tau \in [0, \frac{1}{2}]$. Indeed, by Poincaré's inequality we have

$$\begin{aligned} |u - u_{B_{\tau r}}|^2_{0,B_{\tau r}} &\leqq k(\tau r)^2|u|^2_{1,B_{\tau r}} \\ &\leqq k\tau^{N+2}r^2|u|^2_{1,B_{r/2}} \leqq k\tau^{N+2}|u - u_{B_r}|^2_{0,B_r}. \end{aligned}$$

Remark 3.3. Estimate (3.6) is useful to study the Hölder continuity of solutions to differential problems. Indeed, from the results of [10] it follows that

$$[u]^2_{\alpha,B_r} \leqq k(N) \sup_{xB_r,0<\rho\leqq r} \rho^{-(N+2\alpha)} \int_{B_\rho(x)} |u(y)-u_{B_\rho(x)}|^2 \, dy \tag{3.9}$$

for any function $u \in L^2_{\text{loc}}(R^N)$ and any ball $B_r \subset R^N$.

We conclude this section recalling a technical lemma proved in [11, p. 136].

LEMMA 3.4. *Let ϕ and σ be nonnegative functions on an interval $]0, d]$. Suppose that $\lim_{r\to 0} \sigma(r) = 0$ and*

$$\phi(\tau r) \leqq [K\tau^\alpha + \sigma(r)]\phi(r) + Lr^\beta$$

for any $r \in]0, d]$, any $\tau \in]0, 1]$ and some positive constants α, β, K, L with $\alpha > \beta$. Then, for any $\varepsilon \in]0, \alpha-\beta[$ there exists $r_\varepsilon \leqq d$ such that

$$\phi(\tau r) \leqq (1+K)\tau^{\alpha-\varepsilon}\phi(r) + K_\varepsilon L(\tau r)^\beta \quad \forall r \in]0, r_\varepsilon], \quad \forall \tau \in]0, 1]$$

where

$$K_\varepsilon = \frac{(1+K)^{2\alpha/\varepsilon}}{(1+K)^{(\alpha-\beta)/\varepsilon} - (1+K)}.$$

Furthermore, if $\sigma \equiv 0$, then $r_\varepsilon = d$.

4. Generation in Morrey spaces. The object of this section is the proof of Theorem 1.3, that we obtain as a consequence of Theorem 1.2 and of the local estimates recalled in the previous section. Our first step is the following lemma which adapts the idea of [11] to the present situation.

LEMMA 4.1. *Assume (1.2)-(1.6) and let $\lambda \in C$, $\operatorname{Re} \lambda \geqq V_0$. Suppose $u \in H^2_{\text{loc}}(R^N)$ is a solution of the equation $(\lambda - E)u = f\, L^{2,\mu}(R^N)$, $0 < \mu < N$. Then, for any $x^0 \in R^N$ and any $0 < r \leqq 1$,*

$$\|u\|^2_{2,B_r(x^0),V} \leqq kr^\mu[\|u\|^2_{2,B_1(x^0),V} + |f|^2_{L^{2,\mu}(R^N)}] \tag{4.1}$$

where k is a constant independent of λ, r.

Proof. Let $x^0 \in R^N$, $B_r = B_r(x^0)$, $0 < r \leqq 1$ and $w \in H^2(B_r)$ be the solution of the Dirichlet problem

$$\begin{aligned} (\lambda - E^0)w &= f + (E - E^0)u \quad &&\text{in } B_r, \\ w &= 0 &&\text{on } \partial B_r \end{aligned}$$

where E^0 is defined as in (3.1). Then, by (3.2) and (1.6) we obtain

$$\|w\|^2_{2,B_r,V} \leqq k[|f|^2_{0,B_r} + |(E-E^0)u|^2_{0,B_r}] \leqq k[\sigma(r)\|u\|^2_{2,B_r,V} + r^\mu |f|^2_{L^{2,\mu}(R^N)}] \tag{4.2}$$

where $\sigma : [0, +\infty[\to [0, +\infty[$ is increasing and $\lim_{r\to 0} \sigma(r) = 0$.

On the other hand, the difference $v = u - w$ satisfies the equation $(\lambda - E^0)v = 0$. so, applying (3.7) to every partial derivative $D^\alpha v$, $|\alpha| \leqq 2$ and using property (1.7) of V, we conclude that

$$\|v\|^2_{2,B_{\tau r},V} \leqq k\tau^N \|v\|^2_{2,B_r,V} \quad \forall \tau \in]0, 1]. \tag{4.3}$$

Therefore, (4.2) and (4.3) imply

$$\begin{aligned} \|u\|^2_{2,B_{\tau r},V} &\leqq 2[\|v\|^2_{2,B_{\tau r},V} + \|w\|^2_{2,B_{\tau r},V}] \\ &\leqq k[(\tau^N + \sigma(r))\|u\|^2_{2,B_r,V} + r^\mu |f|^2_{L^{2,\mu}(R^N)}] \end{aligned}$$

for any $0 < \tau, r \leqq 1$. We can now apply Lemma 3.4: by choosing $\varepsilon = (N-\mu)/2$ we obtain

$$\|u\|_{2,B_{\tau r_0},V} \leqq k[\tau^{(N+\mu)/2}\|u\|^2_{2,B_{r_0},V} + (\tau r_0)^{\mu}|f|^2_{L^{2,\mu}(R^N)}] \tag{4.4}$$

for any $\tau \in]0,1]$ and some $r_0 = r_0(\mu) \in]0,1]$. By assuming $\tau = r/r_0$, estimate (4.4) implies (4.1) for $0 < r \leqq r_0$ and then in general, as (4.1) is trivial if $r_0 \leqq r \leqq 1$.

Proof of Theorem 1.3. From Theorem 1.2 it follows that, if $\operatorname{Re}\lambda \geqq \omega_1$, then the equation $(\lambda - E)u = f \in L^2 \cap L^{2,\mu}(R^N)$, $0 < \mu < N$, has a unique solution $u\, H^2(R^N, V)$. Also, from (4.1) we have

$$|u|_{H^{2,\mu}(R^N,V)} \leqq k[\|u\|_{2,R^N,V} + |f|_{L^{2,\mu}(R^N)}]$$

and so, by (1.8),

$$\|u\|_{H^{2,\mu}(R^N,V)} \leqq k\|f\|_{0,\mu,R^N}. \tag{4.5}$$

Since $\lambda u = f + Eu$, from (4.5) we obtain

$$|\lambda - \omega_1| \cdot \|u\|_{0,\mu,R^N} \leqq k\|f\|_{0,\mu,R^N}. \tag{4.6}$$

The remainder of (1.9) may be derived from (4.5) and (4.6) by standard interpolation techniques (see, e.g. [26, p. 7]).

5. Generation in the topology of uniform convergence. The thesis of Theorem 1.4 will be obtained by compactness arguments once we have proved the following lemma. Here, the main idea is similar to the one contained in [28], [29], [14].

LEMMA 5.1. *Assume* (1.2)-(1.6). *There exists* $\omega_2 \geqq \omega_1$ *such that if* $\operatorname{Re}\lambda \geqq \omega_2$ *and* $u \in H^2(R^N, V) \cap C^1(R^N)$ *is a solution of the equation* $(\lambda - E)u = f \in L^\infty(R^N)$, *then* u *satisfies* (1.10).

Proof. Let $x \in R^N$, $r > 0$ and $\theta \in C^\infty(R^N)$ be such that $0 \leqq \theta \leqq 1$, $\theta \equiv 1$ on $B_{r/2} = B_{r/2}(x)$, $\theta \equiv 0$ out of $B_r = B_r(x)$, $|D^\beta\theta|_{0,\infty} \leqq kr^{-|\beta|}$, $|\beta| \leqq 2$. Then

$$\theta u \in H^2(R^N, V), \qquad (\lambda - E)(\theta u) = F,$$

where

$$F = \theta f - \sum a_{ij}[D_i u D_j\theta + D_j u D_i\theta + u D_i D_j \theta] - \sum b_j u D_j \theta. \tag{5.1}$$

In particular, $F \in L^2(R^N) \cap L^{2,\mu}(R^N)$ for any $0 \leqq \mu \leqq N$ and

$$\begin{aligned} |F|_{L^{2,\mu}(R^N)} &\leqq |f|_{L^{2,\mu}(B_r)} + \frac{k}{r^2}|u|_{L^2,\mu(B_r)} + \frac{k}{r}|u|_{H^{1,\mu}(B_r,V)} \\ &\leqq kr^{(N-\mu)/2}[|f|_{0,\infty,R^N} + r^{-2}|u|_{0,\infty,R^N} + r^{-1}(|u|_{1,\infty,R^N} + |V^{1/2}u|_{0,\infty,R^N})]. \end{aligned} \tag{5.2}$$

Now, let us fix $N-2 < \mu < N$ and apply Theorem 1.3. From (1.9) we obtain, for $\operatorname{Re}\lambda \geqq \omega_1$,

$$\|\theta u\|_{H^{2,\mu}(R^N,V)} + |\lambda - \omega_1|^{1/2}\|\theta u\|_{H^1,\mu(R^N,V)} + |\lambda - \omega_1|\ \|\theta u\|_{0,\mu,R^N} \leqq k\|F\|_{0,\mu,R^N}. \tag{5.3}$$

On the other hand, from known properties of Morrey spaces (see, e.g., [14, Lemma 6.2]) it follows that for any $\varepsilon > 0$ there exists a constant $k(\varepsilon)$, independent of r, such that

$$\begin{aligned} &r^{-2}|\theta u|_{0,\infty,B_r} + [|\theta u|_{1,\infty,B_r} + |V^{1/2}\theta u|_{0,\infty,B_r}] \\ &\quad \leqq \varepsilon r^{(\mu-N)/2}\|\theta u\|_{H^{2,\mu}(B_r,V)} \\ &\quad\quad + k(\varepsilon)r^{(\mu-N-2)/2}\|\theta u\|_{H^{1,\mu}(B_r,V)}. \end{aligned} \tag{5.4}$$

Therefore, for $0<r\leqq 1$, from (5.2), (5.3), (5.4) we conclude that for all $\varepsilon>0$

$$r^{-2}|u|_{0,\infty,B_{r/2}}+r^{-1}[|u|_{1,\infty,B_{r/2}}+|V^{1/2}u|_{0,\infty,B_{r/2}}]$$
$$\leqq k_0\left[\varepsilon+\frac{k(\varepsilon)}{r|\lambda-\omega_1|^{1/2}}\right][|f|_{0,\infty,R^N}+r^{-2}|u|_{0,\infty,R^N}+r^{-1}(|u|_{1,\infty,R^N}+|V^{1/2}u|_{0,\infty,R^N})]$$

where k_0 is independent of ε, r and λ. Now, for any $\varepsilon>0$ there exists $\omega_2(\varepsilon)\geqq\omega_1$ such that $r_\varepsilon=k(\varepsilon)\varepsilon^{-1}|\lambda-\omega_1|^{-1/2}\leqq 1$ for $\operatorname{Re}\lambda\geqq\omega_2(\varepsilon)$. Then

$$\begin{aligned}&r_\varepsilon^{-2}|u|_{0,\infty,B_{r_\varepsilon/2}}+r_\varepsilon^{-1}[|u|_{1,\infty,B_{r_\varepsilon/2}}+|V^{1/2}u|_{0,\infty,B_{r_\varepsilon/2}}]\\ &\qquad\leqq 2k_0\varepsilon[|f|_{0,\infty,R^N}+r_\varepsilon^{-2}|u|_{0,\infty,R^N}+r_\varepsilon^{-1}(|u|_{1,\infty,R^N}+|V^{1/2}u|_{0,\infty,R^N})]\end{aligned}\tag{5.5}$$

and this estimate holds for any point x R^N. But $|u|_{0,\infty,R^N}$ is actually attained at some point of R^N and the same is true for the other norms that appear in the left-hand side of (5.5). So, choosing ε sufficiently small and using (5.5) at most three times, we can easily get (1.10). □

We can now complete the proof of our theorem.

Proof of Theorem 1.4. Let us prove existence first. For each $n=1,2,\cdots$, let $\theta_n\in C^\infty(R^N)$ be such that $0\leqq\theta_n\leqq 1$, $\theta_n\equiv 1$ on $B_n(0)$, $\theta_n\equiv 0$ out of $B_{2n}(0)$ and $|D^\beta\theta_n|\leqq kn^{-|\beta|}$, $|\beta|\leqq 2$. Then $f_n=\theta_nf\in L^{2,\mu}(R^N)$ for any $0\leqq\mu\leqq B$. Fix $N-2<\mu<N$ and let u_n be the solution of the problem

$$u_n\in H^{2,\mu}(R^N,V),\qquad(\lambda-E)u_n=f_n,\operatorname{Re}\lambda\geqq\omega_2.$$

In particular, $u_n\in H^2(R^N)\cap C^1(R^N)$ and so, by Lemma 5.1, $\{u_n\}$ is bounded in $C^1(R^N)$. Therefore, there exists a subsequence $\{u_n^*\}$ which converges to a function $u\in C(R^N)$. uniformly on each compact subset of R^N.

Now, if we show that

$$u_n^*\to u\text{ in }H^2(D)\quad\text{for any }D\subset\subset R^N,\tag{5.6}$$

then standard regularity results and Lemma 5.1 will imply that u is a solution of our equation in the desired class, and that satisfies (1.10). In order to prove (5.6), let $D\subset\subset R^N$ and n_0 be such that $B_{n_0}=B_{n_0}(0)\supset D$. Let us set $u_{nm}=u_n^*-u_m^*$. Since $(\lambda-E)u_{nm}=0$ in B_{n_0+1} for each $n,m>n_0$, we can estimate u_{nm} using the Cacciopoli inequalities proved in [15, § 3]. There exists $r_0\in]0,1[$ such that

$$|u_{nm}|_{2,B_{r_0}(x)}+|u_{nm}|_{1,B_{r_0}(x)}\leqq k|u_{nm}|_{0,B_{r_0}(x)}\tag{5.7}$$

for any $x\in D$. By covering B_{n_0} with a finite number of balls of radius r_0 and adding together inequalities (5.7) we conclude that $\{u_n^*\}$ is a Cauchy sequence in $H^2(B_{n_0})$, which in turn implies (5.6).

Next, the uniqueness of solutions can be proved by similar arguments. Indeed, let $v\in H^2_{\text{loc}}(R^N)\cap C^1(R^N)$ be such that $|V^{1/2}v|_{0,\infty,R^N}<+\infty$ and $(\lambda-E)v=0$. Then $\theta_nv\in H^2(R^N)\cap C^1(R^N)$ and $(\lambda-E)(\theta_nv)=g_n$, where

$$g_n=-\sum_{ij}a_{ij}[D_ivD_j\theta_n+D_jvD_i\theta_n+vD_iD_j\theta_n]-\sum_j b_jvD_j\theta_n.$$

Clearly, $g_n\in L^\infty(R^N)$ and $\lim|g_n|_{0,\infty,R^N}=0$. But, by Lemma 5.1

$$|\lambda-\omega_2|\,|\theta_nv|_{0,\infty,R^N}\leqq k|g_n|_{0,\infty,R^N}\quad\forall n=1,2,\cdots$$

and so $v\equiv 0$.

Remark 5.2. In [28], [29], Stewart proves a generation theorem in the space

$$C_0(R^N)=\{u\in C(R^N)\colon\lim_n|u|_{0,\infty,R^N\setminus B_n(0)}=0\}.$$

The analogue of this result for operators with unbounded coefficients may be easily recovered from Theorem 1.4. Indeed, suppose $f \in C_0(R^N)$, $\operatorname{Re} \lambda \geqq \omega_2$ and let $u \in H^2_{\text{loc}}(R^N) \cap C^1(R^N)$ be the solution of equation $(\lambda - E)u = f$ with $|V^{1/2}u|_{0,\infty,R^N} < +\infty$. Define θ_n, $n = 1, 2, \cdots$, as in the proof of Theorem 1.4 and set $u_n = (1 - \theta_n)u$. Then

$$\begin{aligned}(\lambda - E)u_n &= F_n \\ &= (1-\theta_n)f + \sum_{ij} a_{ij}[D_j\theta_n D_i u + D_i\theta_n D_j u + uD_iD_j\theta_n] + \sum_j b_j u D_j\theta_n.\end{aligned}$$

Therefore, by (1.10) applied to u_n we obtain

$$|\lambda - \omega_2|^{1/2}[|u|_{0,\infty,R^N\setminus B_{2n}} + |V^{1/2}u_{0,\infty,R^N\setminus B_{2n}}] + |\lambda - \omega_2| \cdot |u|_{0,\infty,R^N\setminus B_{2n}} \leqq k|F_n|_{0,\infty,R^N\setminus B_{2n}}.$$

Since $|F_n|_{0,\infty,R^N\setminus B_{2n}} \to 0$ as $n \to \infty$, the last inequality implies that u, $V^{1/2}u$, $D_j u \in C_0(R^N)$.

Remark 5.3. Let us go back to the proof of Lemma 5.1. From (5.2), (5.3) we obtain

$$\|u\|_{H^{2,\mu}(B_{r/2},V)} \leqq kr^{(N-\mu)/2}[|f|_{0,\infty,R^N} + r^{-2}|u|_{0,\infty,R^N} + r^{-1}(|u|_{1,\infty,R^N} + |V^{1/2}u|_{0,\infty,R^N})]$$

for $0 < r < 1$. Therefore, choosing r as in the sequel of that proof, we conclude that

$$(5.8) \qquad |\lambda - \omega_2|^{(N-\mu)/4} \sup_{x \in R^N} \|u\|_{H^{2,\mu}(B_{r/2}(x),V)} \leqq k|f|_{0,\infty,R^N}.$$

See also [28], [29] for a similar estimate involving L^p norms.

6. Generation in the Hölder topology. The estimate contained in the following lemma is essential for the proof of Theorem 1.5. We will briefly sketch its proof, which uses the techniques of [12].

LEMMA 6.1. *Assume* (1.2)-(1.5), (1.11). *If* $u \in H^2_{\text{loc}}(R^N)$ *is a solution of the equation* $(\lambda - E)u = f \in C^\alpha(R^N)$, $0 < \alpha < 1$, $\operatorname{Re} \lambda \geqq V_0$, *then*

$$(6.1) \qquad \|u\|_{C^{2,\alpha}(B_1(x^0),V)} \leqq k[\|u\|_{2,B_2(x^0),V} + \|f\|_{\alpha,R^N}]$$

for any $x^0 \in R^n$ *and some constant k independent of* λ.

Proof. We note first that (6.1) will be proved provided we show that

$$(6.2) \qquad \sum_{|\beta| \leqq 2} \int_{B_r(x)} V^{2-|\beta|}|D^\beta u - (D^\beta u)_{B_r(x)}|^2 \, dy \leqq kr^{N+2\alpha}[\|u\|_{2,B_2(x^0),V)} + \|f\|_{\alpha,R^N}$$

for any $x \in B_1(x^0)$ and $0 < r \leqq 1$ (here g_{B_r} denotes the integral mean value of g on B_r). Indeed, from (1.7), (6.2) and (3.9) we obtain

$$\sum_{|\beta| \leqq 2} V^{2-|\beta|}(x^0)[D^\beta u]_{\alpha,B_1(x^0)} \leqq k[\|u\|_{2,B_2(x^0),V} + \|f\|_{\alpha,R^N}]$$

which in turn implies (6.1), recalling (1.3) and (1.7).

Also, since $f \in L^{2,\mu}(R^N)$ for any $0 \leqq \mu \leqq N$, from Lemma 4.1 and (1.11) it follows that

$$(6.3) \qquad \|u\|^2_{2,B_r(x),V} \leqq kr^{N-\alpha}[\|u\|^2_{2,B_1(x),V} + |f|^2_{0,\infty,R^N}]$$

for any $x \in R^N$ and $0 < r \leqq 1$. Here and in the sequel, k denotes a constant independent of λ.

Now, for $x \in B_1(x^0)$ and $0 < r \leqq 1$, let $B_r = B_r(x)$ and $w \in H^2(B_r)$ be the solution of the Dirichlet problem

$$\begin{aligned}(\lambda - E^0)w &= f - f_{B_r} + (E - E^0)u \quad &&\text{in } B_r, \\ w &= 0 &&\text{on } \partial B_r\end{aligned}$$

where E^0 is defined as in (3.1). From (1.3), (1.11) and (6.2) it follows that

$$|(E - E^0)u|^2_{0,B_r} \leqq kr^{2\alpha}\|u\|^2_{2,B_r,V} \leqq kr^{N+\alpha}[\|u\|^2_{2,B_1,V} + |f|^2_{0,\infty,R^N}]$$

so, by (3.2),

$$\|w\|^2_{2,B_r,V} \leqq k\{r^{N+2\alpha}[f]^2_{\alpha,R^N} + |(E-E^0)u|^2_{0,B_r}\} \\ \leqq kr^{N+\alpha}\{\|u\|^2_{2,B_1,V} + \|f\|^2_{\alpha,R^N}\}. \tag{6.4}$$

On the other hand, the difference $v=u-w$, together with the partial derivatives $D^\beta v$, $|\beta|\leqq 2$, satisfies the assumptions of Lemma 3.2. Then, by (3.6) and by the properties (1.3), (1.7) of V, we conclude that

$$\sum_{|\beta|\leqq 2}\int_{B_{\tau r}} V^{2-|\beta|}D^\beta v-(D^\beta v)_{B_{\tau r}}|^2\,dy \leqq k\tau^{N+2}\sum_{|\beta|\leqq 2}\int_{B_r} V^{2-|\beta|}|D^\beta v-(D^\beta v)_{B_r}|^2\,dy \tag{6.5}$$

for any $\tau\in[0,1]$. Since $u=v+w$, (6.4) and (6.5) yield

$$\sum_{|\beta|\le 2}\int_{B_{\tau r}} V^{2-|\beta|}|D^\beta u-(D^\beta u)_{B_{\tau r}}|^2\,dy \\ \leqq k\tau^{N+2}\sum_{|\beta|\leqq r}\int_{B_r} V^{2-|\beta|}|D^\beta u-(D^\beta u)_{B_r}|^2\,dy \\ + kr^{N+\alpha}[\|u\|^2_{2,B_1,V} + \|f\|^2_{\alpha,R^N}] \tag{6.6}$$

for any $\tau\in\,]0,1]$. So, applying Lemma 3.4, we obtain (6.2) with exponent $N+\alpha$ instead of $N+2\alpha$. Therefore, by our introductory remarks, (6.1) holds with exponent $\alpha/2$ instead of α. Now, inserting this information in the above proof, we obtain (6.3) with exponent N instead of $N-\alpha$, (6.6) with exponent $N+2\alpha$ instead of $N+\alpha$ and finally (6.2). □

Proof of Theorem 1.5. From Theorem 1.4 it follows that, for Re $\lambda\geqq\omega_2$, the equation $(\lambda-E)u=f\in C^\alpha(R^N)$ has a unique solution $u\in H^2_{\text{loc}}(R^N)\cap C^1(R^N)$ such that $|V^{1/2}u|_{0,\infty,R^N}<+\infty$. Furthermore, u satisfies (1.10). Therefore, Lemma 6.1 will allow us to conclude that $u\in C^{2,\alpha}(R^N,V)$ and

$$\|u\|_{C^{2,\alpha}(R^N,V)}\leqq k\|f\|_{\alpha,R^N} \tag{6.7}$$

provided we show that

$$\|u\|_{2,B_2(x^0),V}\leqq k\|f\|_{\alpha,R^N} \tag{6.8}$$

for any point $x^0\in R^N$.

In view of this, let us localize our equation by a cutoff function $\theta\in C^\infty(R^N)$ such that $\theta\equiv 1$ on $B_2(x^0)$ and $\theta\equiv 0$ out of $B_4(x^0)$. Then $\theta u\in H^2(R^N,V)$ and $(\lambda-E)(\theta u)=F$, where F is defined as in (5.1). Now, from (1.10) we obtain

$$|F|_{0,R^N}\leqq k|f|_{0,\infty,R^N} \tag{6.9}$$

and so (6.8) follows from (6.9) and (1.8) applied to θu.

Finally, the full estimate (1.12) may be easily recovered from (6.7). □

7. Application to parabolic equations. Let us consider the initial value problem

$$\frac{\partial u}{\partial t}(t,x)-E(t)u(t,x)=f(t,x),\ (t,x)\in[0,T]\times R^N, \\ u(0,x)=u_0(x) \tag{7.1}$$

where

$$E(t)=\sum_{ij=1}^N a_{ij}(t,x)D_iD_j+\sum_{j=1}^N b_j(t,x)D_j-c(t,x) \tag{7.2}$$

is a differential operator with unbounded coefficients. We have already recalled some of the literature on this subject. Recently, motivation for studying these problems has also been provided by nonlinear filtering for diffusion processes (see, e.g., [19]). Indeed, in the case of unbounded observations, the pathwise treatment of the Zakai equation leads to parabolic operators with unbounded lower order coefficients (see [5], [20], [18], [27], [16]).

We analyze (7.1) as an abstract Cauchy problem in a Banach space, using some of the semigroup generation theorems proved in the previous sections.

If X is a Banach space with norm $\| \|_x$, we denote by $L(X)$ the space of bounded linear operators from X into X and by $C([T_0, T_1]; X)$ the space of continuous functions $g:[T_0, T_1] \to X$. We also define, for $0<\alpha<1$,

$$C^\alpha(T_0, T_1]; X)=\left\{u\in C([T_0, T_1]; X): \sup_{\substack{T_0\leqq s,t\leqq T_1\\ s\neq t}} \frac{\|u(t)-u(s)\|_x}{|t-s|^\alpha}<+\infty\right\}.$$

The spaces $C^1([T_0, T_1]; X)$, $C^{1,\alpha}([T_0, T_1]; X)$ are defined similarly. Moreover, we set

$$C(]T_0, T_1]; X)=\bigcap_{0<\varepsilon<T_1-T_0} C([T_0+\varepsilon, T_1]; X)$$

and we give analogous definitions for the spaces $C^\alpha(]T_0, T_1]; X)$, $C^1(]T_0, T_1]; X)$ and $C^{1,\alpha}(]T_0, T_1]; X)$.

Let V be a function having properties (1.3) and suppose that $E(t)$ is an operator of type (7.2), with complex-valued coefficients satisfying the assumptions of § 1, uniformly for $t\in[0, T]$, i.e., for all $(t, x)\in[0, T]\times R^N$

$$\operatorname{Re}\sum_{ij} a_{ij}(t, x)z_j\bar{z}_i\geqq \nu\|z\|^2 \quad \forall z\in C^N, \tag{7.3}$$

$$K_1V(x)-V_0\leqq \operatorname{Re} c(t, x)\leqq K_2V(x)+V_0, \tag{7.4}$$

$$\sum |b_j(t, x)|^2\leqq \nu B^2K_1V(x) \tag{7.5}$$

for some constants $\nu, K_1, K_2, V_0>0$ and $B\in[0,2[$. Also, suppose that, for some $\alpha\in]0,1[$,

$$a_{ij}, b_j/V^{1/2}, c/V\in C^\alpha([0, T]; C(R^N)), \qquad i,j=1,\cdots, N. \tag{7.6}$$

Notice that the structure hypotheses (7.3), $\cdots$, (7.6) are similar to those of [27].

The theorem below treats the existence, uniqueness and regularity of solutions to problem (7.1) with respect to the L^2 topology.

THEOREM 7.1. *Assume* (7.3)-(7.6) *and let* $f\in C^\alpha([0, T]; L^2)(R^N))$. *Then*

(i) *For any* $u_0\in L^2(R^N)$ *problem* (7.1) *has a unique solution*

$$u\in C([0, T]; L^2(R^N))\cap C^1(]0, T]; L^2(R^N))\cap C(]0, R]; H^2(R^N, V));$$

(ii) *For any* $u_0\in H^2(R^N, V)$ *problem* (7.1) *has a unique solution*

$$u\in C^1([0, T]; L^2(R^N))\cap C([0, T]; H^2(R^N, V)).$$

Moreover, in both cases (i) *and* (ii),

$$u\in C^{1,\alpha}(]0, T]; L^2(R^N))\cap C^\alpha(]0, T]; H^2(R^N, V)). \tag{7.7}$$

Proof. Theorem 1.2 implies that $E(T)$, $t\in[0, T]$, generates an analytic semigroup on $L^2(R^N)$ with domain $H^2(R^N, V)$. Furthermore, by (1.8)

$$\|[\lambda-E(t)]^{-1}\|_{\mathscr{L}(L^2(R^N))}\leqq\frac{k}{|\lambda-\omega_1|}$$

for any $\lambda \in C$ with $\operatorname{Re} \lambda > \omega_1$. Also, from (7.6)

$$\|[E(t)-E(s)][\omega_1 - E(\tau)]^{-1}\|_{\mathscr{L}(L^2(R^N))} \leqq k|t-s|^\alpha$$

for any $\tau, s, t \in [0, T]$. Therefore, assertions (i), (ii) and (7.7) follow from classical results on nonautonomous evolution equations (see, e.g. [26]; see also [17] and [1]). □

Remark 7.2. Recalling Remark 2.2, one can easily extend the result of Theorem 7.1 to the weighted space $L^2_\pi(R^N)$, provided that $\pi = \log \Pi$ satisfies (2.17).

We will now derive the analogue of Theorem 7.1 for Hölder spaces. According to the assumptions of Theorem 1.5, we will replace (7.6) by

$$(7.8) \qquad a_{ij},\ b_j/V^{1/2},\ c/V \in C^\alpha([0, T]; C^\beta(R^N)), \qquad i, j = 1, \cdots, N$$

for some $\alpha, \beta \in]0, 1[$. Then arguing as in the proof of Theorem 7.1 we obtain the following result.

THEOREM 7.3. *Assume* (7.3), (7.4), (7.5), (7.8) *and let* $f \in C^\alpha([0, T]; C^\beta(R^N))$.

(i) *If* $u_0 \in h^\beta(R^N, V)$, *then problem* (7.1) *has a unique solution*

$$u \in C([0, T]; C^\beta(R^N)) \cap C^1(]0, T]; C^\beta(R^N)) \cap C^{2,\beta}(]0, T]; C^{2,\beta}(R^N, V)).$$

(ii) *If* $u_0 \in C^{2,\beta}(R^N, V)$ *and* $E(0)+f(\cdot, 0) \in h^\beta(R^N, V)$, *then problem* (7.1) *has a unique solution*:

$$u \in C^1([0, T]; C^\beta(R^N)) \cap C([0, T]; C^{2,\beta}(R^N, V)).$$

Moreover, in both cases (i) *and* (ii),

$$u \in C^{1,\alpha}(]0, T]; C^\beta(R^N)) \cap C^\alpha(]0, T]; C^{2,\beta}(R^N, V)).$$

Remark 7.4. Suppose that a_{ij}, b_j, c $(i, j = 1, \cdots, N)$ satisfy (7.6) and $a_{ij}, b_j, c \in C([0, T]; C^\beta(R^N))$. Then (7.8) fails to be fulfilled, in general. However, (7.8) holds under any of the following assumptions:

$$(7.9) \qquad \begin{array}{l} a_{ij},\ b_j,\ c \text{ are differentiable with respect to } x \text{ and} \\ D_h a_{ij},\ D_h b_j/V^{1/2},\ D_h c/V \in C^\alpha([0, T]; C(R^N)), \qquad i, j, h = 1, \cdots, N; \end{array}$$

$$(7.10) \qquad \begin{array}{l} a_{ij},\ b_j,\ c \text{ are differentiable with respect to } t \text{ and} \\ \dfrac{\partial a_{ij}}{\partial t}, \dfrac{\partial b_j}{\partial t}\Big/ V^{1/2}, \dfrac{\partial c}{\partial t}\Big/ V \in C([0, T]; C^\beta(R^N)), \qquad i, j = 1, \cdots, N. \end{array}$$

REFERENCES

[1] P. ACQUISTAPACE AND B. TERRENI, *On the abstract non-autonomous Cauchy problem in the case of constant domains*, Ann. Mat. Pura Appl., 140 (1985), pp. 1-55.

[2] S. AGMON, *On the eigenfunctions and on the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119-147.

[3] D. G. ARONSON AND P. BESALA, *Uniqueness of solutions of the Cauchy problem for parabolic equations*, J. Math. Anal. Appl., 13 (1966), pp. 516-526.

[4] ———, *Parabolic equations with unbounded coefficients*, J. Differential Equations, 3 (1967), pp. 1-14.

[5] J. S. BARAS, G. O. BLANKENSHIP AND W. E. HOPKINS, JR., *Existence, uniqueness and asymptotic behaviour of solutions to a class of Zakai equations with unbounded coefficients*, IEEE Trans. Automat. Control, 28 (1983), pp. 203-214.

[6] A. BENSOUSSAN AND J. L. LIONS, *Applications of variational inequalities in stochastic control*, North-Holland, Amsterdam, 1982.

[7] P. BESALA, *On the existence of a fundamental solution for a parabolic differential equation with unbounded coefficients*, Ann. Polon. Math., 29 (1975), pp. 403-409.

[8] W. BODANKO, *Sur le problème de Cauchy et les problèmes de Fourier pour les équations paraboliques dans un domaine non borne*, Ann. Polon. Math., 18 (1966), pp. 79-94.

[9] F. E. BROWDER, *On the spectral theory of elliptic differential operators*, I, Math. Ann., 142 (1961), pp. 22-130.

[10] S. CAMPANATO, *Proprietà di holderianità di alcune classi di funzioni*, Ann. Scuola Norm. Sup. Pisa, 17 (1963), pp. 175-188.

[11] ———, *Equazioni ellittiche non variazionali a coefficient continui*, Ann. Math. Pura Appl., 86 (1970), pp. 125-154.

[12] ———, *Generation of analytic semigroups by elliptic operators of second order in Hölder spaces*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 8 (1981), pp. 495-512.

[13] P. CANNARSA, *Analytic semigroups generated by strongly elliptic operators of second order in R^N*, preprint n.53, Accademia Navale, Livorno, 1984.

[14] P. CANNARSA, B. TERRENI AND V. VESPRI, *Analytic semigroups generated by non-variational elliptic systems of second order under Dirichlet boundary conditions*, J. Math. Anal. Appl., 112 (1985), pp. 56-103.

[15] P. CANNARSA AND V. VESPRI, *Analytic semigroups generated on Hölder spaces by second order elliptic systems under Dirichlet boundary conditions*, Ann. Mat. Pura Appl., 140 (1985), pp. 393-415.

[16] ———, *Existence and uniqueness of solutions to a class of stochastic partial differential equations*, Stochastic Anal. Appl., 3 (1985), pp. 315-339.

[17] G. DA PRATO AND E. SINESTRARI, *Holder regularity for non-autonomous abstract parabolic equations*, Israel J. Math., 42 (1982), pp. 1-19.

[18] G. DA PRATO, M. IANNELLI AND L. TUBARO, *An existence theorem for a stochastic partial differential equation arising from filtering theory*, Rend. Sem. Mat. Univ. Padova, 71 (1983), pp. 217-222.

[19] M. H. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering*, NATO Advanced Study Institute Series, Reidel, Dordrecht, June-July 1980.

[20] W. H. FLEMING AND S. K. MITTER, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 8 (1982), pp. 63-77.

[21] R. S. FREEMAN AND M. SCHECHTER, *On the existence, uniqueness and regularity of solutions to general elliptic boundary value problems*, J. Differential Equations, 15 (1974), pp. 213-246.

[22] Y. HIGOUCHI, *A priori estimates and existence theorems on elliptic boundary value problems for unbounded domains*, Osaka J. Math., 5 (1968), pp. 103-135.

[23] S. ITÔ, *Fundamental solutions of parabolic differential equations and boundary value problems*, Japan J. Math., 27 (1957), pp. 55-102.

[24] M. KRZYZANSKI AND A. SZYBIAK, *Construction et etude de la solution fondamentale de l'équation lineaire du type parabolique dont le dernier coefficient est non borne*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. (8), 27 (1959), pp. 1-10.

[25] A. LUNDARDI, *Interpolation spaces between domains of elliptic operators and spaces of continuous functions with applications to nonlinear parabolic equations*, Math. Nachr., to appear.

[26] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[27] S. J. SHEU, *Solution of certain parabolic equations with unbounded coefficients and its application to nonlinear filtering*, Stochastics, 10 (1983), pp. 31-46.

[28] H. B. STEWART, *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 188 (1974), pp. 141-162.

[29] ———, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299-310.

[30] V. VESPRI, *Generation of analytic semigroups in the uniform and L^p topology by elliptic operators in R^n*, preprint n.18, Dipartimento di Matematica, II Universita degli Studi di Roma, 1984.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

A DIFFERENTIAL-DIFFERENCE EQUATION*

JERRY L. FIELDS† AND WALEED A. AL-SALAM†

Abstract. Wimp showed that the hypergeometric polynomials

$$P_n(z) = {}_{p+2}F_{p+1}(-n,\ n+\lambda,\ a_{\mathrm{p}};\ b_{\mathrm{p+1}};\ z), \qquad n=0,1,\cdots,$$

satisfied a certain differential-difference equation. Here we show that all "common" solutions to the standard differential equation and the standard difference equation satisfied by the $P_n(z)$ also satisfy the above mentioned differential-difference equation.

Key words. hypergeometric functions, difference equations, difference-differential equations

AMS(MOS) subject classification. 33A35

1. Notation and introduction.

Let $\mathbf{Z}$ denote the integers, and $\mathbf{Z}^+$ the positive integers. In addition to the usual notation for hypergeometric functions and Meijer G-functions, see [1], [6], we will use the abbreviated notation

$$\Gamma_n(s+c_{\mathbf{P}}) := \prod_{k=n+1}^{p} \Gamma(s+c_k), \qquad \Gamma(s+c_{\mathbf{Q}}) := \Gamma_0(s+c_{\mathbf{Q}}),$$

$$(s)_t := \frac{\Gamma(s+t)}{\Gamma(s)}, \qquad (s+c_{\mathbf{Q}})_t := \prod_{k=1}^{q}(s+c_k)_t,$$

$${}_pF_q(z) := {}_pF_q\left(\begin{matrix} a_{\mathbf{P}} \\ b_{\mathbf{Q}} \end{matrix}\middle|\, z\right) := {}_pF_q\left(\begin{matrix} a_1, \cdots, a_p \\ b_1, \cdots, b_q \end{matrix}\middle|\, z\right)$$

$$:= \sum_{k=0}^{\infty} \frac{(a_{\mathbf{P}})_k}{(b_{\mathbf{Q}})_k}\frac{z^k}{k!},$$

$$G^{m,n}_{p,q}(w) := G^{m,n}_{p,q}\left(w\,\middle|\,\begin{matrix} a_{\mathbf{P}} \\ b_{\mathbf{Q}} \end{matrix}\right) := G^{m,n}_{p,q}\left(w\,\middle|\,\begin{matrix} a_1, \cdots, a_p \\ b_1, \cdots, b_q \end{matrix}\right)$$

$$:= \frac{1}{2\pi i}\int_L \frac{\Gamma(b_{\mathbf{M}}-t)\Gamma(1-a_{\mathbf{N}}+t)w^t}{\Gamma_m(1-b_{\mathbf{Q}}+t)\Gamma_n(a_{\mathbf{P}}-t)\,dt},$$

where L is an upward oriented contour which separates the poles of $\Gamma(b_{\mathbf{M}}-t)$ from those of $\Gamma(1-a_{\mathbf{N}}+t)$, and which runs from $-i\infty$ to $+i\infty$ ($L=L_0$), or begins and ends at $+\infty$ ($L=L_+$), or $-\infty$ ($L=L_-$). In particular, $(a_{\mathbf{Q}})_1 = a_1 \cdot a_2 \cdots a_q$. As hypergeometric functions, the polynomials

$$P_n(z) := {}_{p+2}F_{p+1}\left(\begin{matrix} -n,\ n+\lambda,\ a_{\mathbf{P}} \\ b_{\mathbf{P+1}} \end{matrix}\middle|\, z\right), \qquad n+1 \in \mathbf{Z}^+,$$

$$1-b_j \notin \mathbf{Z}^+, \qquad j=1,\cdots,p+1,$$

satisfy a standard differential equation [1], [6] of the form

$$\begin{aligned} \mathscr{L}_z\{Y_n(z)\} &= \{\delta(\delta+b_{\mathbf{P+1}}-1)_1 - z(\delta-n)(\delta+n+\lambda)(\delta+a_{\mathbf{P}})_1\}Y_n(z) \\ &= 0, \qquad \delta = z\frac{d}{dz}. \end{aligned} \tag{1}$$

* Received by the editors March 17, 1986; accepted for publication (in revised form) May 8, 1986. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under grants A7549 and A2975.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

Under suitable parameter restrictions for $n \geqq p+2$, these polynomials also satisfy a standard difference equation [2], [6] of the form

$$(2) \qquad \mathcal{M}_n\{Y_n(z)\} = \sum_{m=0}^{p+2} (A_m^* + zB_m^*) Y_{n-m}(z) = 0, \quad A_0^* = 1, \quad B_0^* = B_{p+2}^* = 0,$$

where the numbers A_m^*, B_m^* are rational in n and independent of z. In [7], Wimp showed that the $P_n(z)$ also satisfy a differential-difference equation

$$(3) \qquad z(1-z)\frac{dY_n(z)}{dz} = \sum_{m=0}^{p+1} (A_m + zB_m) Y_{n-m}(z), \qquad B_{p+1} = 0,$$

where the numbers A_m, B_m are rational in n, and independent of z. When $z=1$ and $Y_n(z) = P_n(z)$, (3) gives a $p+2$ term difference equation for $P_n(1)$, which is one term less than the $p+3$ term equation given by (2). We will show that any "common" solution of (1) and (2) is also a solution of (3). We will now define what we mean by a "common" solution of (1) and (2).

In [3], it was shown that under suitable parameter restrictions, (1) and (2) have a common global basis in

$$\mathcal{D} = \{z\colon |\arg z| < \pi, |\arg(1-z)| < \pi\},$$

i.e., if n is sufficiently large, and

$$\begin{aligned} &a_k - a_r \notin \mathbf{Z}, \quad k \neq r, \\ &1 - b_j \notin \mathbf{Z}^+, \qquad\qquad k, r = 1, \cdots, p, \quad j = 1, \cdots, p+1, \\ &a_k \neq b_j, \\ &\mathcal{B} = \{G_n(z\,e^{i\pi}), G_n(z\,e^{-i\pi}), L_{n,j}(z), j = 1, \cdots, p+1\}, \end{aligned}$$

$$L_{n,j}(z) = \frac{\Gamma(n+1)}{\Gamma(n+\lambda)} G_{p+3,p+3}^{p+3,2}\left(z \,\middle|\, \begin{matrix} 1-n-\lambda, 1-a_j, 1-a_{\mathbf{P}}, n+1 \\ 0, 1-b_{\mathbf{P+1}}, 1-a_j \end{matrix}\right),$$

$$G_n(z) = \frac{\Gamma(n+1)}{\Gamma(n+\lambda)} G_{p+2,p+2}^{p+2,1}\left(z \,\middle|\, \begin{matrix} 1-n-\lambda, 1-a_{\mathbf{P}}, n+1 \\ 0, 1-b_{\mathbf{P+1}} \end{matrix}\right),$$

where $G_n(z\,e^{i\pi})$ is the analytic continuation of $G_n(z\,e^{-i\pi})$ along an arc which encloses $z=0$, but not $z=1$, then the elements of $\mathcal{B}$ are solutions of both (1) and (2) in $\mathcal{D}$, and are linearly independent both as functions of z, and of n.

Definition. The function $Y_n(z)$ is a common solution to $\mathcal{L}_z\{Y_n(z)\} = 0$ and $\mathcal{M}_n\{Y_n(z)\} = 0$ in $\mathcal{D}$, provided $Y_n(z)$ has a representation

$$Y_n(z) = D_1 G_n(z\,e^{i\pi}) + D_2 G_n(z\,e^{-i\pi}) + \sum_{j=1}^{p} C_j L_{n,j}(z),$$

where the constants $D_1, D_2, C_1, \cdots, C_p$ are independent of z and are periodic functions of n whose period is 1.

In [3] it is shown that $P_n(z)$ is such a common solution.

The following lemma is central to our analysis.

Lemma. *If $S_r(w)$ is a polynomial in w of degree r, and t is any integer $\geqq r$, then $S_r(w)$ can be represented uniquely in the form*

$$(4) \qquad S_r(w) = \sum_{m=0}^{t} Q_m (w+\gamma)_m (w+\gamma+\varepsilon)_{t-m},$$

$$
(5)\qquad
\begin{aligned}
Q_m &= \frac{(t+\varepsilon-2m)}{m!(\varepsilon)_{t+1-m}} \sum_{k=0}^{m} \frac{(-m)_k(m-\varepsilon-t)_k}{k!(1-\varepsilon)_k} S_r(-\gamma-k)\\
&= \frac{(-t-\varepsilon+2m)}{(t-m)!(-\varepsilon)_{1+m}} \sum_{k=0}^{t-m} \frac{(m-t)_k(-m+\varepsilon)_k}{k!(1+\varepsilon)_k} S_r(-\gamma-\varepsilon-k),
\end{aligned}
$$

where γ, ε are arbitrary constants, and $\varepsilon \neq 0, \pm 1, \cdots, \pm(t-1)$.

Proof. In a slightly different form, this lemma occurs in [2]. The uniqueness of the Q_m's follows immediately, as substituting the values $w=-\gamma-j, j=0, 1, \cdots, t$ into $S_r(w)$ leads one to a nonsingular, triangular system of linear equations in the Q_m's. Thus, if the Q_m's exist at all, they are unique.

For the existence of the Q_m, consider the Lagrange representation

$$
(6)\qquad S_r(w) = \sum_{k=0}^{t} \frac{(-1)^k(w+\gamma)_t(w+\gamma+t)}{k!(t-k)!(w+\gamma+k)} S_r(-\gamma-k).
$$

This follows from the fact that

$$
\frac{(w+\gamma)_t(w+\gamma+t)}{w+\gamma+k} = (w+\gamma)_k(w+\gamma+k+1)_{t-k},
$$

so that the right-hand side of (6) is a polynomial in w of degree t. When this polynomial is evaluated at the points $w=-\gamma-j, j=0, 1, \cdots, t$, it agrees with $S_r(w)$ at these points. As $t \geqq r$, this is sufficient to establish (6). By considering the partial fraction decomposition in w, one can show

$$
{}_4F_3\left(\begin{matrix} k-t,\ w+\gamma+\varepsilon,\ 1+\dfrac{\varepsilon-t}{2},\ 1 \\ 1+\varepsilon-k,\ 1-w+\gamma-t,\ \dfrac{\varepsilon-t}{2} \end{matrix}\,\middle|\, 1\right) = \frac{(\varepsilon-k)(w+\gamma+t)}{(\varepsilon-t)(w+\gamma+k)},
$$

$$
1+t-k \in \mathbf{Z}^+, \qquad (\varepsilon-t)(w+\gamma+k) \neq 0.
$$

It is worth noting that this formula also follows from a special case of a limiting form of Dougall's Theorem [1, p. 191, formula 6], i.e. set $d=1$ in

$$
{}_5F_4\left(\begin{matrix} a,\ \dfrac{a}{2}+1,\ b,\ c,\ d \\ \dfrac{a}{2},\ a+1-b,\ a+1-c,\ a+1-d \end{matrix}\,\middle|\, 1\right)
$$
$$
= \frac{\Gamma(a+1-b)\Gamma(a+1-c)\Gamma(a+1-d)\Gamma(a+1-b-c-d)}{\Gamma(a+1)\Gamma(a+1-c-d)\Gamma(a+1-b-d)\Gamma(a+1-b-c)}.
$$

Substituting this expression for $(w+\gamma+t)/(w+\gamma+k)$ into (6), expressing the ${}_4F_3$ as a sum and interchanging summation processes, one obtains the first line of (5). To obtain the second line of (5), it is sufficient to observe that

$$
S_r(w) = \sum_{m=0}^{t} Q_{t-m}(w+\gamma^*)_m(w+\gamma^*+\varepsilon^*)_{t-m},
$$

$$
\gamma^* = \gamma+\varepsilon, \qquad \varepsilon^* = -\varepsilon.
$$

Representing Q_{t-m} in the form given by the first line of (5), and then replacing m by $t-m$, we obtain the second line of (5).

Remark. Applying this lemma to the relationship

$$T(w) := \sum_{m=0}^{p+2} \{A_m^* + zB_m^*\} \frac{(w-n)_m(w+n+\lambda-p-2)_{p+2-m}}{(-n)_m(n+\lambda-p-2)_{p+2-m}}$$

$$= \frac{(n+\lambda)_n}{(n+\lambda-p-2)_n(n+b_{P+2}-1)_1}$$

$$\cdot\{(w+b_{P+2}-1)_1 - z(w-n)(w+n+\lambda-p-2)(w+a_P)_1\},$$

where $b_{p+2}=1$ and $S_r(w)$ is the right-hand side of this relation, determines the numbers A_m^*, B_m^* as convenient closed form expressions. If the A_m^* and B_m^* in (2) are the numbers determined above, and we substitute in the left-hand side of (2) the series expression for $P_n(z)$, interchange the summation processes, we get an equation of the form

$$\sum_{j=0}^{n} \frac{(a_P)_j(n+\lambda-p-2)_j(-z)^j}{(b_{P+1})_j(n+1)_{-j}} T(j).$$

Substituting in this expression the polynomial form of $T(j)$ and rearranging in powers of z, we see that the resulting expression is zero, and hence equation (2) is satisfied.

Similarly, numbers A_m, B_m are determined by

$$\sum_{m=0}^{p+1} \{A_m + zB_m\} \frac{(w-n)_m(w+n+\lambda-p-1)_{p+1-m}}{(-n)_m(n+\lambda-p-1)_{p+1-m}}$$

$$(7) \qquad = \frac{1}{(n+\lambda-p-1)_{p+1}}\{w(w+n+\lambda-p-1)_{p+1} - w(w+b_{P+1}-1)_1$$

$$+ z(-w(w+n+\lambda-p-1)_{p+1}$$

$$+ (w-n)(w+n+\lambda-p-1)(w+a_P)_1)\}.$$

In our main result, we will show that with this determination of A_m, B_m, (3) is satisfied.

2. Main result.

THEOREM. *If $Y_n(z)$ is a common solution to*

$$\mathscr{L}_z\{Y_n(z)\} = 0 \quad \textit{and} \quad \mathscr{M}_n\{Y_n(z)\} = 0$$

in $\mathscr{D}$, then $Y_n(z)$ also satisfies the differential-difference equation

$$(8) \qquad z(1-z)\frac{dY_n(z)}{dz} = \sum_{m=0}^{p+1} (A_m + zB_m) Y_{n-m}(z),$$

where the A_m, B_m are defined in (7).

Proof. In view of the linearity of the various operators and the fact that $Y_n(z)$ can be written as a linear combination of the elements in $\mathscr{B}$, it is sufficient to show that $G_n(z\,e^{\pm i\pi})$ and $L_{n,j}(z)$, $j = 1, \cdots, p+1$, satisfy (8).

For convenience, let $\sigma = p+1$. We will find use for the relationship, s general,

$$\sum_{m=0}^{\sigma} \{A_m + zB_m\} \frac{(n-m+\lambda)_s}{(n-m+1)-s}$$

$$(9) \qquad = (1-z)_s \frac{(n+\lambda)_s}{(n+1)_{-s}} - \frac{(n+\lambda)_{s-\sigma}}{(n+1)_{-s}} s(s-1+b_{P+1})_1 - z\frac{(n+\lambda)_{s+1-\sigma}}{(n+1)_{-s-1}}(s+a_P)_1.$$

To see that (9) holds, let H denote the left-hand side of (9). Using

$$\frac{(n-m+\lambda)_s}{(n-m+1)_{-s}} = \frac{(n+\lambda-\sigma)_s}{(n+1)_{-s}} \frac{(s-n)_m(s+n+\lambda-\sigma)_{\sigma-m}}{(-n)_m(n+\lambda-\sigma)_{\sigma-m}},$$

H becomes a form to which (7) is applicable, and after some algebra, the right-hand side of (9) occurs. Using the Mellin-Barnes representation

$$G_{n-m}(z\,e^{\pm i\pi})=\frac{1}{2\pi i}\int_L \Omega(s)\frac{(n-m+\lambda)_s}{(n-m+1)_{-s}}(z\,e^{\pm i\pi})^s\,ds,$$

$$\Omega(s):=\frac{\Gamma(-s)\Gamma(1-b_{\mathrm{P}+1}-s)}{\Gamma(1-a_{\mathrm{P}}-s)},$$

where the integration contour L can be chosen to be a valid contour for n fixed and $m=0,1,\cdots,\sigma$, we have

$$\begin{aligned}K&:=\sum_{m=0}^{\sigma}(A_m+zB_m)G_{n-m}(z\,e^{\pm i\pi})\\&=\frac{1}{2\pi i}\int_L \Omega(s)(z\,e^{\pm i\pi})^s\sum_{m=0}^{\sigma}(A_m+zB_m)\frac{(n-m+\lambda)_s}{(n-m+1)_{-s}}\,ds.\end{aligned}$$

After substituting (9) into this expression and doing some algebra, we get

$$\begin{aligned}K&=\frac{z(1-z)}{2\pi i}\int_L \Omega(s)\frac{(n+\lambda)_s}{(n+1)_{-s}}\frac{d}{dz}\{(z\,e^{\pm i\pi})^s\}\,ds\\&\quad+\frac{(-1)^p}{2\pi i}\left\{\int_L-\int_{L-1}\right\}\Omega^*(s)\frac{(n+\lambda)_{s+1-\sigma}}{(n+1)_{-s-1}}(z\,e^{\pm i\pi})^{s+1}\,ds\\&=z(1-z)\frac{d}{dz}G_n(z\,e^{\pm i\pi})+R,\end{aligned}$$

$$\Omega^*(s)=\frac{\Gamma(-s)\Gamma(1-b_{\mathrm{P}+1}-s)}{\Gamma(-a_{\mathrm{P}}-s)},$$

$$R=\sum_{\substack{\text{Residue}\\ s\text{ between}\\ L\text{ and }L-1}}\left\{(-1)^p\Omega^*(s)\frac{(n+\lambda)_{s+1-\sigma}}{(n+1)_{-s-1}}(z\,e^{\pm i\pi})^{s+1}\right\}.$$

By inspection, $R=0$, and $G_n(z\,e^{\pm i\pi})$ satisfies (8). Similarly one can show that the $L_{n,j}(z)$ satisfy (8), starting with the Mellin-Barnes representation

$$L_{n,j}(z)=\frac{1}{2\pi i}\int_L \Gamma(a_j+s)\Gamma(1-a_j-s)\Omega(s)\frac{(n+\lambda)_s}{(n+1)_{-s}}z^s\,ds,$$

where L separates the poles of $\Gamma(a_j+s)\Gamma(n+\lambda+s)$ from those of $\Gamma(-s)\Gamma(1-b_{\mathrm{P}+1}-s)$.

COROLLARY. *If $Y_n(z)$ is a common solution to*

$$\mathscr{L}_z\{Y_n(z)\}=0\quad\textit{and}\quad \mathscr{M}_n\{Y_n(z)\}=0$$

in $\mathscr{D}$, and the limits, taken in $\mathscr{D}$,

$$\lim_{z\to1}\frac{dY_n(z)}{dz},\quad \lim_{z\to1}Y_{n-m}(z)=Y_{n-m}(1),\quad m=0,1,\cdots,p+1,$$

as well defined finite numbers for $n\geqq p+1$, then

$$0=\sum_{m=0}^{p+1}(A_m+B_m)Y_{n-m}(1).$$

Remark. If we set $b_{p+2}=1$, and

$$\mathscr{B}_1=\{F_{n,k}(z),\ k=1,\cdots,p+2\},$$

$$F_{n,k}(z)=\frac{(n+1)_{1-b_k}}{(n+1)_{-1+b_k}}z^{1-b_k}{}_{p+3}F_{p+2}\left(\begin{matrix}1,1-b_k-n,1-b_k+n+\lambda,1-b_k+a_P\\ 1-b_k+b_{P+2}\end{matrix}\middle| z\right),$$

then under the conditions

$$b_j-b_k\notin \mathbf{Z},\qquad j\neq k,$$

$\mathscr{B}_1$ forms a basis equivalent to $\mathscr{B}$ in

$$\mathscr{D}_0=\{z\colon 0<|z|<1,\ |\arg z|<\pi\}.$$

That is, considering $\mathscr{B}$. and $\mathscr{B}_1$ as column vectors,

$$\mathscr{B}_1=M\mathscr{B}$$

where M is a nonsingular matrix whose elements are independent of z and are periodic functions of n, of period 1 (see [3]). Thus, the $F_{n,k}(z)$ are common solutions to (1) and (2) in $\mathscr{D}\cap\mathscr{D}_0$, and by analytic continuation, in $\mathscr{D}$.

Hence, the $F_{n,k}(z)$ satisfy (8). Note that $P_n(z)=F_{n,p+2}(z)$.

Similary, if we set

$$\mathscr{B}_\infty=\{H_{n,k}(z),\ k=1,\cdots,p+2\},$$

$$H_{n,k}(z)=\frac{(n+\lambda)_{-a_k}}{(n+1)_{a_k}}z^{-a_k}{}_{p+3}F_{p+2}\left(\begin{matrix}1,1+a_k-b_{P+2}\\ 1+a_k+n,1+a_k-n-\lambda,1+a_k-a_P\end{matrix}\middle|\frac{1}{z}\right),$$

$$k=1,\cdots,p,$$

$$H_{n,p+1}(z)=\frac{\Gamma(n+1)\Gamma(n+\lambda+1-b_{P+2})z^{-n-\lambda}}{\Gamma(n+\lambda)\Gamma(2n+\lambda+1)\Gamma(n+\lambda+1-a_P)}{}_{p+2}F_{p+1}\left(\begin{matrix}n+\lambda+1-b_{P+2}\\ 2n+\lambda+1,n+\lambda+1-a_P\end{matrix}\middle|\frac{1}{z}\right),$$

$$H_{n,p+2}(z)=\frac{\Gamma(n+1)\Gamma(2n+\lambda)\Gamma(n+a_P)z^n}{\Gamma(n+\lambda)\Gamma(n+b_{P+2})}\left(\begin{matrix}-n,1-b_{P+2}\\ {}_{p+2}F_{p+1}1-2n-\lambda,-n+1-a_P\end{matrix}\middle|\frac{1}{z}\right)$$

$$2n+2\lambda\notin\mathbf{Z}^+,\quad n+a_j\notin\mathbf{Z}^+,\quad j=1,\cdots,p,$$

then under the conditions

$$a_j-a_k\notin\mathbf{Z},\qquad j\neq k,$$

$\mathscr{B}_\infty$ forms a basis equivalent to $\mathscr{B}$ in

$$\mathscr{D}_\infty=\{z\colon 1<|z|,\ |\arg(z-1)|<\pi\}.$$

Remark. In [8, p. 159], Wimp observed that it was conjectured by Lewanowicz that the numbers

$$Q_n={}_{p+2}F_{p+1}\left(\begin{matrix}n+\beta_{P+2}\\ 2n+\lambda+1,n+\alpha_P\end{matrix}\middle|1\right),$$

satisfy a $p+2$ term linear recursion relationship with rational coefficients in n. When we compare Q_n with $H_{n,p+1}$ (1), it is clear that the Q_n do satisfy such a linear recursion relationship. A proof by Lewanowicz has now been given in [5].

Remark. If $Y_n(z)$ is a function such that

$$\mathscr{L}_z\{Y_n(z)\}=0 \quad \text{and} \quad \mathscr{M}_n\{Y_n(z)\}=0 \quad \text{in } \mathscr{D},$$

it is an interesting, open question whether $Y_n(z)$ is a common solution to these equations according to our definition. Also, the structure of the linear solution space for (8) is unknown.

Remark. There are q-analogues of all these results which will appear elsewhere.

Remark. Recently, Lewanowicz [4], [5] has derived other differential-difference relations for $P_n(z)$ and demonstrates relations satisfied by the coefficients A_m, B_m.

REFERENCES

[1] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[2] J. L. Fields, Y. L. Luke and J. Wimp, *Recursion formulae for generalized hypergeometric functions*, J. Approx. Theory, 1 (1968), pp. 137-166.

[3] J. L. Fields, *Uniform asymptotic expansions of a class of Meijer G-functions for a large parameter*, this Journal, 14 (1983), pp. 1204-1253.

[4] S. Lewanowicz, *On the differential-difference properties of the extended Jacobi polynomials*, Math. Comp., 44 (1985), pp. 435-441.

[5] ———, *Recurrence relations for hypergeometric functions of unit argument*, Math. Comp., 45 (1985), pp. 521-535.

[6] Y. L. Luke, *The Special Functions and Their Approximations*, Vol. I and II, Academic Press, New York, 1969.

[7] J. Wimp, *Differential difference properties of hypergeometric polynomials*, Math. Comp., 29 (1975), pp. 577-581.

[8] ———, *Computation with Recurrence Relations*, Pitman, London, 1984.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

021

A PROOF OF THE G_2 CASE OF MACDONALD'S ROOT SYSTEM–DYSON CONJECTURE*

DORON ZEILBERGER†

Dedicated to Joe Gillis on the occasion of his 75th birthday

Key words. root systems, constant term identities, hypergeometric summation

AMS(MOS) subject classifications. 31, 05A

We will prove the following theorem.

THEOREM. *Let m and n be integers and x, y and z commuting indeterminates; then the constant term of the Laurent polynomial*

$$F(x,y,z)=\left[\left(1-\frac{x}{y}\right)\left(1-\frac{y}{z}\right)\left(1-\frac{z}{x}\right)\right]^m\left[\left(1-\frac{xy}{z^2}\right)\left(1-\frac{xz}{y^2}\right)\left(1-\frac{yz}{x^2}\right)\right]^n$$
$$\cdot\left[\left(1-\frac{y}{x}\right)\left(1-\frac{z}{y}\right)\left(1-\frac{x}{z}\right)\right]^m\left[\left(1-\frac{z^2}{xy}\right)\left(1-\frac{y^2}{xz}\right)\left(1-\frac{x^2}{yz}\right)\right]^n$$

is

$$C(m,n)=\frac{(3m+3n)!(3n)!(2m)!(2n)!}{(2m+3n)!(m+2n)!(m+n)!m!n!n!}.$$

This is the G_2 case of Macdonald's Root System–Dyson conjecture (see [6, Conjecture 2.3, and (c), p. 994]; see also Morris [7]).

Macdonald [6] showed how Selberg's integral [8] (see [1] for Aomoto's recent brilliant proof) implies his conjecture for all the so-called classical root systems. We will follow the same route and show how the G_2 case follows from a corollary of Selberg's integral that is due to Morris [7, p. 94].

After the first version of this paper was written, I was kindly informed by Dominique Foata that Laurent Habsieger [9] has independently and simultaneously obtained the results of this paper.

We only need the case $n=3$ of Morris' result that spells out to the following.

MORRIS' THEOREM ($n=3$). *Let a, b, c be integers. The constant term of the Laurent polynomial*

$$H(u,v,w;a,b,c)=[(1-u)(1-v)(1-w)]^a\quad\left[\left(1-\frac{u}{v}\right)\left(1-\frac{u}{w}\right)\left(1-\frac{v}{w}\right)\right]^c$$
$$\cdot\left[\left(1-\frac{1}{u}\right)\left(1-\frac{1}{v}\right)\left(1-\frac{1}{w}\right)\right]^b\left[\left(1-\frac{v}{u}\right)\left(1-\frac{w}{u}\right)\left(1-\frac{w}{v}\right)\right]^c$$

is

$$\frac{(a+b+2c)!(a+b+c)!(a+b)!(2c)!(3c)!}{(a+2c)!(b+2c)!(a+c)!(b+c)!a!b!c!c!}.$$

We will need the following easy corollary.

* Received by the editors May 12, 1986; accepted for publication (in revised form) June 25, 1986. This research was partly sponsored by the National Science Foundation

† Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

COROLLARY. *The coefficient of $u^Av^Aw^A$ in $H(u, v, w; a, b, c)$ above is*

$$(-1)^A\frac{(a+b+2c)!(a+b+c)!(a+b)!(2c)!(3c)!}{(a-A+2c)!(b+A+2c)!(a-A+c)!(b+A+c)!(a-A)!(b+A)!c!c!}.$$

Proof. Since $(1-t)^a(1-t^{-1})^b/t^A=(-1)^A(1-t)^{a-A}(1-t^{-1})^{b+A}$, we have

$$H(u, v, w; a, b, c)/u^Av^Aw^A=(-1)^AH(u, v, w; a-A, b+A, c)$$

and taking constant terms, the corollary follows from Morris' theorem.

Finally, we need the formula shown below.

DIXON'S FORMULA (e.g. [5, 1.2.6, Ex. 62, pp. 73 and 489]). *Let M, N, K be integers; then*

$$\sum_A\frac{(-1)^A}{(M+A)!(M-A)!(N+A)!(N-A)!(K+A)!(K-A)!}$$
$$=\frac{(M+N+K)!}{M!N!K!(M+N)!(M+K)!(N+K)!}.$$

To prove the theorem we let $u=x/y$, $v=y/z$, and $w=z/x$. Then $F(x, y, x)=H(u, v, w; m, m, n)$. But $uvw=1$, so the constant term of F is the sum of all the diagonal coefficients of H. Thus by the corollary the constant term of F is

$$\sum_A(-1)^A\frac{(2m+n)!(2m+n)!(2m)!(2n)!(3n)!}{(m-A+2n)!(m+A+2n)!(m-A+n)!(m+A+n)!(m-A)!(m+A)!(n)!(n)!}$$
$$=\frac{(2m+2n)!(2m+n)!(2m)!(2n)!(3n)!}{n!n!}$$
$$\cdot\sum_A\frac{(-1)^A}{(m+2n-A)!(m+2n+A)!(m+n-A)!(m+n+A)!(m-A)!(m+A)!}.$$

Using Dixon's formula with $M=m+2n$, $N=m+n$, $K=m$, we get that this is equal to

$$\frac{(2m+2n)!(2m+n)!(2m)!(2n)!(3n)!}{n!n!}$$
$$\cdot\frac{(3m+3n)!}{(m+2n)!(m+n)!m!(2m+3n)!(2m+2n)!(2m+n)!}$$
$$=\frac{(3m+3n)!(3n)!(2m)!(2n)!}{(2m+3n)!(m+2n)!(m+n)!m!n!n!}. \qquad \text{Q.E.D.}$$

Since F of the theorem is obviously with integer coefficients, our theorem implies the not entirely obvious fact that $C(m, n)$ is an integer, thus solving Askey's problem [2].

The q-Analogue. We will show how Kadell's [4] recent q-analogue of Morris' theorem implies the q-analogue of the G_2 Macdonald-Dyson conjecture [6]. Since the ordinary case is just the special case $q=1$ of the q-analogue, we could have started with the q-analogue right away, giving the ordinary reader the option to plug in $q=1$ throughout. However we feel that this would have been very poor pedagogy. Indeed, the way mathematics is created is by slowly increasing steps of generality. Unfortunately, all too often results are presented in their overpowering full generality right from the start, thus making them very hard to read and understand, let alone use as motivation.

Let

$$(y)_a = (1-y)(1-qy)\cdots(1-q^{a-1}y)$$

and

$$[a]! = \frac{(q)_a}{(1-q)^a} = 1(1+q)(1+q+q^2)\cdots(1+q+\cdots+q^{a-1}).$$

We will prove the following theorem.

q-THEOREM. *Let m and n be integers and x, y and z commuting indeterminates; then the constant term of the Laurent polynomial*

$$F(x,y,z) = \left(\frac{x}{y}\right)_m \left(\frac{z}{y}\right)_m \left(\frac{z}{x}\right)_m \left(\frac{z^2}{xy}\right)_n \left(\frac{xz}{y^2}\right)_n \left(\frac{yz}{x^2}\right)_n$$
$$\cdot \left(q\frac{y}{x}\right)_m \left(q\frac{y}{z}\right)_m \left(q\frac{x}{z}\right)_m \left(q\frac{xy}{z^2}\right)_n \left(q\frac{y^2}{xz}\right)_n \left(q\frac{x^2}{yz}\right)_n$$

is

$$\frac{[3m+3n]![3n]![2m]![2n]!}{[2m+3n]![m+2n]![m+n]![m]![n]![n]!}.$$

We need the following theorem [4].

KADELL'S q-MORRIS THEOREM ($n=3$). *Let a, b, c be integers. The constant term of the Laurent polynomial*

$$H(u,v,w;a,b,c) = (u)_a(v)_a(w)_a \left(\frac{u}{v}\right)_c \left(\frac{u}{w}\right)_c \left(\frac{v}{w}\right)_c \left(\frac{q}{u}\right)_b \left(\frac{q}{v}\right)_b \left(\frac{q}{w}\right)_b \left(q\frac{v}{u}\right)_c \left(q\frac{w}{u}\right)_c \left(q\frac{w}{v}\right)_c$$

is

$$\frac{[a+b+2c]![a+b+c]![a+b]![2c]![3c]!}{[a+2c]![b+2c]![a+c]![b+c]![a]![b]![c]![c]!}.$$

We will need the following easy corollary.

q-COROLLARY. *The coefficient of $u^A v^A w^A$ in $H(u,v,w;a,b,c)$ above is*

$$(-1)^A q^{3A(A-1)/2}$$
$$\cdot \frac{[a+b+2c]![a+b+c]![a+b]![2c]![3c]!}{[a-A+2c]![b+A+2c]![a-A+c]![b+A+c]![a-A]![b+A]![c]![c]!}.$$

Proof. We are really looking for the constant term of $H(u,v,w;a,b,c)/u^A v^A w^A$. But since

$$\frac{(t)_a(q(1/t))_b}{t^A} = (-1)^A q^{A(A-1)/2}(q^A t)_{a-A}\left(\frac{q}{q^A t}\right)_{b+A}.$$

it follows that

$$\frac{H(u,v,w;a,b,c)}{u^A v^A w^A} = (-1)^A q^{3A(A-1)/2} H(q^A u, q^A v, q^A w; a-A, b+A, c)$$

and the corollary follows from Kadell's q-Morris Theorem.

Finally, we need the following formula.

THE q-DIXON FORMULA ([3], [5, p. 489]).

$$\sum_A \frac{(-1)^A q^{A(3A-1)/2}}{[M+A]![M-A]![N+A]![N-A]![K+A]![K-A]!} = \frac{[M+N+K]!}{[M]![N]![K]![M+N]![M+K]![N+K]!}.$$

To prove the q-Theorem we let $u=q(y/z)$, and $v=z/x$ and $w=x/y$. Then $F(x,y,z)=H(u,v,w;m,m,n)$. But $uvw=q$ so the constant term of F is the weighted sum of all the diagonal coefficients of H, where the coefficient of $u^Av^Aw^A$ gets multiplied by q^A.

Thus by the corollary the constant term of F is

$$\sum_A q^A(-1)^A q^{3A(A-1)/2} \cdot \frac{[2m+2n]![2m+n]![2m]![2n]![3n]!}{[m-A+2n]![m+A+2n]![m-A+n]![m+A+n]![m-A]![m+A]![n]![n]!}$$
$$= \frac{[2m+2n]![2m+n]![2m]![2n]![3n]!}{[n]![n]!} \cdot \sum_A \frac{(-1)^A q^{A(3A-1)/2}}{[m+2n-A]![m+2n+A]![m+n-A]![m+n+A]![m-A]![m+A]!}.$$

Using the q-Dixon formula with $M=m+2n$, $N=m+n$, $K=m$, we get that this is equal to

$$\frac{[2m+2n]![2m+n]![2m]![2n]![3n]!}{[n]![n]!} \cdot \frac{[3m+3n]!}{[m+2n]![m+n]!m![2m+3n]![2m+2n]![2m+n]!}$$
$$= \frac{[3m+3n]![3n]![2m]![2n]!}{[2m+3n]![m+2n]![m+n]![m]![n]![n]!}. \qquad \text{Q.E.D.}$$

Acknowledgment. I heartily thank Richard Askey for rekindling my interest in the Macdonald conjecture.

REFERENCES

[1] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, this Journal, to appear.

[2] R. ASKEY, *Advanced problem* 6514, Amer. Math. Monthly, 93 (1986), pp. 304-305.

[3] F. H. JACKSON, *Summation of q-hypergeometric series*, Messenger of Math., 47 (1917), pp. 101-112.

[4] K. W. J. KADELL, *A proof of Askey's conjectured q-analog of Selberg's integral and a conjecture of Morris*, preprint.

[5] D. E. KNUTH, *Fundamental Algorithms*, in The Art of Computer Programming, vol. 1, second edition, Addison-Wesley, Reading, MA, 1973.

[6] I. G. MACDONALD, *Some conjectures for root systems and finite reflection groups*, this Journal, 13 (1982), pp. 988-1007.

[7] W. MORRIS, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin, Madison, 1982.

[8] A. SELBERG, *Bemerkinger om et multiplet integral*, Normat, 26 (1944), pp. 71-78.

[9] LAURENT HABSIEGER, *La q-conjecture de Macdonald-Morris pour G_2*, C.R. Acad. Sci Paris Sér. I Math. to appear.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

022

AN ANALYTIC CONTINUATION OF THE HYPERGEOMETRIC SERIES*

WOLFGANG BÜHRING†

Abstract. The points $z=\frac{1}{2}(1\pm i\sqrt{3})$ of the complex z-plane are on the boundary for each of the convergence domains of the various hypergeometric series which appear in the transformation or continuation formulas of the hypergeometric function ${}_2F_1(a, b; c; z)$. This paper presents a continuation formula containing series in powers of $1/(z-\frac{1}{2})$ with the convergence domain $|z-\frac{1}{2}|>\frac{1}{2}$, which contains the two points in question in its interior. The coefficients of the power series are determined by a three-term recurrence relation and are represented explicitly in terms of terminating hypergeometric series. If $2c=a+b+1$, then one term of the recurrence relation disappears and the series become hypergeometric series.

Key words. special functions, hypergeometric series, hypergeometric functions, continuation formulas, hypergeometric differential equations

AMS(MOS) subject classifications. 33A30, 34A20, 34A30, 30B40

1. Introduction. This paper is concerned with the hypergeometric differential equation

$$z(1-z)w''(z)+\{c-(a+b+1)z\}w'(z)-abw(z)=0 \tag{1}$$

and its solution

$${}_2F_1(a, b; c; z)=\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n n!}z^n. \tag{2}$$

Here z is the complex variable and a, b, c are complex parameters, with c not equal to a negative integer or zero. The notation $(a)_n$ means the Pochhammer symbol

$$(a)_n=a(a+1)\cdots(a+n-1)=\frac{\Gamma(a+n)}{\Gamma(a)}. \tag{3}$$

In addition to the gamma function $\Gamma(z)$ we shall require its logarithmic derivative $\psi(z)=\Gamma'(z)/\Gamma(z)$. Following common practice, we use the same symbol ${}_2F_1$ to denote the hypergeometric function as well as the hypergeometric series on the right-hand side of (2) which represents the function inside the unit circle.

By means of the well-known transformation or continuation formulas it is possible to express the hypergeometric function in terms of one or two other hypergeometric functions with suitably changed parameters and the variable z replaced by either $z/(z-1)$ or $1-z$ or $1/z$ or $1/(1-z)$ or $(z-1)/z$. The corresponding series have different convergence domains and so each of the formulas gives an analytic continuation of the original hypergeometric series on the right-hand side of (2). An example, which is relevant for the later discussion, is the continuation formula

$$\begin{aligned}\{\Gamma(c)\}^{-1}{}_2F_1(a, b; c; z)&=\frac{\Gamma(b-a)}{\Gamma(b)\Gamma(c-a)}(-z)^{-a}{}_2F_1\left(a, 1-c+a; 1-b+a; \frac{1}{z}\right)\\&\quad+\frac{\Gamma(a-b)}{\Gamma(a)\Gamma(c-b)}(-z)^{-b}{}_2F_1\left(b, 1-c+b; 1-a+b; \frac{1}{z}\right),\end{aligned} \tag{4}$$

$$(|\arg(-z)|<\pi, \quad b-a \text{ not an integer}).$$

* Received by the editors June 3, 1985; accepted for publication (in revised form) May 20, 1986.

† Physikalisches Institut, Universität Heidelberg, D-6900 Heidelberg, West Germany.

Following Olver [3] we consider $\{\Gamma(c)\}^{-1}{}_2F_1(a, b; c; z)$ rather than ${}_2F_1$ itself in order to avoid the restriction that c be not equal to a negative integer or zero.

There are two points of the complex z-plane which are on the boundary of the convergence domain for each of the hypergeometric series appearing in the various formulas mentioned so far. It is the purpose of this paper to communicate an appropriate continuation formula which expresses ${}_2F_1(a, b; c; z)$ in terms of power series such that the two points in question, $z=\frac{1}{2}(1\pm i\sqrt{3})=\exp(\pm i\pi/3)$, are interior points of the convergence domain. These series are not hypergeometric, but their coefficients satisfy a simple three-term recurrence relation. They also may be represented in terms of terminating hypergeometric series.

Our general results are presented in § 2; some interesting special cases are discussed in § 3. The proofs are supplied in § 4-§ 6, and the concluding remarks in § 7 deal with the computational aspects of this work.

2. General results. It would be a simple matter but of little use to consider a Taylor series solution in powers of $z-z_0$ in a neighborhood of an ordinary point z_0 of the differential equation, since then the problem remains to evaluate the two initial coefficients which are equal to ${}_2F_1(a, b; c; z_0)$ and $(ab/c)\ {}_2F_1(a+1, b+1; c+1; z_0)$, respectively. Such a problem does not occur with series in powers of $1/(z-z_0)$. For since they converge in a neighborhood of $z=\infty$, in a similar way to the hypergeometric series on the right of (4), the required initial coefficients may immediately be obtained by comparison with (4). Following this idea in detail, we can prove

THEOREM 1. *If $b-a$ is not an integer, we have for $|\arg(z_0-z)|<\pi$ the continuation formula*

$$\begin{aligned}\{\Gamma(c)\}^{-1}{}_2F_1(a, b; c; z)&=\frac{\Gamma(b-a)}{\Gamma(b)\Gamma(c-a)}(z_0-z)^{-a}\sum_{n=0}^{\infty}d_n(a, z_0)(z-z_0)^{-n}\\&\quad+\frac{\Gamma(a-b)}{\Gamma(a)\Gamma(c-b)}(z_0-z)^{-b}\sum_{n=0}^{\infty}d_n(b, z_0)(z-z_0)^{-n},\end{aligned}\tag{5}$$

where the series converge outside the circle $|z-z_0|=\max(|z_0|, |z_0-1|)$ and the coefficients $d_n(s, z_0)$ are given by the three-term recurrence relation

$$\begin{aligned}d_n(s, z_0)&=\{n(n+2s-a-b)\}^{-1}(n+s-1)\\&\quad\cdot(\{(n+s)(1-2z_0)+(a+b+1)z_0-c\}d_{n-1}(s, z_0)\\&\quad+z_0(1-z_0)(n+s-2)d_{n-2}(s, z_0))\end{aligned}\tag{6}$$

with starting values

$$d_{-1}(s, z_0)=0, \qquad d_0(s, z_0)=1.\tag{7}$$

The series in (5) may be rewritten as

$$\sum_{n=0}^{\infty}d_n(s, z_0)(z-z_0)^{-n}=\sum_{n=0}^{\infty}\frac{(s)_n}{(1+2s-a-b)_n}e_n(s, z_0)(z-z_0)^{-n},\tag{8}$$

where the coefficients $e_n(s, z_0)$ now obey the recurrence relation

$$\begin{aligned}e_n(s, z_0)&=n^{-1}(\{(n+s)(1-2z_0)+(a+b+1)z_0-c\}e_{n-1}(s, z_0)\\&\quad+z_0(1-z_0)(n-1+2s-a-b)e_{n-2}(s, z_0))\end{aligned}\tag{9}$$

with starting values

$$e_{-1}(s, z_0)=0, \qquad e_0(s, z_0)=1.\tag{10}$$

The coefficients d_n are also given by

$$d_n(s, z_0) = \frac{(s)_n(1+s-c)_n}{(1+2s-a-b)_n n!} {}_2F_1(-n, a+b-2s-n; c-s-n; z_0) \tag{11}$$

or by

$$\begin{aligned} d_n(s, z_0) = (-1)^n \frac{(s)_n(s+c-a-b)_n}{(1+2s-a-b)_n n!} \\ \cdot {}_2F_1(-n, a+b-2s-n; 1+a+b-s-c-n; 1-z_0). \end{aligned} \tag{12}$$

A formula corresponding to (5) for the case when $b-a$ is an integer can be derived from (5) by a limiting process in a similar way as described in [2]. The case $b=a$ is covered by

COROLLARY 1. *For* $|\arg(z_0-z)|<\pi$ *there holds*

$$\begin{aligned} \{\Gamma(c)\}^{-1} {}_2F_1(a, a; c; z) = \frac{1}{\Gamma(a)\Gamma(c-a)}(z_0-z)^{-a} \sum_{n=0}^{\infty} \frac{(a)_n}{n!}(e_n(a, z_0) \\ \cdot \{2\psi(1+n) - \psi(a+n) - \psi(c-a) \\ + \ln(z_0-z)\} - f_n(z_0))(z-z_0)^{-n}, \end{aligned} \tag{13}$$

where the series converges for $|z-z_0| > \max(|z_0|, |z_0-1|)$ *and the coefficients* $e_n(a, z_0)$ *are given by* (9)–(10) *with* $b=a$ *and the* $f_n(z_0)$ *by the recurrence relation*

$$\begin{aligned} f_n(z_0) = n^{-1}(\{(n+a)(1-2z_0) + (2a+1)z_0 - c\} f_{n-1}(z_0) \\ + z_0(1-z_0)(n-1) f_{n-2}(z_0) + (1-2z_0) e_{n-1}(a, z_0) \\ + 2z_0(1-z_0) e_{n-2}(a, z_0)) \end{aligned} \tag{14}$$

with starting values

$$f_{-1}(z_0) = f_0(z_0) = 0. \tag{15}$$

The case when $b-a$ is an integer different from zero is less simple and is therefore omitted.

3. Special cases. We discuss some special cases of Theorem 1.

(i) For $z_0=0$ or $z_0=1$ the recurrence relation (6) simplifies and becomes a two-term recurrence relation, the series in (5) are then hypergeometric. Consequently (5) reduces to (4) if $z_0=0$ or to another well-known continuation formula if $z_0=1$.

(ii) The case when $z_0=\frac{1}{2}$ is unique in so far as then two singular points of the differential equation are on the boundary of the convergence domain of the series. Although (6) simplifies, it remains a three-term recurrence relation. The series converge outside the circle $|z-\frac{1}{2}|=\frac{1}{2}$, which means that the convergence domain has been enlarged considerably as compared with the convergence domain $|z|>1$ of the hypergeometric series on the right of (4). As a consequence, both points in question, $z=\frac{1}{2}(1\pm i\sqrt{3})$, are now inside the convergence domain.

(iii) If $z_0=\frac{1}{2}$ and $c=\frac{1}{2}(a+b+1)$, which implies that the two singular points on the boundary of the convergence domain have the same characteristic exponents, then we have

COROLLARY 2. *If $b-a$ is not an even integer, there holds for $|\arg(\frac{1}{2}-z)|<\pi$ the continuation formula*

$$\begin{aligned}
&\{\Gamma(\tfrac{1}{2}a+\tfrac{1}{2}b+\tfrac{1}{2})\}^{-1}{}_2F_1(a, b; \tfrac{1}{2}a+\tfrac{1}{2}b+\tfrac{1}{2}; z)\\
&\quad=\frac{2^{b-a-1}\Gamma(\frac{1}{2}b-\frac{1}{2}a)}{\Gamma(b)\sqrt{\pi}}(\tfrac{1}{2}-z)^{-a}{}_2F_1(\tfrac{1}{2}a+\tfrac{1}{2}, \tfrac{1}{2}a; \tfrac{1}{2}a-\tfrac{1}{2}b+1; \{2z-1\}^{-2})\\
&\quad+\frac{2^{a-b-1}\Gamma(\frac{1}{2}a-\frac{1}{2}b)}{\Gamma(a)\sqrt{\pi}}(\tfrac{1}{2}-z)^{-b}{}_2F_1(\tfrac{1}{2}b+\tfrac{1}{2}, \tfrac{1}{2}b; \tfrac{1}{2}b-\tfrac{1}{2}a+1; \{2z-1\}^{-2}),
\end{aligned}\tag{16}$$

where the series on the right converge for $|z-\frac{1}{2}|>\frac{1}{2}$.

For $a=-\nu$ and $b=\nu+1$, if z is replaced by $\frac{1}{2}(1-z)$, this reduces to (8.1.5) of [1] with $\mu=0$.

A continuation formula valid when $b-a$ is an even integer may be obtained from (16) by a limiting process. The case $b=a$ is covered by

COROLLARY 3. *For $|\arg(\frac{1}{2}-z)|<\pi$ there holds*

$$\begin{aligned}
&\{\Gamma(a+\tfrac{1}{2})\}^{-1}{}_2F_1(a, a; a+\tfrac{1}{2}; z)\\
&\quad=\frac{1}{\Gamma(a)\sqrt{\pi}}\left(\frac{1}{2}-z\right)^{-a}\sum_{n=0}^{\infty}\frac{(\frac{1}{2}a+\frac{1}{2})_n(\frac{1}{2}a)_n}{(n!)^2}\\
&\qquad\cdot\left\{\psi(1+n)-\psi(a+2n)+2\ln(2)+\ln\left(\frac{1}{2}-z\right)\right\}(2z-1)^{-2n},
\end{aligned}\tag{17}$$

the series on the right being convergent for $|z-\frac{1}{2}|>\frac{1}{2}$.

The case $b-a=2m$ with $m=1, 2, 3\cdots$ is covered by

COROLLARY 4. *For $|\arg(\frac{1}{2}-z)|<\pi$ and for $m=1, 2, 3\cdots$ there holds*

$$\begin{aligned}
&\{\Gamma(a+m+\tfrac{1}{2})\}^{-1}{}_2F_1(a, a+2m; a+m+\tfrac{1}{2}; z)\\
&\quad=\frac{2^{a-1+2m}}{\Gamma(a+2m)\sqrt{\pi}}(1-2z)^{-a}\sum_{n=0}^{m-1}(-1)^n\frac{(\frac{1}{2}a+\frac{1}{2})_n(\frac{1}{2}a)_n(m-n-1)!}{n!}(2z-1)^{-2n}\\
&\quad+(-1)^m\frac{2^{a-1}}{\Gamma(a)\sqrt{\pi}}(1-2z)^{-a-2m}\sum_{n=0}^{\infty}\frac{(\frac{1}{2}a+\frac{1}{2}+m)_n(\frac{1}{2}a+m)_n}{(m+n)!\,n!}\\
&\quad\cdot\{\psi(1+m+n)+\psi(1+n)-\psi(\tfrac{1}{2}a+\tfrac{1}{2}+m+n)-\psi(\tfrac{1}{2}a+m+n)\\
&\qquad\qquad+2\ln(1-2z)\}(2z-1)^{-2n},
\end{aligned}\tag{18}$$

the infinite series being convergent for $|z-\frac{1}{2}|>\frac{1}{2}$.

Corollary 4 contains Corollary 3 as a special case if the empty sum is interpreted as zero. The formulas (17) and (18) can be written in various other ways by means of the duplication formulas of ψ and Γ so that $\psi(a+2m+2n)$ and $(a+2m)_{2n}$ or $(a)_{2m+2n}$ appears.

4. Proof of Theorem 1. The hypergeometric function ${}_2F_1(a, b; c; z)$ is a solution of the differential equation (1) which, when rewritten in the variable

$$x=z-z_0,\tag{19}$$

reads

$$\{-x^2+(1-2z_0)x+z_0(1-z_0)\}u''+\{c-(a+b+1)z_0-(a+b+1)x\}u'-abu(x)=0.\tag{20}$$

Relative to the regular singular point at infinity the Frobenius ansatz

$$u(x)=\sum_{n=0}^{\infty}d_n(s, z_0)x^{-s-n}\tag{21}$$

yields the recurrence relation (6) for the coefficients $d_n(s, z_0)$ with the characteristic exponents $s \in \{a, b\}$ as expected.

Next we observe that, when $|z|$ is sufficiently large, $x^{-s-n} = z^{-s-n}\{1-(z_0/z)\}^{-s-n}$ may be expanded in powers of $1/z$ for any fixed z_0. Then $u(z)$ becomes a power series in $1/z$ multiplied by z^{-s}. Since there is (apart from a constant factor) only one solution of the hypergeometric differential equation with this analytical structure, $u(x)$ must be equal to the well-known solution which corresponds to the special case when $z_0 = 0$. We therefore have, for each of the two possible values of s,

$$(22) \qquad z^{-s}{}_2F_1(s, 1+s-c; 1+2s-a-b; z^{-1}) = (z-z_0)^{-s} \sum_{n=0}^{\infty} d_n(s, z_0)(z-z_0)^{-n},$$

by means of which the continuation formula (5) follows from (4).

The explicit representation for $d_n(s, z_0)$ may be obtained from (22) by applying

$$(23) \qquad z^{-s-n} = \left\{(z-z_0)\left(1+\frac{z_0}{z-z_0}\right)\right\}^{-s-n} = (z-z_0)^{-s-n} \sum_{k=0}^{\infty} \frac{(s+n)_k}{k!}\left(\frac{-z_0}{z-z_0}\right)^k$$

to the series on the left. Collecting the terms with equal powers of $z-z_0$ we obtain

$$(24) \qquad (z-z_0)^{-s} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(s)_k(1+s-c)_k}{(1+2s-a-b)_k k!} \frac{(s+k)_{n-k}(-z_0)^{n-k}}{(n-k)!}(z-z_0)^{-n},$$

which in view of

$$(25) \qquad \frac{(s+k)_{n-k}}{(n-k)!} = (-1)^k \frac{(s)_n(-n)_k}{(s)_k n!}$$

is equal to

$$(26) \qquad (z-z_0)^{-s} \sum_{n=0}^{\infty} \frac{(s)_n}{n!}(-z_0)^n(z-z_0)^{-n} \sum_{k=0}^{n} \frac{(-n)_k(1+s-c)_k}{(1+2s-a-b)_k k!}(z_0)^{-k}.$$

The sum over k is a terminating hypergeometric series, and comparison of (26) with the right-hand side of (22) then shows that

$$(27) \qquad d_n(s, z_0) = (n!)^{-1}(s)_n(-z_0)^n{}_2F_1(-n, 1+s-c; 1+2s-a-b; 1/z_0).$$

Application of the appropriate continuation formula of the hypergeometric function now yields (11) or (12), respectively.

5. Proof of Corollary 1. The starting point is (5) with the series rewritten in terms of the coefficients $e_n(s, z_0)$ according to (8). Then a is replaced by $a-\varepsilon$ and b by $a+\varepsilon$, so that the $e_n(s, z_0)$ depend on ε via s only. Introducing

$$(28) \qquad f_n(z_0) = \left\{\frac{d}{ds} e_n(s, z_0)\right\}_{s=a} = -\left\{\frac{d}{d\varepsilon} e_n(a-\varepsilon, z_0)\right\}_{\varepsilon=0}$$

and performing the limit $\varepsilon \to 0$ we obtain (13). The relation (14) follows if (9), with $a+b=2a$, is differentiated with respect to s and then s put equal to a.

6. Proof of Corollaries 2–4. When $z_0 = \frac{1}{2}$ and $c = \frac{1}{2}(a+b+1)$, the recurrence relation (6) reads

$$(29) \qquad d_n(s, \tfrac{1}{2}) = 2^{-2}\{n(n+2s-a-b)\}^{-1}(n+s-1)(n+s-2)d_{n-2}(s, \tfrac{1}{2}),$$

and the series in question becomes

$$(30) \qquad {}_2F_1(\tfrac{1}{2}s+\tfrac{1}{2}, \tfrac{1}{2}s; 1+s-\tfrac{1}{2}a-\tfrac{1}{2}b; \{2z-1\}^{-2}).$$

The factor in front of the series in (5) is

$$\frac{\Gamma(a+b-2s)}{\Gamma(a+b-s)\Gamma(c-s)}(z_0-z)^{-s}. \tag{31}$$

In the context of Corollary 2 it becomes, using the gamma duplication formula,

$$\frac{2^{a+b-2s-1}\Gamma(\frac{1}{2}a+\frac{1}{2}b-s)}{\Gamma(a+b-s)\sqrt{\pi}}(\tfrac{1}{2}-z)^{-s}. \tag{32}$$

With the two possible values of s the two connection coefficients on the right of (16) follow immediately.

Corollary 2 can also be verified in a different way by means of known transformation formulas. For if $c=\frac{1}{2}(a+b+1)$, the hypergeometric functions on the right of (4) appear with parameters such that the quadratic transformation formula (according to [1, (15.3.16)])

$$ {}_2F_1(A, B; 2B; Z)=\left\{\frac{2}{2-Z}\right\}^A {}_2F_1\left(\frac{1}{2}A+\frac{1}{2}, \frac{1}{2}A; B+\frac{1}{2}; \left\{\frac{Z}{2-Z}\right\}^2\right) \tag{33}$$

can be applied to each of them. The required result then follows in view of (32).

Corollaries 3 and 4 follow from (16) by tedious but standard calculations.

7. Concluding remarks. From a computational point of view the most efficient method for evaluating the hypergeometric function is the Miller algorithm. The relevant formulas are now easily accessible in the monograph by Wimp [4].

Nevertheless our Theorem 1 offers a method for computing the hypergeometric function at the two points in question which before were practically inaccessible by power series. If $z_0=\frac{1}{2}$ is used, the convergence ratio is $q=|\frac{1}{2}/(z-\frac{1}{2})|$, which for the two exceptional points has the value $q=1/\sqrt{3}=0.577\cdots$.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] Y. L. Luke, *The Special Functions and Their Approximations*, Vol. 1, Academic Press, New York, 1969.

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

[4] J. Wimp, *Computation with Recurrence Relations*, Pitman, Boston, 1984.

SIAM J. MATH. ANAL.
Vol. 18, No. 3, May 1987

023

ERROR BOUNDS FOR ASYMPTOTIC EXPANSIONS OF THE RATIO OF TWO GAMMA FUNCTIONS*

C. L. FRENZEN†

Abstract. Error bounds are obtained for asymptotic expansions of the ratio of two gamma functions $\Gamma(x+a)/\Gamma(x+b)$ for the case of real, bounded a, b and large positive x. In particular an assertion of Luke about a result of Fields is rigorously justified by showing that the error made in truncating Fields' asymptotic expansion is numerically less than and has the same sign as the first neglected term. Use is made of some results for completely monotonic functions and enveloping series.

Key words. error bound, asymptotic expansion, gamma function

AMS (MOS) subject classifications. Primary 41A60; secondary 33A15

1. In 1951 Tricomi and Erdélyi [12] derived an asymptotic expansion for the ratio of two gamma functions in the form

$$\frac{\Gamma(z+a)}{\Gamma(z+b)}=\sum_{j=0}^{m-1}(-1)^j\frac{\Gamma(b-a+j)}{\Gamma(b-a)j!}B_j^{(a-b+1)}(a)z^{a-b-j}+O(z^{a-b-m}), \tag{1.1}$$

valid for all $m\geqq 1$ as $z\to\infty$ with $|\arg(z+a)|<\pi$. In (1.1) the quantities a and b are bounded complex numbers and the $B_j^{(\sigma)}(a)$ are the generalized Bernoulli polynomials defined by

$$\left(\frac{t}{e^t-1}\right)^\sigma e^{at}=\sum_{j=0}^{\infty}\frac{t^j}{j!}B_j^{(\sigma)}(a),\quad B_0^{(\sigma)}(a)=1,\quad |t|<2\pi. \tag{1.2}$$

A similar result was obtained by Fields [3], who showed that for all $m\geqq 1$

$$\frac{\Gamma(z+a)}{\Gamma(z+b)}=\sum_{j=0}^{m-1}\frac{\Gamma(1-2\rho+2j)}{\Gamma(1-2\rho)(2j)!}B_{2j}^{(2\rho)}(\rho)w^{2\rho-1-2j}+O(w^{2\rho-1-2m}), \tag{1.3}$$

as $w\to\infty$ with $|\arg(w+\rho)|<\pi$, where $2w=2z+a+b-1$ and $2\rho=a-b+1$. Furthermore, in a later paper, Fields [4] gave an improved order estimate for the remainder in (1.3). Several other authors have investigated the asymptotic expansion of the left-hand side of (1.1) for essentially the special case $a=0$, $b=\frac{1}{2}$ and $z=n$, where n is a positive integer (see [1], [6], [7] and [9]). Recently Luke [6] pointed out that when a, b and z are all real, simple adjustments using the functional equation for the gamma function can always be made so that $0<\rho<\frac{1}{2}$ $(0<a-b+1<1)$ in (1.3). In this case the generalized Bernoulli polynomials $B_{2k}^{(2\rho)}(\rho)$ are positive when k is even and negative when k is odd so that the series in (1.3) is alternating. Luke assumed, as in the case of certain alternating convergent series, that consecutive partial sums of the right-hand side of (1.3) yield upper and lower bounds for the left-hand side and his numerical calculations supported this assumption for the special case $a=\frac{1}{2}$, $b=1$ (i.e., $\rho=\frac{1}{4}$). However, while this is always the case for large enough values of real positive z in an alternating asymptotic expansion, it is not known in general at what point this occurs (see [8, p. 68] for details).

The purpose of this paper is to rigorously justify Luke's assertion about Fields' result, and consequently to establish computable error bounds for an asymptotic

* Received by the editors January 28, 1985; accepted for publication (in revised form) January 7, 1986.

† Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

expansion of the ratio of two gamma functions for the case of real bounded a, b and real z. These results make use of some of the properties of completely monotonic functions and enveloping series. Error bounds for a one parameter family of asymptotic approximations for the ratio of two gamma functions are also established which include (1.1) in a certain parameter range of a, b. Since only the real case will be considered, from now on we take $z=x$.

2. A function $f(t)$ is said to be *completely monotonic over* (t_1, t_2), where $-\infty \leqq t_1 < t_2 \leqq +\infty$, if

$$(2.1) \qquad (-1)^n f^{(n)}(t) \geqq 0, \qquad t_1 < t < t_2, \quad n = 0, 1, 2, \cdots.$$

J. Dubourdieu [2] showed that strict inequality holds in (2.1) for all nonconstant functions completely monotonic over (t_1, ∞); that is, if $f(t)$ satisfies (2.1) with $t_2 = \infty$ and is not constant, then

$$(2.2) \qquad (-1)^n f^{(n)}(t) > 0, \qquad t_1 < t < \infty, \quad n = 0, 1, 2, \cdots.$$

A function $f(t)$ is said to be *monotonic of order* N if (2.1) holds for $n = 0, 1, 2, \cdots, N$. Examples of commonly occurring completely monotonic functions are e^{-t} and $(t-t_1)^{-\gamma}$, $\gamma \geqq 0$. If $f(t)$ is also continuous at $t = t_1$, then it is called *completely monotonic over* $[t_1, t_2)$ and similar definitions exist for $(t_1, t_2]$ and $[t_1, t_2]$. In the standard case, which is the one appropriate to this paper, $t_1 = 0$ and $t_2 = \infty$. These values for t_1 and t_2 will be assumed from now on. Completely monotonic functions are often useful in asymptotic analysis because if $f(t)$ is completely monotonic over $[0, \infty)$, then

$$(2.3) \qquad |f^{(n)}(t)| \leqq |f^{(n)}(0)|, \qquad 0 \leqq t < \infty, \quad n = 0, 1, 2, \cdots.$$

A detailed study of the concept of completely monotonic functions can be found in [13, Chap. IV]. Next we discuss the notion of an enveloping series.

The series $a_0 + a_1 + a_2 + \cdots$ is said to envelop the number A if the relations

$$(2.4) \qquad |A - (a_0 + a_1 + \cdots + a_n)| < |a_{n+1}|, \qquad n = 0, 1, 2, \cdots$$

are satisfied. The enveloping series may be convergent (to A) or divergent. If A, $a_0, a_1, \cdots$ are all real and

$$(2.5) \qquad A - (a_0 + a_1 + \cdots + a_n) = \theta_n a_{n+1}, \qquad 0 < \theta_n < 1, \quad n = 0, 1, 2, \cdots,$$

then A is enveloped by the series $a_0 + a_1 + \cdots$ and in fact lies between two consecutive partial sums. Note also that (2.5) implies that the error made in approximating A by truncating the series $a_0 + a_1 + \cdots$ is numerically less than and has the same sign as the first neglected term. Following Pólya and Szegö [10, p. 33], we say that the series $a_0 + a_1 + \cdots$ envelops A in the *strict sense* if (2.5) holds. The terms of a strictly enveloping series have necessarily alternating signs.

By noting that if $f(t)$ is a nonconstant completely monotonic function over $[0, \infty)$, then $|f(t)|, |f'(t)|, |f''(t)|, \cdots$ are strictly decreasing in the interval $(0, t)$, $t > 0$, we have the following result (see Pólya and Szegö [10, Prob. 140, p. 33]):

LEMMA 1. *If $f(t)$ is a nonconstant completely monotonic function over* $[0, \infty)$, *then $f(t)$ is enveloped in the strict sense by its Maclaurin series for* $0 < t < \infty$.

These results are, for the present, all we shall need.

3. The ratio of two gamma functions can be defined by the following integral representation (see [8, p. 118]):

$$(3.1) \qquad \frac{\Gamma(x+a)}{\Gamma(x+b)} = \frac{1}{\Gamma(b-a)} \int_0^\infty e^{-xt} e^{-at} (1 - e^{-t})^{b-a-1} \, dt, \qquad x + a > 0, \quad b - a > 0.$$

The variable x will be assumed to be large and positive. The case of large negative x can be reduced to the above form by using the relationship $\Gamma(x)\Gamma(1-x)=\pi/\sin \pi x$. To consider a one parameter family of asymptotic expansions of the left-hand side of (3.1), let c be an arbitrary constant and rewrite (3.1) as

$$(3.2)\quad \frac{\Gamma(x+a)}{\Gamma(x+b)}=\frac{1}{\Gamma(b-a)}\int_0^\infty e^{-t(x+c)}t^{-(a-b+1)}G^{[c]}(t)\,dt, \qquad x+a>0,\quad b-a>0,$$

where

$$(3.3)\quad \begin{aligned} G^{[c]}(t)&=e^{t(c-b+1)}\left(\frac{t}{e^t-1}\right)^{a-b+1}\\ &=\sum_{j=0}^{\infty}\frac{t^j}{j!}B_j^{(a-b+1)}(c-b+1), \qquad |t|<2\pi, \end{aligned}$$

from (1.2). Using (3.3) and Watson's lemma in (3.2), and then employing the method of extraction of the singular part to remove the restriction on a and b (see [8, p. 119]), we obtain

$$(3.4)\quad \begin{aligned} \frac{\Gamma(x+a)}{\Gamma(x+b)}&=\sum_{j=0}^{m-1}\frac{\Gamma(b-a+j)}{\Gamma(b-a)j!}B_j^{(a-b+1)}(c-b+1)(x+c)^{a-b-j}\\ &\quad+R_m^{[c]}(a,b;x), \qquad x+\min\{a,c\}>0, \end{aligned}$$

where

$$(3.5)\quad R_m^{[c]}(a,b;x)=O((x+c)^{a-b-m}), \qquad x\to\infty,$$

for all $m\geqq 1$. With the relationship

$$(3.6)\quad B_j^{(\sigma)}(\sigma-d)=(-1)^jB_j^{(\sigma)}(d),$$

we note that (3.4) becomes a real version of the Tricomi-Erdélyi expansion (1.1) for $c=0$, and that it becomes a real version of Fields' expansion (1.3) for $c=(a+b-1)/2$.

To close this section, we mention two special cases of (3.4). If $a-b+1=m$, where m is a positive integer, then $R_m^{[c]}(a,b;x)$ is zero and (3.4) is exact. Also if $a-b=-p$, p a positive integer, and $|x+c|>|b-c+1|$, then with $m\to\infty$ the series in (3.4) is convergent and sums to $\{(x+a)(x+a+1)\cdots(x+a+p-1)\}^{-1}$.

4. In this section we obtain the main result of this paper - a proof of Luke's assertion about Fields' expansion. First assume that $|b-a|$ is not a nonnegative integer; for this case see the remarks at the end of the previous section. By using the functional relation $x\Gamma(x)=\Gamma(x+1)$ for the gamma function, as Luke noted, it is always possible to reduce consideration of $\Gamma(x+a)/\Gamma(x+b)$ to the case where $0<a-b+1<1$. Putting $c=(a+b-1)/2$ in (3.2) and (3.3) yields

$$(4.1)\quad \begin{aligned} \frac{\Gamma(x+a)}{\Gamma(x+b)}=\frac{1}{\Gamma(b-a)}\int_0^\infty e^{-t(x+(a+b-1)/2)}t^{-(a-b+1)}G^{[(a+b-1)/2]}(t)\,dt, \qquad x+a>0,\\ 0<a-b+1<1, \end{aligned}$$

where

$$(4.2)\quad \begin{aligned} G^{[(a+b-1)/2]}(t)&=e^{t((a-b+1)/2)}\left(\frac{t}{e^t-1}\right)^{a-b+1}\\ &=\left(\frac{\sinh t/2}{t/2}\right)^{-(a-b+1)}. \end{aligned}$$

Substituting the Maclaurin series of $G^{[(a+b-1)/2]}(t)$ into (4.1), requiring $x+\min\{a, (a+b-1)/2\}>0$ and using Watson's lemma then gives Fields' expansion (1.3). We now need the following result.

LEMMA 2. *The function* $h(t)=(\sinh\sqrt{t}/\sqrt{t})^{-\alpha}$ *is completely monotonic over* $[0,\infty)$ *for* $\alpha>0$.

Proof. Recall the well-known infinite product expansion of $\sin z/z$, valid for all complex z (see [11, p. 114]):

$$\frac{\sin z}{z}=\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2\pi^2}\right). \tag{4.3}$$

Put $z=i\sqrt{t}$, $0\leqq t<\infty$, in (4.3) to obtain

$$\frac{\sinh\sqrt{t}}{\sqrt{t}}=\prod_{n=1}^{\infty}\left(1+\frac{t}{n^2\pi^2}\right). \tag{4.4}$$

Since $0\leqq t<\infty$ the product in (4.4) is strictly positive, and consequently

$$h(t)=\left(\frac{\sinh\sqrt{t}}{\sqrt{t}}\right)^{-\alpha}=\prod_{n=1}^{\infty}\left(1+\frac{t}{n^2\pi^2}\right)^{-\alpha},\qquad 0\leqq t<\infty. \tag{4.5}$$

It is easy to show that $h(t)$ may be differentiated to yield

$$h'(t)=-\alpha h(t)\sum_{n=1}^{\infty}\frac{1}{t+n^2\pi^2},\qquad 0\leqq t<\infty. \tag{4.6}$$

Now the sum in (4.6) defines a function $g(t)$ which is completely monotonic over $[0,\infty)$;

$$g(t)=\sum_{n=1}^{\infty}\frac{1}{t+n^2\pi^2}, \tag{4.7}$$

and so $(-1)^k g^{(k)}(t)\geqq 0$, $k=0,1,2,\cdots,0\leqq t<\infty$.

To prove that $h(t)$ is completely monotonic, we proceed by induction. From (4.5) and (4.6), $(-1)^k h^{(k)}\geqq 0$ for $k=0$ and $k=1$. Now suppose that $(-1)^k h^{(k)}(t)\geqq 0$ for $k=0, 1, 2, \cdots, n$. By Leibniz's rule,

$$\begin{aligned} h^{(n+1)}(t)&=(-\alpha h(t)g(t))^{(n)}\\ &=-\alpha\sum_{j=0}^{n}\binom{n}{j}h^{(j)}(t)g^{(n-j)}(t)\\ &=-\alpha(-1)^n\sum_{j=0}^{n}\binom{n}{j}|h^{(j)}(t)g^{(n-j)}(t)|,\end{aligned} \tag{4.8}$$

where we have used the inductive hypothesis and the complete monotonicity of $g(t)$. Equation (4.8) implies that $(-1)^{n+1}h^{(n+1)}(t)\geqq 0$, $0\leqq t<\infty$, thus completing the proof. □

Lemma 1 now implies that $h(t)$ is enveloped in the strict sense by its Maclaurin series for all $0<t<\infty$. Upon replacing t by t^2 and noting that $0<t^2<\infty$, we see that the even function $h(t^2)=(\sinh t/t)^{-\alpha}$ is also strictly enveloped by its Maclaurin series for $0<t<\infty$. It now follows immediately from (4.2) that $G^{[(a+b-1)/2]}(t)$ is strictly enveloped by its Maclaurin series when $0<a-b+1<1$. Consequently, (4.2), (1.2), (3.6) and (2.5) combine to give

$$\begin{gathered} G^{[\sigma]}(t)-\sum_{j=0}^{n}a_j t^{2j}=\theta_n(t)a_{n+1}t^{2n+2},\\ \sigma=\frac{a+b-1}{2},\quad \rho=\frac{a-b+1}{2},\quad a_j=\frac{B_{2j}^{(2\rho)}(\rho)}{(2j)!}\\ 0<\theta_n(t)<1,\quad n=1,2,3,\cdots,\quad 0<t<\infty,\quad 0<a-b+1<1.\end{gathered} \tag{4.9}$$

Multiplying (4.9) by $e^{-t(x+\sigma)}t^{-2p}/\Gamma(b-a)$, integrating from zero to infinity with respect to t and using (4.1) then gives

$$\begin{gathered}\frac{\Gamma(x+a)}{\Gamma(x+b)}-\sum_{j=0}^{n} b_j(x+\sigma)^{a-b-2j}=\frac{a_{n+1}}{\Gamma(b-a)}\int_0^\infty e^{-t(x+\sigma)}t^{2n+2-2p}\theta_n(t)\,dt,\\ b_j=\frac{\Gamma(b-a+2j)}{\Gamma(b-a)}a_j,\quad x+\min\{a,\sigma\}>0,\quad n=1,2,3,\cdots,\quad 0<a-b+1<1.\end{gathered}\tag{4.10}$$

Denoting the left-hand side of (4.10) by I_n, we see that

$$|I_n|<\frac{|a_{n+1}|}{\Gamma(b-a)}\int_0^\infty e^{-t(x+\sigma)}t^{2n+1+b-a}\,dt=|b_{n+1}|(x+\sigma)^{a-b-2n-2}.\tag{4.11}$$

Here we have used the fact that $\theta_n(t)$ is continuous on $(0,\infty)$. Thus, the asymptotic series $\sum_{j=0}^{\infty} b_j(x+\sigma)^{a-b-2j}$ envelops $\Gamma(x+a)/\Gamma(x+b)$. Moreover, as

$$\begin{aligned}\text{Sign}\,(I_n)&=\text{Sign}\left(\frac{a_{n+1}}{\Gamma(b-a)}\int_0^\infty e^{-t(x+\sigma)}t^{2n+1+b-a}\,dt\right)\\ &=\text{Sign}\,(b_{n+1}(x+\sigma)^{a-b-2n-2}),\end{aligned}\tag{4.12}$$

it follows from Pólya and Szegö [10, Prob. 144, p. 33] that $\Gamma(x+a)/\Gamma(x+b)$ is enveloped in the strict sense by the given asymptotic series. This proves Luke's assertion about Fields' expansion.

5. We conclude by using the notion of complete monotonicity to establish error bounds for the one parameter family of asymptotic approximations obtained from (3.2). Write $G^{[c]}(t)$, given in (3.3), as

$$G^{[c]}(t)=\left(\frac{e^{((a-c)/(a-b+1))t}-e^{((a-c)/(a-b+1)-1)t}}{t}\right)^{-(a-b+1)}.\tag{5.1}$$

We shall need the following result.

LEMMA 3. *The function* $f(t)=e^{-c_1t}-e^{-c_2t}/t$ *is completely monotonic over* $[0,\infty)$ *if* $-c_2<-c_1\leqq 0$.

Proof. By straightforward calculation one finds

$$f^{(k)}(t)=\frac{(-1)^k k!}{t^{k+1}}\{r_k(c_1t)-r_k(c_2t)\}\tag{5.2}$$

where

$$r_k(t)=e^{-t}\sum_{j=0}^{k}\frac{t^j}{j!}.\tag{5.3}$$

It is easily shown that $r_k(t)$ is strictly decreasing on $0<t<\infty$, and since $0\leqq c_1t<c_2t$ it follows from (5.2) that $(-1)^kf^{(k)}(t)\geqq 0$, $k=0,1,2,\cdots,0\leqq t<\infty$. Consequently $f(t)$ is completely monotonic over $[0,\infty)$. □

The case of interest to us is that of a completely monotonic function raised to a positive power (see (5.1)), so we assume $a-b+1<0$. To use Lemma 3 on $G^{[c]}(t)$ we must require $(a-c)/(a-b+1)\leqq 0$, or since $a-b+1<0$, $c\leqq a$. Under these assumptions (5.1) then represents a completely monotonic function raised to a positive power. If $-(a-b+1)$ is a positive integer and $c\leqq a$, then $G^{[c]}(t)$ is a completely monotonic function raised to a positive integer power and so is again completely monotonic. (The product of completely monotonic functions is completely monotonic.) In this case, by Lemma 1, $G^{[c]}(t)$ is strictly enveloped by its Maclaurin series for $0<t<\infty$, and proceeding as in § 4, we obtain

$$\frac{\Gamma(x+a)}{\Gamma(x+b)}-\sum_{j=0}^{n}\frac{\Gamma(b-a+j)}{\Gamma(b-a)j!}B_j^{(a-b+1)}(c-b+1)(x+c)^{a-b-j}$$

$$=\theta_n\frac{\Gamma(b-a+n+1)}{\Gamma(b-a)(n+1)!}B_{n+1}^{(a-b+1)}(c-b+1)(x+c)^{a-b-n-1}, \tag{5.4}$$

$$0<\theta_n<1,\quad n=0,1,2,\cdots,\quad -(a-b+1)=p,\ p \text{ a positive integer},\ c\leqq a,\ x+c>0.$$

We see from (5.4) that in this case the error made by truncating the expansion has the same sign and is numerically less than the first term neglected.

Suppose $-(a-b+1)$ is positive and not an integer. Lorch and Newman [5, p. 45] have shown that if $-(a-b+1)>1$, then $(f(t))^{-(a-b+1)}$ is monotonic of order 5 (at least) when $f(t)$ is completely monotonic. Coupling their result with (5.1) and Lemma 3 yields the following result:

$$\frac{\Gamma(x+a)}{\Gamma(x+b)}-\sum_{j=0}^{n}\frac{\Gamma(b-a+j)}{\Gamma(b-a)j!}B_j^{(a-b+1)}(c-b+1)(x+c)^{a-b-j}$$

$$=\theta_n\frac{\Gamma(b-a+n+1)}{\Gamma(b-a)(n+1)!}B_{n+1}^{(a-b+1)}(c-b+1)(x+c)^{a-b-n-1}, \tag{5.5}$$

$$0<\theta_n<1,\quad n=0,1,2,3,\quad -(a-b+1)>1,\quad c\leqq a,\quad x+c>0.$$

Equation (5.5) implies that for the first four terms of the asymptotic approximation given there, the error has the same sign and is numerically less than the first term neglected. The above results, in certain parameter ranges of a, b which can be obtained by using the functional equation for the gamma function, can be used to provide error bounds for the Tricomi-Erdélyi expansion (1.1) where $c=0$.

It is natural to ask if the function $G^{[c]}(t)$ in (5.1) is completely monotonic when $-(a-b+1)$ is positive, not an integer, and $c\leqq a$. It can be shown by using some further results on completely monotonic functions that there exist positive noninteger $-(a-b+1)>1$ such that $G^{[c]}(t)$ is not completely monotonic over $[0,\infty)$. To determine the order of monotonicity of $G^{[c]}(t)$ for general $c\leqq a$ and $-(a-b+1)$ is a difficult question which, in view of the results in § 4, will not be pursued here.

Finally we remark that the results in § 4 and this section can be used to provide some new inequalities for the ratio of two gamma functions, but we make no attempt here to compare them with previous work in this area.

Acknowledgments. I am grateful to Professor R. Wong for mentioning this problem to me, and also would like to thank the referees for pointing out the one-parameter family of asymptotic expansions given in (3.2) and (3.3) and for correcting an error in the original derivation of (4.10).

REFERENCES

[1] K. O. Bowman and L. R. Shenton, *Asymptotic series and Stieltjes continued fractions for a gamma function ratio*, J. Comput. Appl. Math., 4 (1978), pp. 105-111.

[2] J. Dubourdieu, *Sur un théorème de M. S. Bernstein relatif à la transformation de Laplace-Stieltjes*, Composito Math., 7 (1939), pp. 96-111.

[3] J. L. Fields, *A note on the asymptotic expansion of a ratio of gamma functions*, Proc. Edinburgh Math. Soc., 15 (1966), pp. 43-45.

[4] J. L. FIELDS, *The uniform asymptotic expansion of a ratio of gamma functions*, Proc. Internat. Conf. on Constructive Function Theory, Varna, May 19-25, 1970.

[5] L. LORCH AND D. J. NEWMAN, *On the composition of completely monotonic functions and completely monotonic sequences and related questions*, J. London Math. Soc., 28 (1983), pp. 31-45.

[6] Y. L. LUKE, *On the ratio of two gamma functions*, Jñānābha, 9/10 (1980), pp. 143-148.

[7] J. H. MCCABE, *On an asymptotic series and corresponding continued fraction for a gamma function ratio*, J. Comput. Appl. Math., 9 (1983), pp. 125-130.

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

[9] G. M. PHILLIPS AND B. N. SAHNEY, *An error estimate for least squares approximation*, BIT, 15 (1975), pp. 426-430.

[10] G. PÓLYA AND G. SZEGÖ, *Problems and Theorems in Analysis* 1, Springer-Verlag, New York, 1972.

[11] E. C. TITCHMARSH, *The Theory of Functions*, Oxford University Press, London, 1939.

[12] F. G. TRICOMI AND A. ERDÉLYI, *The asymptotic expansion of a ratio of gamma functions*, Pacific J. Math., 1 (1951), pp. 133-142.

[13] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

001

ASYNCHRONOUS EXPONENTIAL GROWTH IN TRANSITION PROBABILITY MODELS OF THE CELL CYCLE*

G. F. WEBB† AND A. GRABOSCH‡

Abstract. An analysis is given of the population dynamics of the transition probability model of the cell cycle. The model incorporates a probabilistic phase of indeterminate duration and a deterministic phase of constant duration. Individual cells increase in size according to a prescribed growth law. Sufficient conditions are established for the population to have the property of asynchronous exponential growth. The methods of proof use the theory of positive operators and semigroups of operators.

Key words. functional equation, semigroup of linear operators, positive operators, spectral bound, Banach lattice, exponential steady state

AMS(MOS) subject classifications. 92A15, 34K30

1. Introduction. During recent years an effort has been made to give a mathematical description of the cell cycle and the kinetics of proliferating cell populations. There are many unanswered questions concerning the biological processes occurring between cell birth and cell division. It is recognized that passage through the cell cycle divides into several phases. Typically there is an interphase after cell birth (G_1), a period of DNA synthesis (S), a second interphase (G_2) and a mitotic phase (M). A variety of models have been proposed to explain when and why a cell progresses from one phase into the next. Some models require cells to spend a fixed period of time in one or more phases. Others require that cell size reach a critical value. Other models concern inherited properties of cell descendants. Still others involve random transitions. Some recent mathematical treatments of such processes in cell population models have been given by Diekmann et al. [3], [4], Gyllenberg [9], Gyllenberg and Heijmans [10], Hannsgen et al. [11], Hannsgen and Tyson [12], Heijmans [13], [14], Jagers [15], Kimmel et al. [18], Lasota and Mackey [19], Lebowitz and Rubinow [20], Rotenberg [22], [23], Tyson and Hannsgen [28], [29] and Webb [32], [33].

In this paper we will give a rigorous mathematical analysis of what is known as the transition probability model of the cell cycle. This model was introduced by Smith and Martin in [26]. General discussions of this model can be found in the article of Brooks [1] and Brooks et al. [2]. The main idea of this model is the decomposition of the cell generation cycle into two kinetic states, a probabilistic A-phase and a deterministic B-phase. New cells enter the A-phase, from which they exit by random to the B-phase. The B-phase has a fixed duration normalized to 1. After passage through the B-phase a cell divides into two daughter cells of equal size, both of which enter the A-phase. The A-phase roughly corresponds to part of (G_1), and the B-phase to the rest of the cycle.

In this paper our main objective will be to determine growth laws for individual cells and transition probabilities for passage from A-phase to B-phase which are compatible with the asynchronous exponential growth of the total cell population. Roughly speaking, anynchronous or balanced exponential growth means that the density $n(x, t)$ of the total cell population at time t (with respect to some observed property x of individual cells) is asymptotic to $e^{\lambda_0 t} n_0(x)$, where λ_0 is the intrinsic

* Received by the editors August 12, 1985; accepted for publication June 19, 1986.

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

‡ Institut für Mathematik, Universität Tübingen, D-7400 Tübingen, West Germany.

growth constant and $n_0(x)$ is the exponential steady state. The model we analyze has been discussed by Hannsgen et al. in [11]. In their paper they study the exponential steady state under general growth laws and transition probabilities. One of their conclusions is that no such exponential steady state exists if individual cells satisfy an exponential growth law. In [12] Hannsgen and Tyson show that if individual cells satisfy a linear growth law and the transition probability corresponds to a single random event, then an exponential steady state exists and the population converges to it in a weak sense. In [33] this convergence is shown to hold in a strong sense. In [6] this result is extended to a general growth law.

There are two essential elements in our model. The first is a growth function g which describes completely the growth of individual cells. The size of individual cells (or of some cell component) increases with time. In both phases of the cell cycle individual cells obey the same growth law. We suppose the following hypothesis on g:

(1.1) g is continuously differentiable on $[0,\infty)$, there exist constants $\underline{g}$ and $\bar{g}$ such that $0<\underline{g}\leqq g(x)\leqq \bar{g}$ for all $x\geqq 0$, and there exists a constant $\delta>0$ such that $2g(x)-g(2x)>\delta$ for all $x\geqq 0$.

The assumption $2g(x)-g(2x)>\delta$ was made by Heijmans in [14] for an age-size structured model of cell populations. The weaker assumption $2g(x)-g(2x)>0$ was made by Diekmann et al. in [3] and [4] for a size structured model. Both assumptions rule out the exponential growth function $g(x)=kx$. The second essential element of our model is a cumulative probability distribution function f. The cell generation time T of individual cells can be described as a random variable satisfying $\Pr\{T>t\}=\int_t^\infty f(s)\,ds$. We suppose the following hypothesis on f:

(1.2) f is a continuous function on $[0,\infty)$, $f(s)=0$ for $0\leqq s\leqq 1, f(s)>0$ for $s>1, \int_1^\infty f(s)\,ds=1$, and there exist positive constants M and p such that $|f(s)|\leqq M e^{-ps}$, $s\geqq 0$.

We define $M(t,x)$ as *the size of a cell which at t time units before had size x.* $M(t,x)$ is the solution of the initial value problem

$$\frac{\partial}{\partial t}M(t,x)=g(M(t,x)), \qquad M(0,x)=x\geqq 0. \tag{1.3}$$

$M(1,0)$ is the minimum division size (since 1 unit of time must be spent in the B-phase), $\frac{1}{2}M(1,0)$ is the minimum birth size, $\frac{1}{2}M(1,\frac{1}{2}M(1,0))$ is the minimum birth size in the second generation, and so forth. We define $T(x,y)$ as *the time required for a cell to grow from size x to size y.* In terms of g

$$T(x,y)=\int_x^y \frac{1}{g(u)}\,du, \qquad 0\leqq x\leqq y. \tag{1.4}$$

We define $m(t,x)$ as *the size of a cell which grows in t time units to size x.* For $t<T(0,x)$, $m(t,x)$ is the backward solution of (1.3), which means that

$$\frac{\partial}{\partial t}m(t,x)=-g(m(t,x)), \qquad m(0,x)=x\geqq 0. \tag{1.5}$$

We collect some basic properties of M, m and T:

(1.6) $y=M(t,x)$ iff $x=m(t,y)$ iff $t=T(x,y)$, $0\leqq x\leqq y$, $0\leqq t\leqq T(0,y)$,

$$(1.7)\qquad M_2(t,x)=\frac{M_1(t,x)}{g(x)}=\frac{g(M(t,x))}{g(x)},\qquad x\geqq 0,\quad t\geqq 0,$$

$$(1.8)\qquad m_2(t,x)=\frac{-m_1(t,x)}{g(x)}=\frac{g(m(t,x))}{g(x)},\qquad x\geqq 0,\quad 0\leqq t\leqq T(0,x),$$

$$(1.9)\qquad x-t\bar{g}\leqq m(t,x)\leqq x,\qquad x\geqq 0,\quad 0\leqq t\leqq T(0,x),$$

$$(1.10)\qquad x\leqq M(t,x)\leqq x+t\bar{g},\qquad x\geqq 0,\quad t\geqq 0.$$

Several interpretations of transition probability distribution functions f are given in [11]. Take $f(t)=f_A(t-1)$, $t\geqq 1$, $f(t)=0$, $0\leqq t\leqq 1$. The *two-transition probability model of Brooks et al.* [2] has $f_A(t)=(pq/(q-p))(e^{-pt}-e^{-qt})$, $p, q>0$, $p\neq q$. Here f_A is the convolution of two densities $f_1(t)=p\,e^{-pt}$, $f_2(t)=q\,e^{-qt}$, each corresponding to a single random transition. The *Kendall model* [17] has $f_A(t)=(p/(g-1)!)(pt)^{g-1}\,e^{-pt}$, $p>0$, $g\in\mathbb{N}$. This is the gamma distribution function. Cells divide as soon as a fixed number g events have occurred. These events can happen independently in any order and all with the same constant probability per unit time. The *Rahn model* [21] has $f_A(t)=gp\,e^{-pt}(1-e^{-pt})^{g-1}$, $p>0$, $g\in\mathbb{N}$. This is the Yule distribution function. The idea is the same as in the Kendall model except that the g events have to occur in a specified order. We mention that the *one-transition probability model of Smith and Martin* $f_A(t)=p\,e^{-pt}$, $p>0$, is not included in our development, since (1.2) requires f to be continuous. This case is treated in [6] and [33].

The derivation of our model is similar to the derivation in [11]. Let $\int_{x_1}^{x_2} n(x,t)\,dx$ be the rate at which cells divide with size between x_1 and x_2 at time t. Let $\int_{x_1}^{x_2}\tilde{n}(x,t)\,dx$ be the rate at which cells are born with size between x_1 and x_2 at time t. Fix $t>0$ and $x>M(1,0)$. We write a balance equation for the rate at which cells divide at time t with size between x and $x+\Delta x$. These cells were born σ time units ago, that is, at time $t-\sigma$, where $\sigma>0$. They must have had a birth size which leads to a size between x and $x+\Delta x$ at time t. Thus, they must have had a birth size between $m(\sigma,x)$ and $m(\sigma,x+\Delta x)$. Of those cells born between $t-\sigma-\Delta\sigma$ and $t-\sigma$ the probability that an individual cell has generation time between σ and $\sigma+\Delta\sigma$ is $f(\sigma)\Delta\sigma$. Thus, the rate of dividing cells at time t with size between x and $x+\Delta x$ is

$$\begin{aligned}\int_x^{x+\Delta x} n(y,t)\,dy&=\int_0^\infty\int_{m(\sigma,x)}^{m(\sigma,x+\Delta x)}\tilde{n}(y,t-\sigma)\,dy f(\sigma)\,d\sigma\\&=\int_0^\infty\int_x^{x+\Delta x}\tilde{n}(m(\sigma,u),t-\sigma)m_2(\sigma,u)\,du f(\sigma)\,d\sigma.\end{aligned}$$

Divide by Δx and let $\Delta x\to 0$ to obtain

$$n(x,t)=\int_0^\infty \tilde{n}(m(\sigma,x),t-\sigma)m_2(\sigma,x)f(\sigma)\,d\sigma.$$

For a cell dividing with size x at σ time units after birth σ satisfies $1=T(0,M(1,0))\leqq\sigma\leqq T(0,x)$. Consequently,

$$n(x,t)=\int_1^{T(0,x)}\tilde{n}(m(\sigma,x),t-\sigma)m_2(\sigma,x)f(\sigma)\,d\sigma.$$

We must describe $\tilde{n}$ in terms of n. The number of newborn cells at time t with size between x and $x+\Delta x$ is exactly twice the number of dividing cells at time t with size between $2x$ and $2x+2\Delta x$. Thus,

$$\int_x^{x+\Delta x}\tilde{n}(y,t)\,dy=2\int_{2x}^{2x+2\Delta x} n(y,t)\,dy=4\int_x^{x+\Delta x} n(2y,t)\,dy.$$

Divide by Δx and let $\Delta x \to 0$ to obtain $\tilde{n}(x, t) = 4n(2x, t)$. We are thus led to the following functional equation for the density $n(x, t)$:

(1.11)

$$n(x,t) = \begin{cases} 4\displaystyle\int_1^{T(0,x)} n(2m(\sigma,x), t-\sigma)m_2(\sigma,x)f(\sigma)\,d\sigma, & t \geqq 0, \quad x > M(1,0), \\ 0, & t \geqq 0, \quad 0 \leqq x \leqq M(1,0), \\ \phi(x,t), & t < 0, \quad x \geqq 0. \end{cases}$$

In (1.11) ϕ prescribes the distribution of dividing cells before time 0.

In order to state our main result we introduce some notation. Let $0 < \tau < p/2\bar{g}$. Define the Banach space $Y \equiv \{h \in C([0,\infty); \mathbf{R})\colon h(0) = 0$, and $\lim_{x\to\infty} e^{\tau x}h(x) = 0\}$ with norm $\|h\|_Y \equiv \sup_{x \geqq 0} e^{\tau x}|h(x)|$. Define the Banach space $X \equiv L^1((-\infty, 0]; Y)$.

PROPOSITION 1.1. *There exists a unique real solution λ_0 of the equation $1 = 2\int_1^\infty e^{-\lambda_0 s} f(s)\,ds$, there exists a unique solution x^* of the equation $1 = T(x^*/2, x^*)$, and there exists a unique function $h_0 \in Y$ satisfying the normalizing condition $1 = \int_{x^*}^\infty h_0(x)\,dx$ and the integral equation*

$$\text{(1.12)} \qquad h_0(x) = \begin{cases} 4\displaystyle\int_1^{T(x^*/2,x)} e^{-\lambda_0 s}h_0(2m(s,x))f(s)m_2(s,x)\,ds\,dx, & x \geqq x^*, \\ 0, & 0 \leqq x < x^*. \end{cases}$$

Let $\phi \in X$. There exists a unique solution of (1.11) *satisfying $n(\cdot, t) \in Y$ for $t \geqq 0$. Further, there exists a constant $c(\phi)$ (depending on ϕ) such that*

$$\text{(1.13)} \qquad \lim_{t\to\infty} \|e^{-\lambda_0 t}n(\cdot, t) - c(\phi)h_0\|_Y = 0.$$

The method of proof of Proposition 1.1 will employ the theory of semigroups of linear operators and the theory of positive operators. The application of these theories to the study of structured population dynamics has been developed by a number of authors. Our approach has been particularly influenced by Greiner [7], Diekmann et al. [3], [4] and Heijmans [13], [14]. We state the basic ideas we will use. Let $T(t)$, $t \geqq 0$ be a strongly continuous semigroup of bounded linear operators in a Banach space X and let A be its infinitesimal generator. The *spectral bound* of A is $s(A) \equiv \sup\{\operatorname{Re}\lambda\colon \lambda \in \sigma(A)\}$(see [7]). The *essential growth bound* of $T(t)$, $t \geqq 0$ is $\omega_1(A) \equiv \lim_{t\to\infty}(1/t)\log(\alpha[T(t)])$, where α is the Kuratowski measure of noncompactness (see [31]). Let L be a positive bounded linear operator in a Banach lattice X (see [25]). L is *nonsupporting* if and only if for each $\phi \in X_+$, $\phi \neq 0$, $F \in X_+^*$, $F \neq 0$, there is an integer n_1 such that for $n \geqq n_1$, $(F, L^n\phi) > 0$ (see [14]). The following result is proved in [7] and [33].

PROPOSITION 1.2. *Let $T(t)$, $t \geqq 0$ be a strongly continuous semigroup of positive bounded linear operators in the Banach lattice X.* (1) *If $\omega_1(A) < s(A)$, then $s(A) \in P\sigma(A)$ and there exists $\phi \in X_+$, $\phi \neq 0$, such that $A\phi = s(A)\phi$.* (2) *If $\omega_1(A) < s(A)$ and $s(A)$ is a simple eigenvalue of A (that is, $s(A)$ is a simple pole of $(\lambda I - A)^{-1}$ and $N(s(A)I - A)$ is one-dimensional), then $\lim_{t\to\infty} e^{-s(A)t}T(t) = P_0$, where P_0 is the projection of X onto $N(s(A)I - A)$ given by $(1/2\pi i)\int_\Gamma (\lambda I - A)^{-1}\,d\lambda$ (Γ is a positively oriented circle about $s(A)$ enclosing no other point of $\sigma(A)$).*

The following result is proved in [24] (see also [14]).

PROPOSITION 1.3. *Let L be a positive nonsupporting bounded linear operator in the Banach lattice X and let the spectral radius $r = r(L)$ be a pole of $(\lambda I - L)^{-1}$.* (1) *$r > 0$ and r is a simple eigenvalue of L.* (2) *There exists $\phi \in X_+$, $\phi \neq 0$, such that $L\phi = r\phi$ and*

$\langle F, \phi\rangle > 0$ *for all* $F \in X_+^*$, $F \neq 0$. (3) *There exists* $F \in X_+^*$, $F \neq 0$, *such that* $L^*F = rF$ *and* $\langle F, \phi\rangle > 0$ *for all* $\phi \in X_+$, $\phi \neq 0$.

2. The semigroup of operators. We observe that X and Y are Banach lattices with the natural ordering. Define $G: X \to Y$ as follows: for $\phi \in X$

$$(2.1)\qquad (G\phi)(x) = \begin{cases} 4\displaystyle\int_1^{T(0,x)} \phi(-s)(2m(s,x))m_2(s,x)f(s)\,ds, & x > M(1,0), \\ 0, & 0 \leqq x \leqq M(1,0). \end{cases}$$

PROPOSITION 2.1. *G is a bounded linear operator from X to Y.*

Proof. Let $\phi \in X$. Since $T(0, M(1,0)) = 1$, $(G\phi)(x)$ is a continuous function of x. Since $m(s,x) \geqq x - s\bar{g}$,

$$\begin{aligned} e^{\tau x}|(G\phi)(x)| &= e^{\tau x}4\int_1^{T(0,x)} |\phi(-s)(2m(s,x))|\, e^{2\tau m(s,x)}\, e^{-2\tau m(s,x)} m_2(s,x)f(s)\,ds \\ &\leqq 4Me^{-\tau x}(\bar{g}/g)\int_1^{T(0,x)} \|\phi(-s)\|_Y e^{(2\tau\bar{g}-p)s}\,ds \\ &\leqq 4Me^{-\tau x}(\bar{g}/g)\|\phi\|_X. \end{aligned}$$

Thus, $\lim_{x\to\infty} e^{\tau x}|(G\phi)(x)| = 0$ and $\|G_\phi\|_Y \leqq 4M(\bar{g}/g)\|\phi\|_X$. □

The problem (1.11) can be formulated as

$$(2.2)\qquad n(t) = Gn_t, \qquad t \geqq 0, \quad n_0 = \phi$$

where $n: (-\infty, \infty) \to Y$, $n_t \in X$, $n_t(\theta) = n(t+\theta)$, $\theta \leqq 0$, $\phi \in X$ and $n(x,t) = n(t)(x)$. The proof of the following proposition follows directly from the results in [30].

PROPOSITION 2.2. *For each $\phi \in X$ there exists a unique function $n: (-\infty, \infty) \to Y$ satisfying* (2.2). *If $\phi \in X_+$, then $n_t \in X_+$ for all $t \geqq 0$. The solutions of* (2.2) *define a strongly continuous semigroup of bounded linear operators in X by the formula $T(t)\phi = n_t$. The infinitesimal generator of $T(t)$, $t \geqq 0$ is $A\phi = +\phi'$, $D(A) = \{\phi \in X$: ϕ is locally absolutely continuous, $\phi'(\theta)$ exists for almost all $\theta \leqq 0$, $\phi' \in X$, and $\phi(0) = G\phi\}$.*

PROPOSITION 2.3. *The essential growth bound of $T(t)$, $t \geqq 0$ satisfies $\omega_1(A) \leqq 0$.*

Proof. By virtue of Proposition 2.4 in [33] it suffices to show that for t sufficiently large there exists a representation $T(t) = U(t) + V(t)$, where $|U(t)| \leqq C$ for some constant C independent of t and $V(t)$ is compact. Let $\phi \in X$ and define $n_j: (-\infty, \infty) \to Y$, $j = 1, 2, 3, 4$, as follows:

$$n_1(t)(x) = \begin{cases} 4\displaystyle\int_{t-T(0,x)}^0 \phi(u)(2m(t-u,x))m_2(t-u,x)f(t-u)\,du, & t \geqq 0, \quad x \geqq 0, \\ \phi(t)(x), & t < 0, \quad x \geqq 0, \end{cases}$$

$$n_2(t)(x) = \begin{cases} 4\displaystyle\int_0^{t-1} n_1(u)(2m(t-u,x))m_2(t-u,x)f(t-u)\,du, & t \geqq 1, \quad x \geqq 0, \\ 0, & t < 1, \quad x \geqq 0, \end{cases}$$

$$n_3(t)(x) = \begin{cases} 16\displaystyle\int_{-1}^0\int_0^{w+1} \phi(w)(2m(u-w), 2m(t-u,x))m_2(u-w, 2m(t-u,w)) & \\ \quad\cdot f(u-w)m_2(t-u,x)f(t-u)\,du\,dw, & t \geqq 1, \quad x \geqq 0, \\ 0, & t < 1, \quad x \geqq 0, \end{cases}$$

$$n_4(t)(x)=\begin{cases}16\displaystyle\int_0^{t-2}\int_{w+1}^{t-1} n(w)(2m(u-w),2m(t-u,x))m_2(u-w,2m(t-u,x)) & \\ \quad\cdot f(u-w)m_2(t-u,x)f(t-u)\,du\,dw, & t\geqq 2,\quad x\geqq 0,\\ 0, & t<2,\quad x\geqq 0\end{cases}$$

(all integrals are taken as 0 whenever the lower limit of integration exceeds the upper limit of integration). For $\phi\in X$, $t\geqq 0$, $x\geqq 0$,

$$\begin{aligned}n(t)(x)&=4\int_{t-T(0,x)}^{t-1} n(u)(2m(t-u,x))m_2(t-u,x)f(t-u)\,du\\ &=n_1(t)(x)+4\int_0^{t-1} n(u)(2m(t-u,x))m_2(t-u,x)f(t-u)\,du\\ &=n_1(t)(x)+n_2(t)(x)+16\int_0^{t-1}\int_0^{u-1} n(w)(2m(u-w,2m(t-u,x)))\\ &\quad\cdot m_2(u-w,2m(t-u,x))f(u-w)\,dw\,m_2(t-u,x)f(t-u)\,du\\ &=n_1(t)(x)+n_2(t)(x)+n_3(t)(x)+n_4(t)(x).\end{aligned}$$

Define$(U(t)\phi)(\theta)(x)=\sum_{j=1}^{3} n_j(t+\theta)(x)$,$(V(t)\phi)(\theta)(x)=n_4(t+\theta)(x)$.Then,$(U(t)\phi+V(t)\phi)(\theta)(x)=n(t+\theta)(x)=(T(t)\phi)(\theta)(x)$.

Since $m(t+\theta-u,x)\geqq x-(t+\theta-u)\bar{g}$, there exists a constant C independent of t, θ and x such that

$$\begin{aligned}e^{\tau x}|n_1(t+\theta)(x)|&\leqq Ce^{-\tau x}\int_{t+\theta-T(0,x)}^{0}\|\phi(u)\|_Y e^{2r(t+\theta-u)\bar{g}}\,e^{-p(t+\theta-u)}\,du\\ &\leqq C\|\phi\|_X e^{(2\tau\bar{g}-p)(t+\theta)}.\end{aligned}$$

Consequently,

$$\begin{aligned}\int_{-\infty}^{0}\sup_{x\geqq 0} e^{\tau x}|n_1(t+\theta)(x)|\,d\theta&\leqq\int_{-\infty}^{-t}\|\phi(t+\theta)\|_Y\,d\theta+C\|\phi\|_X\int_{-t}^{0} e^{(2\tau\bar{g}-p)(t+\theta)}\,d\theta\\ &\leqq(1+C)\|\phi\|_X.\end{aligned}$$

Similarly, from (2.3) we see that there exists a constant C, which is changing, but is independent of t, θ and x, such that

$$\begin{aligned}e^{\tau x}|n_2(t+\theta)(x)|&\leqq Ce^{\tau x}\int_0^{t+\theta-1}|n_1(u)(2m(t+\theta-u,x))|\,e^{-p(t+\theta-u)}\,du\\ &\leqq Ce^{\tau x}\int_0^{t+\theta-1} e^{-2\tau m(t+\theta-u,x)}\|\phi\|_X e^{(2\tau\bar{g}-p)u}\,e^{-p(t+\theta-u)}\,du\\ &\leqq C\|\phi\|_X e^{-\tau x}\,e^{(2\tau\bar{g}-p)(t+\theta)}(t+\theta-1).\end{aligned}$$

Consequently,

$$\begin{aligned}\int_{-\infty}^{0}\sup_{x\geqq 0} e^{\tau x}|n_2(t+\theta)(x)|\,d\theta&\leqq C\|\phi\|_X\int_{1-t}^{0} e^{(2\tau\bar{g}-p)(t+\theta)}(t+\theta-1)\,d\theta\\ &\leqq C\|\phi\|_X.\end{aligned}$$

Since $2m(u-w, 2m(t+\theta-u, x)) \geqq m(u-w, 2m(t+\theta-u, x)) \geqq 2x+(w+u-2(t+\theta))\bar{g}$,

$$e^{\tau x}|n_3(t+\theta)(x)| \leqq C\,e^{\tau x}\int_{-1}^{0}\int_{0}^{w+1}\|\phi(w)\|_Y\, e^{-2\tau x}\, e^{-\tau(w+u-2(t+\theta))\bar{g}}\, e^{-p(t+\theta-w)}\,du\,dw$$

$$\leqq C\,e^{-\tau x}\|\phi\|_X\, e^{(2\tau\bar{g}-p)(t+\theta)}.$$

Consequently,

$$\int_{-\infty}^{0}\sup_{x\geqq 0} e^{\tau x}|n_3(t+\theta)(x)|\,d\theta \leqq C\|\phi\|_X.$$

Thus, $|U(t)| \leqq C$, where C is a constant independent of t.

It remains to show that for $t>2$, $V(t)$ is compact from X to X. Let M be a bounded set in X. We substitute $v=2m(u-w, 2m(t-u, x))$ in the integral which defines $n_4(t)(x)$. Since $\partial v/\partial u = 2g(m(u-w, 2m(t-u, x)))[-1+2g(m(t-u, x))/g(2m(t-u, x))]$ and $2g(y)-g(2y)>\delta>0$ for all $y\geqq 0$, we have $\partial v/\partial u \geqq 2\delta\underline{g}/\bar{g}>0$. Define $u=g(v, t, x)$ if and only if $v=2m(u-w, 2m(t-u, x))$ and observe that $\partial q/\partial v \leqq \bar{g}/2\delta\underline{g}$. Thus, for $\theta \geqq 2-t$, $x\geqq 0$

$$(V(t)\phi)(\theta)(x) = 16\int_{0}^{t+\theta-2}\int_{r(x,w,t+\theta)}^{s(x,w,t+\theta)} n(w)(v)z(v, w, t+\theta, x)\,dv\,dw$$

where $s(x, w, t+\theta)=2m(t+\theta-w-1, 2m(1, x))$, $r(x, w, t+\theta)=2m(1, 2m(t-w-1, x))$ and

$$z(v, w, t+\theta, x) = m_2(q(v, t+\theta, x)-w, 2m(t+\theta-q(v, t+\theta, x), x))f(q(v, t+\theta, x)-w)$$

$$\cdot\, m_2(t+\theta-q(v, t+\theta, x), x)f(t+\theta-q(v, t+\theta, x))\partial q(v, t+\theta, x)/\partial v.$$

Since $|n(w)(v)| \leqq e^{-\tau v}\|n(w)\|_Y = e^{-\tau v}\|Gn_w\|_Y \leqq e^{-\tau v}|G|\,\|T(w)\phi\|_X$, $(V(t)\phi)(\theta)(x)$ is equicontinuous in θ for $\theta\in[2-t, 0]$ and in x for x in bounded intervals independently of $\phi\in M$. We need only show that $\lim_{x\to\infty} e^{\tau x}|(V(t)\phi)(\theta)(x)|=0$ uniformly for $\phi\in M$, $\theta\in[2-t, 0]$. Since $r(x, w, t+\theta)\geqq 4(x-(t+\theta-w-\frac{1}{2})\bar{g})$,

$$e^{\tau x}|(V(t)\phi)(\theta)(x)| \leqq C\,e^{\tau x}\int_{0}^{t+\theta-2}\int_{r(x,w,t+\theta)}^{s(x,w,t+\theta)} e^{-\tau v}\, e^{-p(t+\theta-w)}\,dv\,dw$$

$$\leqq C\,e^{\tau x}\, e^{-p(t+\theta)}\int_{0}^{t+\theta-2} e^{-\tau r(x,w,t+\theta)}\, e^{pw}\,dw$$

$$\leqq C\,e^{-3\tau x}$$

where C is independent of $\phi\in M$, $\theta\in[2-t, 0]$ and $x\geqq 0$. □

3. Spectral properties of the infinitesimal generator. Let $\lambda=\lambda_0$ be the unique real root of the equation $1=2\int_1^\infty e^{-\lambda s}f(s)\,ds$. Notice that λ_0 is positive. We will establish that $s(A)=\lambda_0$ is a simple eigenvalue of A. For $\lambda>2\tau\bar{g}-p$ and $h\in Y$ define $L_\lambda h$ by

$$(3.1)\quad (L_\lambda h)(x)=\begin{cases} 4\displaystyle\int_{1}^{T(0,x)} e^{-\lambda s}h(2m(s, x))m_2(s, x)f(s)\,ds, & x>M(1, 0),\\ 0, & 0\leqq x\leqq M(1, 0).\end{cases}$$

PROPOSITION 3.1. *Let $\lambda>2\tau\bar{g}-p$. (1) L_λ is a positive compact operator from Y to Y. (2) $\lambda\in P_\sigma(A)$ and $A\phi=\lambda\phi$, $\phi\neq 0$, if and only if there exists $h\in Y$, $h\neq 0$ such that $\phi(\theta)=e^{\lambda\theta}h$ and $L_\lambda h=h$.*

Proof. Obviously L_λ is positive. To prove L_λ is compact let M be a bounded subset of Y. The substitution $\sigma = 2m(s, x)$ in (3.1) yields

$$(3.2)\qquad (L_\lambda h)(x) = \frac{2}{g(x)} \int_0^{2m(1,x)} e^{-\lambda T(\sigma/2,x)} h(\sigma) f\left(T\left(\frac{\sigma}{2}, x\right)\right) d\sigma.$$

From (3.2) we see that $(L_\lambda h)(x)$ is equicontinuous in x for bounded intervals of x and for h in M. From (3.1),

$$\begin{aligned} e^{\tau x}|(L_\lambda h)(x)| &\leqq C e^{\tau x} \int_1^{T(0,x)} e^{-\lambda s} \|h\|_Y e^{-2\tau m(s,x)}\, e^{-ps}\, ds \\ &\leqq C e^{-\tau x} \int_1^{T(0,x)} e^{-(\lambda+p-2\tau\bar{g})s}\, ds \leqq C e^{-\tau x} \end{aligned}$$

where C is independent of $h \in M$ and $x \geqq M(1, 0)$. Thus, $\lim_{x\to\infty} e^{\tau x}(L_\lambda h)(x) = 0$ uniformly for $h \in M$. Hence, L_λ is compact and (1) is proved.

To prove (2) observe by Proposition 2.2 that $A\phi = \lambda\phi$ if and only if $\phi' = \lambda\phi$ and $\phi(0) = G\phi$ if and only if $\phi(\theta) = e^{\lambda\theta}\phi(0)$ and $\phi(0) = G(e^{\lambda\theta}\phi(0)) = L_\lambda\phi(0)$. □

PROPOSITION 3.2. *Let $\gamma(x) = \frac{1}{2}M(1, x)$, $x \geqq 0$. (1) γ has a unique fixed point y^* and $y^* > 0$. (2) Let $x \geqq 0$ and define $y_1 = \gamma(x)$, $y_{n+1} = \gamma(y_n)$, $n = 1, 2, \cdots$. If $0 \leqq x < y^*$, then $\{y_n\}$ increases to y^*, and if $y^* < x$, then $\{y_n\}$ decreases to y^*.*

Proof. Our proof follows [14, Lemma 6.5]. Observe that $\gamma'(x) = \frac{1}{2}(g(2\gamma(x))/g(x))$. Since $2g(x) - g(2x) > 0$, we have that $\gamma'(y) < 1$ for any fixed point y of γ. Hence, γ can have at most one fixed point. Since $y_{n+1} \leqq \frac{1}{2}(y_n + \bar{g})$, the sequence $\{y_n\}$ is bounded. If $0 \leqq x < \gamma(x)$, then $\{y_n\}$ is increasing to a fixed point of γ. If $x > \gamma(x)$, then $\{y_n\}$ decreases to a fixed point of γ. The claims (1) and (2) now follow immediately. □

PROPOSITION 3.3. *Let $\lambda > 2\tau\bar{g} - p$, let $x_1 = \frac{1}{2}M(1, 0)$, $x_{n+1} = \gamma(x_n)$, $n = 1, 2, \cdots$, and let $x^* \equiv \lim_{n\to\infty} 2x_n = 2y^*$. (1) If $h \in Y$, then $(L_\lambda^n h)(x) = 0$ for $0 \leqq x \leqq 2x_n$, $n = 1, 2, \cdots$. (2) If $h \in Y_+$ such that $h(x_0) > 0$ for some $x_0 > x^*$ and $y_1 = \frac{1}{2}M(1, x_0/2)$, $y_{n+1} = \gamma(y_n)$, $n = 1, 2, \cdots$, then $(L_\lambda^n(x) > 0$ for $2y_n \leqq x < \infty$, $n = 1, 2, \cdots$.*

Proof. By (3.1) $(L_\lambda h)(x) = 0$ for $0 \leqq x \leqq 2x_1$. By (3.2)

$$(3.3)\qquad (L_\lambda^2 h)(x) = \frac{2}{g(x)} \int_0^{2m(1,x)} e^{-\lambda T(\sigma/2,x)} (L_\lambda h)(\sigma) f\left(T\left(\frac{\sigma}{2}, x\right)\right) d\sigma.$$

Since $(L_\lambda h)(x) = 0$ for $0 \leqq x \leqq 2x_1$, $(L_\lambda^2 h)(x) = 0$ for $2m(1, x) \leqq 2x_1$, that is, for $x \leqq M(1, x_1) = 2x_2$. An induction argument proves (1) in the general case. By (3.2) $(L_\lambda h)$ $\cdot (x) > 0$ for $2m(1, x) \geqq x_0$, that is, for $x \geqq M(1, x_0/2) = 2y_1$. By (3.3), $(L_\lambda^2 h)(x) > 0$ for $2m(1, x) \geqq M(1, x_0/2)$, that is, for $x \geqq M(1, \frac{1}{2}M(1, x_0/2)) = 2y_2$. An induction argument proves (2) in the general case. □

Define $Z = \{h \in C([x^*, \infty); \mathbf{R})\colon h(x^*) = 0 \text{ and } \lim_{x\to\infty} e^{\tau x}h(x) = 0\}$ with norm $\|h\|_Z = \sup_{x \geqq x^*} e^{\tau x}|h(x)|$. For $\lambda > 2\tau\bar{g} - p$ and $h \in Z$ define $K_\lambda h$ by

$$(3.4)\qquad (K_\lambda h)(x) = \frac{2}{g(x)} \int_{x^*}^{2m(1,x)} e^{-\lambda T(\sigma/2,x)} h(\sigma) f\left(T\left(\frac{\sigma}{2}, x\right)\right) d\sigma, \qquad x \geqq x^*.$$

PROPOSITION 3.4. *Let $\lambda > 2\tau\bar{g} - p$. (1) If $h \in Y$ such that $h(x) = 0$ for $0 \leqq x \leqq x^*$, then $(L_\lambda h)|_{[x^*,\infty)} = K_\lambda h|_{[x^*,\infty)}$. (2) K_λ is a compact positive nonsupporting operator from Z to Z. (3) The spectral radius $r(K_\lambda)$ of K_λ is positive and is a pole of the resolvent of K_λ.*

Proof. We have (1) from (3.2), (3.4), and the fact $2m(1, x^*) = x^*$. That K_λ is a compact positive operator from Z to Z then follows from Proposition 3.1 (1). To prove (2) it remains to show that K_λ is nonsupporting. Let $h \in Z_+$, $h \neq 0$, $F \in F_+^*$, $F \neq 0$. There exists $x_0 > x^*$ such that $h(x_0) > 0$. Define $\{y_n\}$ as in (2) of Proposition 3.3 so that

$\{2y_n\}$ decreases to x^* and $(K_\lambda^n h)(x) > 0$ for $x \geqq 2y_n$. Let $0 \leqq q_n(x) \leqq 1$ such that $q_n \in Z$ and $q_n(x) = 1$ for $a_n \equiv y_n \leqq x \leqq b_n \equiv 2y_n + n$. Define $F_n \in Z_+^*$ by $\langle F_n, k\rangle = \langle F, k \cdot q_n\rangle$, $k \in Z$. For each $k \in Z$, $|\langle F_n - F, k\rangle| = |\langle F, k \cdot q_n - k\rangle| \leqq \|F\|_{Z^*} \|k \cdot q_n - k\|_Z \to 0$ as $n \to \infty$ (since $k(x^*) = \lim_{x\to\infty} k(x) = 0$). Assume there exists a subsequence such that $\langle F, K_\lambda^{m_n} h\rangle = 0$ for all n. Since $0 \leqq \langle F_{m_n}, K_\lambda^{m_n} h\rangle \leqq \langle F, K_\lambda^{m_n} h\rangle$, we have $\langle F_{m_n}, K_\lambda^{m_n} h\rangle = 0$ for all n. For each n there exists a nondecreasing function f_{m_n} on $[a_{m_n}, b_{m_n}]$ such that $\langle F_{m_n}, k\rangle = \int_{a_{m_n}}^{b_{m_n}} k(x)\, df_{m_n}(x)$ for all $k \in Z$ (see [27, § 6.1.4]). Since $(K_\lambda^{m_n} h)(x) >$ for $a_{m_n} \leqq x \leqq b_{m_n}$ and $\langle F_{m_n}, K_\lambda^{m_n} h\rangle = 0$, we must have $f_{m_n} \equiv 0$ on $[a_{m_n}, b_{m_n}]$. Then, $\langle F, k\rangle = \lim_{n\to\infty} \langle F_{m_n}, k\rangle = 0$ for all $k \in Z$, which is a contradiction. Thus, K_λ is nonsupporting and (2) is proved. To show that $r(K_\lambda) > 0$ let $h \in Z_+$ such that $h \neq 0$. From (3.4)

$$\int_{x^*}^{\infty} (K_\lambda h)(x)\, dx = \int_{x^*}^{\infty} \frac{2}{g(x)} \int_{x^*}^{2m(1,x)} e^{-\lambda T(\sigma/2,x)} h(\sigma) f\left(T\left(\frac{\sigma}{2}, x\right)\right) d\sigma\, dx$$

$$\begin{aligned} &= 4 \int_{x^*}^{\infty} \int_{1}^{T(x^*/2,x)} e^{-\lambda s} h(2m(s,x)) f(s) m_2(s,x)\, ds\, dx \\ (3.5) \qquad &= 4 \int_{1}^{\infty} \int_{M(s,x^*/2)}^{\infty} e^{-\lambda s} h(2m(s,x)) f(s) m_2(s,x)\, dx\, ds \\ &= \left(2 \int_1^{\infty} e^{-\lambda s} f(s)\, ds\right) \int_{x^*}^{\infty} h(y)\, dy. \end{aligned}$$

If $\|h\|_Z = 1$, then

$$\left(2 \int_1^{\infty} e^{-\lambda s} f(s)\, ds\right)^n \int_{x^*}^{\infty} h(y)\, dy = \int_{x^*}^{\infty} (K_\lambda h)^n(x)\, dx \leqq \frac{e^{-\tau x^*}}{\tau} |K_\lambda^n|.$$

Consequently, $\lim_{n\to\infty} |K_\lambda^n|^{1/n} = r(K_\lambda) > 0$. Since K_λ is compact, $r(K_\lambda)$ is a pole of the resolvent of K_λ (see [16, p. 185]). Thus, (3) is proved. □

PROPOSITION 3.5. *The spectral bound of* $T(t)$, $t \geqq 0$ *satisfies* $\omega_1(A) < s(A) = \lambda_0$.

Proof. By Propositions 1.3 (2) and 3.4 (2) there exists $h_0 \in Z_+$, $h_0 \neq 0$, such that $K_{\lambda_0} h_0 = r(K_{\lambda_0}) h_0$. From (3.5)

$$r(K_{\lambda_0}) \int_{x^*}^{\infty} h_0(x)\, dx = \left(2 \int_1^{\infty} e^{-\lambda_0 s} f(s)\, ds\right) \int_{x^*}^{\infty} h_0(y)\, dy.$$

Since $h_0 \in Z_+$, $h_0 \neq 0$, we have $r(K_{\lambda_0}) = 1$. Define $h \in Y$ by $h(x) = 0$, $0 \leqq x \leqq x^*$, $h(x) = h_0(x)$, $x > x^*$. By Proposition 3.4 (1), $L_{\lambda_0} h = h$. By Proposition 3.1 (2), $\lambda_0 \in P\sigma(A)$ and by Proposition 2.3 $\omega_1(A) \leqq 0 < \lambda_0 \leqq s(A)$. By Proposition 1.2 (1) there exists $\phi \in X_+$, $\phi \neq 0$, such that $A\phi = s(A)\phi$. By Proposition 3.1 (2), $L_{s(A)}\phi(0) = \phi(0)$, where $\phi(0) \in Y_+$, $\phi(0) \neq 0$. A calculation similar to (3.5) now shows that $2\int_1^{\infty} e^{-s(A)r} f(r)\, dr = 1$, which means $s(A) = \lambda_0$. □

PROPOSITION 3.6. *The spectral bound* $s(A) = \lambda_0$ *is a simple eigenvalue of* A.

Proof. By Proposition 3.1 (2), $\phi \in N(\lambda_0 I - A)$, $\phi \neq 0$, if and only if $\phi(\theta) = e^{\lambda_0 \theta} h$, $h \neq 0$, and $L_{\lambda_0} h = h$. By Proposition 3.3 $h(x) = 0$ for $0 \leqq x \leqq x^*$ and $h(x) > 0$ for $x > x^*$. By Proposition 3.4 (1) $K_{\lambda_0} h|_{[x^*,\infty)} = h|_{[x^*,\infty)}$. Since $r(K_{\lambda_0}) = 1$, Proposition 1.3 (1) yields that $N(I - K_{\lambda_0})$ is one-dimensional. Thus, $N(\lambda_0 I - A)$ is one-dimensional. To prove that λ_0 is a simple eigenvalue it suffices to show that if $\phi \in X$ such that $(\lambda_0 I - A)^2 \phi = 0$, then $(\lambda_0 I - A)\phi = 0$. Assume $\psi \equiv (\lambda_0 I - A)\phi \neq 0$ and $(\lambda_0 I - A)^2 \phi = 0$. Since $\psi \in N(\lambda_0 I - A)$, the argument above shows that $\psi(\theta) = e^{\lambda_0 \theta} h$, where $h(x) = 0$ for $0 \leqq x \leqq x^*$ and $h(x) > 0$ for $x > x^*$. Since $(\lambda_0 I - A)\phi = \psi$, we have that $\phi(\theta) = e^{\lambda_0 \theta} \phi(0) - \theta\, e^{\lambda_0 \theta} h$.

Since $\phi \in D(A)$, $\phi(0) = G\phi = G(e^{\lambda_0\theta}\phi(0)) - G(\theta\, e^{\lambda_0\theta}h) = L_{\lambda_0}\phi(0) + k$, where $k = -G(\theta\, e^{\lambda_0\theta}h)$. From (2.1)

$$k(x) = 4\int_1^{T(0,x)} s\, e^{-\lambda_0 s}h(2m(s,x))m_2(s,x)f(s)\, ds$$

$$= \frac{2}{g(x)}\int_{x^*}^{2m(1,x)} T\left(\frac{\sigma}{2}, x\right) e^{-\lambda_0 T(\sigma/2,x)}h(\sigma)f\left(T\left(\frac{\sigma}{2}, x\right)\right) d\sigma.$$

Thus, $k(x) = 0$ for $0 \leqq x \leqq x^*$, $(L_{\lambda_0}^n k)(x) = 0$ for $0 \leqq x \leqq x^*$, $n = 1, 2, \cdots$, and $k|_{[x^*,\infty)} \in Z_+$, $k|_{[x^*,\infty)} \neq 0$. Observe that $\phi(0) = L_{\lambda_0}^n\phi(0) + \sum_{j=0}^{n-1} L_{\lambda_0}^j k$, $n = 1, 2, \cdots$. By Proposition 3.3 (1), $\lim_{n\to\infty} (L_{\lambda_0}^n\phi(0))(x) = 0$ for $0 \leqq x < x^*$. Thus, $\phi(0)(x) = 0$ for $0 \leqq x \leqq x^*$. By Proposition 1.3 (3), there exists $F \in Z^*$ such that $K_{\lambda_0}^* F = F$ and $\langle F, l\rangle > 0$ for all $l \in Z_+$, $l \neq 0$. Then, $\langle F, \phi(0)|_{[x^*,\infty)}\rangle = \langle F, L_{\lambda_0}\phi(0)|_{[x^*,\infty)} + k|_{[x^*,\infty)}\rangle = \langle K_{\lambda_0}^* F, \phi(0)|_{[x^*,\infty)}\rangle + \langle F, k|_{[x^*,\infty)}\rangle = \langle F, \phi(0)|_{[x^*,\infty)}\rangle + \langle F, k|_{[x^*,\infty)}\rangle$. Consequently, $\langle F, k|_{[x^*,\infty)}\rangle = 0$, which is a contradiction. □

Proof of Proposition 1.1. The proof follows immediately from the preceding propositions. Let $h \in Z_+$ such that $K_{\lambda_0}h = h$ and $\int_{x^*}^\infty h(x)\, dx = 1$. Since $N(I - K_{\lambda_0})$ is one-dimensional, such an h is uniquely determined. Let $h_0(x) = 0$ for $0 \leqq x \leqq x^*$, $h_0(x) = h(x)$ for $x > x^*$ and h_0 is the unique solution of (1.12). For $\phi \in X$, $(P_0\phi)(\theta) = c(\phi)\, e^{\lambda_0\theta}h_0$, where $c(\phi)$ is a constant depending only on ϕ. Thus, $GP_0\phi = c(\phi)G(e^{\lambda_0\theta}h_0) = c(\phi)L_{\lambda_0}h_0 = c(\phi)h_0$. Proposition 1.2 then implies (1.13), since $\lim_{t\to\infty} \|e^{-\lambda_0 t}n(t) - c(\phi)h_0\|_Y = \lim_{t\to\infty} \|G(e^{-\lambda_0 t}T(t)\phi - P_0\phi)\|_Y = 0$. □

Remark 3.1. The value of $c(\phi)$ in (1.12) can be determined in terms of h_0. Define $N(t) = \int_0^\infty n(t)(x)\, dx$, $t \in \mathbf{R}$, and $\Phi(t) = \int_0^\infty \phi(t)(x)\, dx$, $t < 0$. Then

$$N(t) = \begin{cases} \displaystyle\int_{M(1,0)}^\infty 4\int_1^{T(0,x)} n(t-s)(2m(s,x))m_2(s,x)f(s)\, ds\, dx, & t \geqq 0, \\ \Phi(t), & t < 0. \end{cases}$$

A calculation similar to (3.5) shows that $N(t) = 2\int_1^\infty N(t-s)f(s)\, ds$, $t \geqq 0$. If we define $\mathscr{G}\colon \mathscr{X} \equiv L^1((-\infty, 0]; \mathbf{R}) \to \mathbf{R}$ by $\mathscr{G}\Phi = 2\int_1^\infty \Phi(-s)f(s)\, ds$, $\Phi \in \mathscr{X}$, then $N(t) = \mathscr{G}N_t$, $t \geqq 0$, and $N_0 = \Phi$. As in [30] this functional equation yields a semigroup of operators in $\mathscr{X}$ by the formula $\mathscr{T}(t)\Phi = N_t$. The infinitesimal generator of $\mathscr{T}(t)$, $t \geqq 0$ is $\mathscr{A}\Phi = \Phi'$, $D(\mathscr{A}) = \{\Phi \in X\colon \Phi' \in X \text{ and } \Phi(0) = \mathscr{G}\Phi\}$. For $t > 1$

$$(\mathscr{T}(t)\Phi)(\theta) = \begin{cases} 2\displaystyle\int_{-\infty}^0 \Phi(u)f(t+\theta-u)\, du + 2\int_0^{t+\theta-1} N(u)f(t+\theta-u)\, du, & 1-t \leqq \theta \leqq 0, \\ 2\displaystyle\int_{-\infty}^{t+\theta-1} \Phi(u)f(t+\theta-u)\, du, & -t \leqq \theta \leqq 0, \\ \Phi(t+\theta), & \theta < -t. \end{cases}$$

An argument similar to Proposition 2.3 shows that $\omega_1(\mathscr{A}) \leqq 0$. Further, $\lambda \in P\sigma(\mathscr{A})$, $\lambda > 0$, if and only if λ satisfies the characteristic equation $1 = 2\int_1^\infty e^{-\lambda s}f(s)\, ds$. If λ_0 is the unique real root of the characteristic equation, then $\lim_{t\to\infty} e^{-\lambda_0 t}\mathscr{T}(t)\Phi = \mathscr{P}_0\Phi$, where $\mathscr{P}_0 = (1/2\pi i)\int_\Gamma (\lambda I - \mathscr{A})^{-1}\, d\lambda$. For $\lambda > 0$, $\lambda \notin P\sigma(\mathscr{A})$, $(\lambda I - \mathscr{A})^{-1}\Phi = \Psi$ if and only if $\Psi(\theta) = e^{\lambda\theta}\Psi(0) + \chi(\theta)$, where $\chi(\theta) = \int_\theta^0 e^{\lambda(\theta - s)}\Phi(s)\, ds$ and $\Psi(0) = \mathscr{G}\Psi$. Thus,

$$((\lambda I - \mathscr{A})^{-1}\Phi)(\theta) = \frac{e^{\lambda\theta}\mathscr{G}\chi}{1 - 2\int_1^\infty e^{-\lambda s}f(s)\, ds} + \chi(\theta).$$

By the Residue Theorem

$$(\mathcal{P}_0\Phi)(\theta) = \frac{2\, e^{\lambda_0\theta} \int_1^\infty \int_{-s}^0 e^{-\lambda_0(s+r)}\Phi(r)\, dr\, f(s)\, ds}{2 \int_1^\infty s\, e^{-\lambda_0 s} f(s)\, ds}.$$

By (1.13)

$$\begin{aligned}\lim_{t\to\infty} \int_0^\infty e^{-\lambda_0 t} n(t)(x)\, dx &= c(\phi) \int_0^\infty G(e^{\lambda_0\theta} h)(x)\, dx \\ &= \lim_{t\to\infty} e^{-\lambda_0 t} N(t) \\ &= \mathcal{G} P_0 \Phi.\end{aligned}$$

Thus, $c(\phi) = \mathcal{G}\mathcal{P}_0\Phi / \int_0^\infty G(e^{\lambda_0\theta}h)(x)\, dx$. We observe that $\mathcal{G}\mathcal{P}_0\Phi$, the exponential steady state of the total population of dividing cells of all sizes, is independent of g, the growth law of individual cells.

REFERENCES

[1] R. F. Brooks, *Variability in the cell cycle and the control of proliferation*, in The Cell Cycle, P. John, ed., Cambridge Univ. Press, 1981, pp. 35-62.

[2] R. F. Brooks, D. C. Bennett and J. A. Smith, *Mammalian cell cycles need two random transitions*, Cell, 19 (1980), pp. 493-504.

[3] O. Diekmann, H. Heijmans and H. Thieme, *On the stability of the cell size distribution*, J. Math. Biol., 19 (1984), pp. 227-248.

[4] ———, *On the stability of the cell size distribution* II, Internat. J. Comput. Math. Appl., 12A (1986), pp. 491-512.

[5] O. Diekmann, H. Lauwerier, T. Aldenberg and J. Metz, *Growth, fission, and the stable size distribution*, J. Math. Biol., 18 (1983), pp. 135-148.

[6] A. Grabosch, *A two phase transition probability model of the cell cycle*, to appear.

[7] G. Greiner, *A typical Perron-Frobenius theorem with application to an age-dependent population equation*, in Infinite-Dimensional Systems, Proceedings, Retzhof 1983, F. Kappel and W. Schappacher, eds., Lecture Notes in Math., 1076, Springer-Verlag, Berlin, Heidelberg, New York, 1984.

[8] G. Greiner, J. Voigt and M. Wolff, *On the spectral bound of the generator of semigroups of positive operators*, J. Operator Theory, 5 (1981), pp. 245-256.

[9] M. Gyllenberg, *The size and scar distributions of the yeast saccharomyces cerevisiae*, J. Math. Biol., 24 (1986), pp. 81-101.

[10] M. Gyllenberg and H. Heijmans, *An abstract delay-differential equation modelling size dependent cell growth and division*, to appear.

[11] K. Hannsgen, J. Tyson and L. Watson, *Steady-state distributions in probabilistic models of the cell division cycle*, SIAM J. Appl. Math., 45 (1985), pp. 523-540.

[12] K. Hannsgen and J. Tyson, *Stability of the steady-state size distribution in a model of cell growth and division*, J. Math. Biol., 22 (1985), pp. 293-301.

[13] H. Heijmans, *An eigenvalue problem related to cell growth*, J. Math. Anal. Appl., 111 (1985), pp. 253-280.

[14] ———, *The dynamical behavior of the age-size-distribution of a cell population*, to appear.

[15] P. Jagers, *Balanced exponential growth: What does it mean and when is it there?* in Biomathematics and Cell Kinetics, Development in Cell Biology, Vol. 2, A. Valleron and P. Macdonald, eds., Elsevier/North-Holland, Amsterdam, 1978, pp. 21-29.

[16] T. Kato, *Perturbation Theory for Linear Operations*, Springer-Verlag, Berlin, Heidelberg, New York, 1966.

[17] D. G. Kendall, *On the role of variable generation time in the development of a stochastic birth process*, Biometrika, 35 (1948), pp. 316-330.

[18] M. Kimmel, Z. Darzynkiewicz, O. Arino and F. Traganos, *Analysis of a model of cell cycle based on unequal division of mitotic constituents to daughter cells during cytokinesis*, J. Theoret. Biol., 101 (1984).

[19] A. Lasota and M. C. Mackey, *Globally asymptotic properties of proliferating cell populations*, J. Math. Biol., 19 (1984), pp. 43-62.

[20] J. L. Lebowitz and S. L. Rubinow, *A theory for the age and generation time distribution of a microbial population*, J. Math. Biol., 1 (1974), pp. 17-36.

[21] O. RAHN, *A chemical explanation of the variability of the growth rate*, J. Gen. Physiol., 15 (1932), pp. 257-277.

[22] M. ROTENBERG, *Theory of distributed quiescent state in the cell cycle*, J. Theoret. Biol., 96 (1982), 495-509.

[23] ———, *Correlations, asymptotic stability and the G_0 theory of the cell cycle*, in Biomathematics and Cell Kinetics, Development in Cell Biology, Vol. 2, A.-J. Valleron and P. D. M. Macdonald, eds., Elsevier/North-Holland Biomedical Press, Amsterdam, 1978, pp. 59-69.

[24] I. SAWASHIMA, *On spectral properties of some positive operators*, Nat. Sci. Dep. Ochanomizu Univ., 15 (1964), pp. 53-64.

[25] H. SCHAEFER, *Banach Lattices and Positive Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1974.

[26] J. A. SMITH AND L. MARTIN, *Do cells cycle?*, Proc. Nat. Acad. Sci. U.S.A., 70 (1973), pp. 1263-1267.

[27] B. SZ.-NAZY, *Introduction to Real Functions and Orthogonal Expansions*, University Texts in the Mathematical Sciences, Oxford Univ. Press, New York, 1965.

[28] J. TYSON AND K. HANNSGEN, *The distribution of cell size and generation time in a model of the cell cycle incorporating size control and random transitions*, J. Theoret. Biol., 113 (1985), pp. 29-62.

[29] ———, *Cell growth and division: Global asymptotic stability of the size distribution in probabilistic models of the cell cycle*, to appear.

[30] G. F. WEBB, *Volterra integral equations and nonlinear semigroups*, Nonlinear Anal., 1 (1977), pp. 415-417.

[31] ———, *Theory of Nonlinear Age-Dependent Population Dynamics*, in Pure and Applied Mathematics Series of Monographs and Textbooks, Vol. 89, Marcel Dekker, New York, Basel, 1985.

[32] ———, *A model of proliferating cell populations with inherited cycle length*, J. Math. Biol., 23 (1986), pp. 269-282.

[33] ———, *An operator-theoretic formulation of asynchronous exponential growth*, to appear.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

002

ON DETERMINING THE PREDICTOR OF NONFULL-RANK MULTIVARIATE STATIONARY RANDOM PROCESSES*

A. G. MIAMEE†

Abstract. Algorithms for determining the generating function and the predictor for some nonfull-rank multivariate stationary stochastic processes are obtained. In fact, it is shown that the well-known algorithms given by Wiener and Masani [Acta Math., 99 (1958), pp. 93-137] for the full-rank case are valid in certain nonfull-rank cases exactly in the same form.

Key words. nonfull-rank multivariate stationary processes, generating function, best linear predictor

AMS(MOS) subject classification. Primary 60G10

1. Introduction. One of the important problems in the prediction theory of multivariate stationary stochastic processes is to obtain some algorithm for determining the best linear predictor in terms of the past observations. Wiener and Masani [9], [10] solved this problem for the full-rank case, when the spectral density f of the processes is bounded above and away from zero, in the sense that there exist positive numbers c and d such that

$$c\mathbf{I} \leqq \mathbf{f}(\theta) \leqq d\mathbf{I}. \tag{1.1}$$

Masani [2] improved their work substantially showing that the same algorithm is valid if in lieu of (1.1) one assumes that

$$\text{(1.2)}\qquad \begin{aligned} &\text{(i)} \quad \mathbf{f} \in \mathbf{L}_\infty; \\ &\text{(ii)} \quad \mathbf{f}^{-1} \in \mathbf{L}_1. \end{aligned}$$

Several other authors proved the validity of the same algorithm under more general settings, cf. for example Salehi [8], Pourahmadi [6]. However, all these results are under the severe restriction of full-rank and there has been no extension of Wiener and Masani's algorithm beyond the full-rank case.

The purpose of this note is to show that the algorithm remains valid exactly in the same manner for the nonfull-rank processes which satisfy the following conditions:

$$\text{(1.3)}\qquad \begin{aligned} &\text{(i)} \quad \text{The range of } \mathbf{f}(\theta) \text{ is constant a.e. Lebesgue measure,} \\ &\text{(ii)} \quad \mathbf{f} \in \mathbf{L}_\infty, \\ &\text{(iii)} \quad \mathbf{f}^{\#} \in \mathbf{L}_1, \end{aligned}$$

where $\mathbf{A}^{\#}$ stands for the generalized inverse (to be defined later) of the matrix $\mathbf{A}$. In the full-rank case these conditions clearly reduce to the conditions (1.2), and hence our result generalizes Masani's algorithm in [2].

Masani's assumption and approach rests on a characterization [2, Thm. 2.8], for full-rank minimal multivariate stationary stochastic processes. Our motivation and assumptions are based on a characterization of J_0-regularity (which we shall call "purely minimal") due to Makagon and Weron [1]. We will employ Wiener and Masani's algorithm to find the predictor of an associated full-rank process (to be clarified later), which is produced using the technique of Miamee and Salehi [5], and using this we will obtain our algorithm for the nonfull-rank process.

* Received by the editors March 18, 1985; accepted for publication (in revised form) May 20, 1986. This research was partially supported by Air Force Office of Scientific Research grant # F49620 82 C 0009.

† Isfahan University of Technology, Isfahan, Iran. Present address, Department of Mathematics, Hampton University, Hampton, Virginia 23668.

In § 2 we set down the necessary preliminaries. Section 3 is devoted to establishing our algorithm for determining the generating function and in § 4 we will show the validity of Wiener and Masani's algorithm for the best linear predictor.

2. Preliminaries. In this section we set down notation and preliminaries. Most of these are standard and can be found in [4], [9] and [10]. Let H be a complex Hilbert space and q a positive integer. H^q denotes the Cartesian product of q-copies of H, endowed with a *Gramian* structure as follows: For any two vectors $\mathbf{X}=(x^1,\cdots x^q)^T$ and $\mathbf{Y}=(y^1,\cdots,y^q)^T$, the superscript T stands for the matrix transpose, in H^q their *Gramian* matrix (X, Y) is defined by

$$(\mathbf{X},\mathbf{Y})=[(x^i,x^j)]_{i,j=1}^q.$$

It is easy to verify that it has the following properties:

$$(\mathbf{X},\mathbf{X})\geqq 0,$$

$$(\mathbf{X},\mathbf{X})=\mathbf{0}\ \Leftrightarrow\ \mathbf{X}=\mathbf{0},$$

$$\left(\sum_{i=1}^m \mathbf{A}_i\mathbf{X}_i,\ \sum_{j=1}^n \mathbf{B}_j\mathbf{X}_j\right)=\sum_{i=1}^m\sum_{j=1}^n \mathbf{A}_i(\mathbf{X}_i,\mathbf{X}_j)\mathbf{B}_j^*,$$

where $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}_i$, $\mathbf{Y}_j$ are in H^q, $\mathbf{A}_i$, $\mathbf{B}_j$ are constant $q\times q$ matrices, and $\mathbf{A}\geqq 0$ means $\mathbf{A}$ is a *nonnegative definite* matrix. We say that $\mathbf{X}$ is *orthogonal* to $\mathbf{Y}$ if $(\mathbf{X},\mathbf{Y})=\mathbf{0}$. It is well known that H^q is a Hilbert space with the *inner product*

$$((\mathbf{X},\mathbf{Y}))=\text{trace}\,(\mathbf{X},\mathbf{Y}).$$

A subset $\mathbf{M}$ of H^q is called a *subspace* if it is closed and $\mathbf{AX}+\mathbf{BY}\in\mathbf{M}$, whenever $\mathbf{X}$ and $\mathbf{Y}$ are in $\mathbf{M}$, $\mathbf{A}$ and $\mathbf{B}$ are $q\times q$ constant matrices. It is easy to see that $\mathbf{M}$ is a subspace if and only if $\mathbf{M}=M^q$ for some subspace M of H. For any $\mathbf{X}$ in H^q, $(\mathbf{X}|\mathbf{M})$ denotes the projection of $\mathbf{X}$ onto $\mathbf{M}$, and that is the vector whose kth coordinate is $(x^k|M)$, which is the usual projection of x^k onto the subspace M.

A bisequence $\mathbf{X}_n$, $n\in Z$, in H^q is called a *q-variate stationary stochastic process* if the Gramian $(\mathbf{X}_m,\mathbf{X}_n)$ depends *only* on $m-n$.

It is well known that every q-variate stationary stochastic process X_n has a nonnegative matrix valued measure $\mathbf{F}$ on $[0,2\pi]$, called its *spectral measure* such that

$$(\mathbf{X}_m,\mathbf{X}_n)=\frac{1}{2\pi}\int_0^{2\pi} e^{-i(m-n)\theta}\,d\mathbf{F}(\theta).$$

$\mathbf{f}$ stands for the Radon-Nikodym derivative of the absolutely continuous (a.c.) part of $\mathbf{F}$ with respect to the normalized Lebesgue measure $d\theta$, and it is called the *spectral density* of the process.

To every stationary stochastic process $\mathbf{X}_n$, $n\in Z$ the following subspaces are attached:

$\mathbf{M}(+\infty)=\overline{\text{sp}}\,(\mathbf{X}_n,\ -\infty<n<\infty)$, i.e., the subspace of H^q generated by all $\mathbf{X}_n$, $n\in Z$,

$$\mathbf{M}(n)=\overline{\text{sp}}\,(\mathbf{X}_k,\ -\infty<k\leqq n),$$

$$\mathbf{M}(-\infty)=\bigcap_n \mathbf{M}(n),$$

$$\mathbf{M}'(n)=\overline{\text{sp}}\,(\mathbf{X}_k,\ k\neq n).$$

A q-variate stationary stochastic process is called

(a) *Nondeterministic* if $\mathbf{M}(+\infty)\neq\mathbf{M}(n)$ for some and hence all n in Z;

(b) *Purely nondeterministic* if $\mathbf{M}(-\infty)=\mathbf{0}$;

(c) *Minimal* if $\mathbf{M}'(n)\neq\mathbf{M}(+\infty)$ for some and hence all $n\in Z$;

(d) *Purely minimal* if $\bigcap_n \mathbf{M}'(n)=\mathbf{0}$.

If $\mathbf{X}_n$ is nondeterministic then $\mathbf{X}_n \notin \mathbf{M}(n-1)$ for all n, and hence it has a nonzero *one-sided innovation process*

$$\mathbf{g}_n = \mathbf{X}_n - (\mathbf{X}_n|\mathbf{M}(n-1)).$$

If $\mathbf{X}_n$ is minimal then $\mathbf{X}_n \notin \mathbf{M}'(n)$ for all n, and hence it has a nonzero *two-sided innovation process*

$$\boldsymbol{\phi}_n = \mathbf{X}_n - (\mathbf{X}_n|\mathbf{M}'(n)).$$

The corresponding *one-sided and two-sided predictor error matrices* are defined by

$$\mathbf{G} = (\mathbf{g}_0, \mathbf{g}_0) \quad \text{and} \quad \boldsymbol{\Sigma} = (\boldsymbol{\phi}_0, \boldsymbol{\phi}_0),$$

respectively. $\hat{\mathbf{X}}_\nu = (\mathbf{X}_\nu|\mathbf{M}(0))$ is called the *best linear predictor of lag* ν. Clearly $\mathbf{X}_n$ is nondeterministic if and only if $\mathbf{G} \neq \mathbf{0}$ and minimal if and only if $\boldsymbol{\Sigma} \neq \mathbf{0}$. A nondeterministic (purely nondeterministic) process $\mathbf{X}_n$ is said to be *nondeterministic* (*purely nondeterministic*), *of full-rank* if $\mathbf{G}$ is invertible. The process is called *full-rank minimal* if it is minimal and its two-sided predictor error matrix $\boldsymbol{\Sigma}$ is invertible.

It is useful to note that we have the following inclusions between these various classes of processes.

PROPOSITION. *The following inclusions are valid*:

(a) *Purely nondeterministic processes* $\subseteq$ *nondeterministic processes*;

(b) *Purely minimal processes* $\subseteq$ *minimal processes* $\subseteq$ *nondeterministic processes*;

(c) *Purely minimal processes* $\subseteq$ *purely nondeterministic processes.*

Proof. (a) This is clear and well known. (b) If a process is purely minimal, then $\bigcap_n \mathbf{M}'(n) = \mathbf{0}$ which clearly implies that $\mathbf{M}'(n) \neq \mathbf{M}(-\infty)$ for some n, and hence the process must be minimal. This proves the first inclusion in (b). Now if a process is minimal then $\mathbf{M}'(n) \neq \mathbf{M}(+\infty)$ for some n. Since $\mathbf{M}(n-1) \subseteq \mathbf{M}'(n)$ we have $\mathbf{M}(n-1) \neq \mathbf{M}(+\infty)$. That is, the process must be nondeterministic.(c) Since $\mathbf{M}(n-1) \subseteq \mathbf{M}'(n)$ we have $\bigcap_n \mathbf{M}(n) \subseteq \bigcap_n \mathbf{M}'(n)$, which then shows that purely minimality is stronger than being purely nondeterministic.

Remarks. (a) Considering the well-known characterization for minimal and that of purely nondeterministic univariate processes, one can see that neither of these two classes is a subset of the other one.

(b) All the inclusions stated in the above proposition are proper. This for the first inclusion in (b) is shown via the next example, and for the rest of them it is either well-known or can be easily verified.

Example. Let Y_n be orthonormal and Z any vector orthogonal to all Y_n's, and let $X_n = Y_n + Z$. Then

(i) X_n is *stationary*, because $(X_m, X_n) = \delta_{m-n,0} + (Z, Z)$ depends only on $m-n$.

(ii) X_n is *minimal*, because otherwise X_0 must belong to $\overline{\text{sp}}\,\{X_k\colon k \neq 0\} = M'(0)$ and hence there exists a sequence p_n of finite sums of the form

$$p_n = \sum_{k\neq 0} a_k^n X_k = \sum_{k\neq 0} a_k^n (Y_k + Z)$$

such that $X_0 = \lim_{n\to\infty} p_n$. Now since

$$p_n = \sum_{k\neq 0} a_k^n Y_k + \left(\sum_{k\neq 0} a_k^n\right) Z = Q_n + r_n Z \quad \text{and} \quad Q_n \perp r_n Z$$

the convergence of p_n implies the convergence of both sequences Q_n and r_n to some Q and r, respectively. Taking limit on both sides of the last equation, we get

$$X_0 = Q + rZ$$

which gives

$$Y_0 - Q = (r-1)Z, \qquad Y_0 - Q \perp (r-1)Z.$$

We see that $Y_0 - Q = 0$ or $Y_0 = Q$. But since each Q_n is in $\overline{\mathrm{sp}}\{Y_k: k \neq 0\}$, we get $Y_0 \in \overline{\mathrm{sp}}\{Y_k: k \neq 0\}$ which is impossible, because Y_n is an orthonormal sequence.

(iii) X_n is *not purely minimal.* In fact, its spectral distribution $F(\theta) = \theta + \delta_0$, where δ_0 is the unit mass concentrated at 0, is not a.c. A purely time domain proof of this fact may also be given. Q.E.D.

It is known that

$$\mathbf{M}(n) = \sum_{k=0}^{\infty} \overline{\mathrm{sp}}\,(\mathbf{g}_{n-k}) + \mathbf{M}(-\infty).$$

Consider $\mathbf{G}$ as a linear operator on C^q to C^q, C being the complex plane. Let $\mathbf{J}$ be the matrix of the projection on C^q onto the range of $\mathbf{G}$, and we put $(\sqrt{\mathbf{G}} + \mathbf{J}^{\perp})^{-1} = \mathbf{H}$. The *normalized one-sided innovations* are defined by $\mathbf{h}_n = \mathbf{H}\mathbf{g}_n$. One can show that [4]

$$\mathbf{X}_n = \sum_{k=0}^{\infty} \mathbf{A}_k \sqrt{\mathbf{G}}\,\mathbf{h}_{n-k} + (\mathbf{X}_n | \mathbf{M}(-\infty)).$$

Although $\mathbf{A}_k$'s in this decomposition *are not unique*, the coefficients $\mathbf{A}_k\sqrt{\mathbf{G}}$ are in fact *unique* and this enables us to associate the following function to our process

$$\mathbf{\Phi}(e^{i\theta}) = \sum_{k=0}^{\infty} \mathbf{A}_k \sqrt{\mathbf{G}}\, e^{ik\theta};$$

this is called the *generating function* of the process.

We shall be concerned with the class $\mathbf{L}_p (1 \leqq p \leqq \infty)$ of all $q \times q$ matrix valued functions $\mathbf{g}$ on $[0, 2\pi]$ whose entries are in the usual Lebesgue space L_p. $\mathbf{L}_2^{0+}$ will denote the subspace of $\mathbf{L}_2$ consisting of those matrix valued functions whose nth Fourier coefficient vanishes for $n < 0$, i.e.,

$$\int e^{-in\theta} \mathbf{g}(\theta)\, d\theta = \mathbf{0} \quad \text{for all } n < 0.$$

For any $q \times q$ matrix $\mathbf{A}$ its *generalized inverse* $\mathbf{A}^{\#}$ is defined to be

$$\mathbf{A}^{\#} = p_{N(\mathbf{A})^{\perp}} \mathbf{A}^{-1} p_{R(\mathbf{A})},$$

where $N(\mathbf{A})$ denotes the null space of $\mathbf{A}$, $\mathbf{A}^{-1}$ denotes the (many valued) inverse of $\mathbf{A}$, and $R(\mathbf{A})$ stands for the range of $\mathbf{A}$. One can see that $\mathbf{A}^{\#}$ has the following properties [7]:

$$\mathbf{A}\mathbf{A}^{\#}\mathbf{A} = \mathbf{A}, \qquad \mathbf{A}^{\#}\mathbf{A}\mathbf{A}^{\#} = \mathbf{A}^{\#},$$
$$(\mathbf{A}^{\#}\mathbf{A})^* = \mathbf{A}^{\#}\mathbf{A}, \qquad (\mathbf{A}\mathbf{A}^{\#})^* = \mathbf{A}\mathbf{A}^{\#},$$
$$N(\mathbf{A})^{\perp} = R(\mathbf{A}^{\#}), \qquad R(\mathbf{A})^{\perp} = N(\mathbf{A}^{\#}).$$

For the ease of reference we state the following two theorems which are due to Masani [2, Thm. 2.8] and to Makagon and Weron [1], respectively.

THEOREM 1. *Let* $\mathbf{X}_n$, $n \in Z$, *be a q-variate stationary stochastic process with spectral distribution* $\mathbf{F}$. $\mathbf{X}_n$ *is full-rank minimal if and only if* $\mathbf{f} = \mathbf{F}'$ *is invertible and* $\mathbf{f}^{-1} \in \mathbf{L}_1$.

THEOREM 2. *Let* $\mathbf{X}_n$, $n \in Z$ *be a q-variate stationary stochastic process with spectral measure* $\mathbf{F}$. *The process* $\mathbf{X}_n$ *is purely minimal if and only if*

(i) $\mathbf{F}$ *is a.c. with respect to Lebesque measure*; *with spectral density* $\mathbf{f}$,

(ii) $R(\mathbf{f}(\theta))$ *is constant a.e. Lebesque measure,*

(iii) $\mathbf{f}^{\#} \in \mathbf{L}_1$.

3. Determination of the generating function. In this section we give an algorithm for determining the generating function of a (not necessarily full-rank) stationary stochastic process. The result of this section extends Masani's algorithm developed in [2] to the nonfull-rank case. Our technique is essentially that used by Miamee and Salehi in [5], where the following formula for the two-sided prediction error matrix $\mathbf{\Sigma}$ of a purely minimal process was obtained:

$$\mathbf{\Sigma}=\left[\frac{1}{2\pi}\int_0^{2\pi}\mathbf{f}^{\#}(\theta)\,d\theta\right]^{\#}.$$

We will continue this work under the assumption that our process is purely minimal or equivalently assuming that conditions (i), (ii) and (iii) of Theorem 2 are valid. Let $\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p, \mathbf{h}_{p+1}, \cdots, \mathbf{h}_q$ be an orthonormal basis for the q-dimensional complex Euclidean space C^q such that

$$R=R(\mathbf{f}(\theta))=\overline{\mathrm{sp}}\,(\mathbf{h}_i, 1\leqq i\leqq p)\quad \text{a.e.}\,(d\theta),$$

and

$$N=R^{\perp}=N(\mathbf{f}(\theta))=\overline{\mathrm{sp}}\,(\mathbf{h}_i, p+1\leqq i\leqq q).$$

Let $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_q$ be the standard basis of C^q. Define the unitary operator $\mathbf{U}$ on C^q by $\mathbf{U}\mathbf{h}_i=\mathbf{e}_i$, $1\leqq i\leqq q$. Letting $R_1=\overline{\mathrm{sp}}\,(\mathbf{e}_i, 1\leqq i\leqq p)$ then $R_1^{\perp}=\overline{\mathrm{sp}}\,(\mathbf{e}_i, p+1\leqq i\leqq q)$. Clearly $\mathbf{U}$ maps R onto R_1 and $R^{\perp}$ onto $R_1^{\perp}$ and $\mathbf{U}^*$ maps R_1 onto R and $R_1^{\perp}$ onto $R^{\perp}$. As usual we will identify any linear operator on C^q with its matrix with respect to the standard basis of C^q. By our choice of $\mathbf{U}$ we have

$$\mathbf{U}\mathbf{f}(\theta)\mathbf{U}^*=\begin{bmatrix}\mathbf{g}(\theta) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\tag{3.1}$$

where $\mathbf{g}(\theta)$ is a $p\times p$ nonnegative matrix valued function whose rank is a.e. equal to p. Let

$$\mathbf{Y}_n=\mathbf{U}\mathbf{X}_n,\qquad n\in Z$$

be a new stationary stochastic process, then we have

$$\begin{aligned}(\mathbf{Y}_m, \mathbf{Y}_n)=(\mathbf{U}\mathbf{X}_m, \mathbf{U}\mathbf{X}_n)&=\mathbf{U}\left(\frac{1}{2\pi}\int_0^{2\pi}e^{-i(m-n)\theta}\mathbf{f}(\theta)\,d\theta\right)\mathbf{U}^*\\ &=\frac{1}{2\pi}\int_0^{2\pi}e^{-i(m-n)\theta}\mathbf{U}\mathbf{f}(\theta)\mathbf{U}^*\,d\theta\\ &=\frac{1}{2\pi}\int_0^{2\pi}e^{-i(m-n)\theta}\begin{bmatrix}\mathbf{g}(\theta) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}d\theta.\end{aligned}\tag{3.2}$$

This shows that, for $p+1\leqq k\leqq q$, the kth component Y_n^k of $\mathbf{Y}_n$ is zero for all $n\in Z$. The p-variate stationary stochastic process $\mathbf{Z}_n=(Y_n^1, \cdots, Y_n^p)^T$ has spectral density $\mathbf{g}$. Since $\mathbf{U}$ takes R onto R_1 and $R^{\perp}$ onto $R_1^{\perp}$, one can see that

$$\begin{bmatrix}\mathbf{g}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}=\begin{bmatrix}\mathbf{g} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}^{\#}=(\mathbf{U}\mathbf{f}\mathbf{U}^*)^{\#}=\mathbf{U}^*\mathbf{f}^{\#}\mathbf{U}.\tag{3.3}$$

Now since $\mathbf{X}_n$ is assumed to be purely minimal, Theorem 2 implies that $\mathbf{f}^{\#}(\theta)$ is integrable. Thus (3.2) implies that $\mathbf{g}^{-1}$ is integrable and hence by Theorem 1, $\mathbf{Z}_n$ is full-rank minimal.

We are going to utilize Masani's algorithm to obtain the generating function $\mathbf{\Psi}$ and predictor $\hat{\mathbf{Z}}_\nu$ of this full-rank minimal process $\mathbf{Z}_n$, and then use this to get the generating function $\mathbf{\Phi}$ and predictor $\hat{\mathbf{X}}_\nu$ of our process $\mathbf{X}_n$. The following lemma, which reveals the close tie between $\mathbf{\Psi}$ and $\mathbf{\Phi}$, is crucial in the development of our algorithm.

LEMMA. *Let $\mathbf{X}_n$, $n \in Z$ be a purely minimal stationary stochastic process with spectral density* $\mathbf{f}$. *Let* $\mathbf{g}$ *be the spectral density of the corresponding full-rank minimal process* $\mathbf{Z}_n$ *discussed above. If* $\mathbf{\Phi}$ *and* $\mathbf{\Psi}$ *are the generating functions of* $\mathbf{X}_n$ *and* $\mathbf{Z}_n$, *respectively, then*

$$\mathbf{\Phi} = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U},$$

where $\mathbf{U}$ *is the unitary matrix obtained above.*

Proof. We first note that, since $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as generating functions are optimal (cf. Lemma 3.7 and Definition 4.1 in [3]). Now from (3.1) we get

$$\mathbf{f} = \mathbf{U}^* \begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U} = \left(\mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U} \right) \left(\mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U} \right)^* \tag{3.4}$$

so $\mathbf{f}$ has two factorization, namely the one in (3.4) and

$$\mathbf{f} = \mathbf{\Phi}\mathbf{\Phi}^*,$$

where both $\mathbf{\Phi}$ and

$$\mathbf{\delta} = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}$$

belong to $\mathbf{L}_2^{0+}$; to complete the proof it suffices to show that the factor $\mathbf{\delta}$ is also optimal (cf. uniqueness Theorem 4.4 of [3]). To prove this, we first note that since the zeroth coefficient $\mathbf{\Psi}_+(0)$ of $\mathbf{\Psi}$ is nonnegative definite and

$$\mathbf{\delta}_+(0) = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi}_+(0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}$$

we have

$$\mathbf{\delta}_+(0) \geqq \mathbf{0}. \tag{3.5}$$

On the other hand, if

$$\mathbf{f} = \mathbf{\gamma}\mathbf{\gamma}^*, \qquad \mathbf{\gamma} \in \mathbf{L}_2^{0+} \tag{3.6}$$

is another factorization of $\mathbf{f}$, then

$$\begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{U}\mathbf{f}\mathbf{U}^* = (\mathbf{U}\mathbf{\gamma}\mathbf{U}^*)(\mathbf{U}\mathbf{\gamma}\mathbf{U}^*)^* \tag{3.7}$$

but $\mathbf{g} = \mathbf{\Psi}\mathbf{\Psi}^*$ implies that

$$\begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^*. \tag{3.8}$$

Since $\mathbf{\Psi}$ is the generating function of $\mathbf{Z}_n$ one can prove that the function

$$\begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

is the generating function of $\mathbf{Y}_n$. In fact we know that the generating function $\mathbf{\Phi}$ of a q-variate stationary stochastic process $\mathbf{X}_n$ is given by

$$\mathbf{\Phi} = \sum_{n=0}^{\infty} \mathbf{A}_n \sqrt{\mathbf{G}} e^{in\theta},$$

where $\mathbf{A}_n$'s are the coefficients in the representation

$$\mathbf{X}_0 = \sum_{n=0}^{\infty} \mathbf{A}_n \mathbf{g}_{-n} + (\mathbf{X}_0 | \mathbf{M}(-\infty))$$

of $\mathbf{X}_n$ in terms of its innovation process

$$\mathbf{g}_n = \mathbf{X}_n - (\mathbf{X}_n | \mathbf{M}(n-1))$$

and $\mathbf{G} = (\mathbf{g}_0, \mathbf{g}_0)$ is the predictor error matrix. Comparing $\mathbf{Z}_n$ with $\mathbf{Y}_n = [\mathbf{Z}_n | \mathbf{0}]^T$, we note that

$$\mathbf{g}_n^{\mathbf{Y}} = \begin{bmatrix} \mathbf{g}_n^{\mathbf{Z}} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{G}^{\mathbf{Y}} = \begin{bmatrix} \mathbf{G}^{\leqq} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \sqrt{\mathbf{G}^{\mathbf{Y}}} = \begin{bmatrix} \sqrt{\mathbf{G}^{\mathbf{Y}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{Y}_0 = \begin{bmatrix} \mathbf{Z}_0 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{\infty} \mathbf{A}_n^{\mathbf{Z}} \mathbf{g}_n^{\mathbf{Z}} + (\mathbf{Z}_0 | \mathbf{M}^{\mathbf{Z}}(-\infty)) \\ \mathbf{0} \end{bmatrix}$$

$$= \sum_{n=0}^{\infty} \begin{bmatrix} \mathbf{A}_n^{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{g}_n^{\mathbf{Z}} \\ \mathbf{0} \end{bmatrix} + (\mathbf{Y}_0 | \mathbf{M}^{\mathbf{Y}}(-\infty)).$$

Although the coefficients arising in this sum are not unique, they will give us the generating function uniquely, and we have

$$\mathbf{\Phi}^{\mathbf{Y}} = \sum_{n=0}^{\infty} \left(\begin{bmatrix} \mathbf{A}_n^{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \sqrt{\mathbf{G}^{\mathbf{Y}}} \right) e^{-in\theta}$$

$$= \sum_{n=0}^{\infty} \begin{bmatrix} \mathbf{A}_n^{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{\mathbf{G}^{\mathbf{Y}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} e^{-in\theta} = \sum_{n=0}^{\infty} \begin{bmatrix} \mathbf{A}_n^{\mathbf{Z}} \sqrt{\mathbf{G}^{\mathbf{Y}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} e^{-in\theta}$$

$$= \begin{bmatrix} \sum_{n=0}^{\infty} \mathbf{A}_n^{\mathbf{Z}} \sqrt{\mathbf{G}^{\mathbf{Z}}}\, e^{in\theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}^{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Thus

$$\begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

is the optimal factor of

$$\begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

(3.7) and (3.8) together with the optimality of

$$\begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

imply that

$$\begin{bmatrix} \mathbf{\Psi}_+(0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{\Psi}_+(0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \geqq (\mathbf{U}\boldsymbol{\gamma}_+(0)\mathbf{U}^*)(\mathbf{U}\boldsymbol{\gamma}_+(0)\mathbf{U}^*)^*.$$

This in turn implies that

$$(\boldsymbol{\delta}_+(0))^2 = \left(\mathbf{U}^* \begin{bmatrix} \mathbf{\Psi}_+(0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U} \right) \left(\mathbf{U}^* \begin{bmatrix} \mathbf{\Psi}_+(0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U} \right) \geqq \boldsymbol{\gamma}_+(0)\boldsymbol{\gamma}_+(0)^*.$$

This together with (3.5) shows that $\boldsymbol{\delta}$ is the optimal factor of $\mathbf{f}$. Thus by the uniqueness theorem mentioned above

$$\mathbf{\Phi} = \boldsymbol{\delta} = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}. \tag{3.9}$$

Q.E.D

Now we are ready to give the algorithm determining the generating function of our purely minimal q-variate stationary stochastic process $\mathbf{X}_n$:

Step 1. Since $\mathbf{f}$ satisfies condition (i) in (1.3) and its range is constant with dimension p, one can easily see that the rank of the covariance matrix $\mathbf{\Gamma}_0 = \int_0^{2\pi} \mathbf{f}(\theta)\, d\theta$ is also p. In fact, it can be shown that a unitary matrix $\mathbf{U}$ has the property (3.1) if and only if it can reduce $\mathbf{\Gamma}_0$ to the form

$$\mathbf{U\Gamma}_0\mathbf{U}^* = \begin{bmatrix} \mathbf{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{3.10}$$

with $\mathbf{\Sigma}_0$ being a $p \times p$ nonsingular matrix. Using this fact, we can find our unitary matrix $\mathbf{U}$ of this section by finding a $\mathbf{U}$ as in (3.10).

Step 2. Taking the unitary matrix $\mathbf{U}$ obtained in Step 1, we can form the full-rank density $\mathbf{g}$ as in (3.1). Because $\mathbf{f}$ satisfies conditions (ii) and (iii) of (1.3), as we saw before, $\mathbf{g}$ has the properties (i) and (ii) of (1.2). Thus we can use Masani's algorithm developed in § 4 of [2] to find the generating function $\mathbf{\Psi}$, of the process with density $\mathbf{g}$.

Step 3. We can find the generating function $\mathbf{\Phi}$ of our original process with density $\mathbf{f}$, via the formula (3.9) above.

Remark. One can similarly extend the other available extension of Masani's algorithm (such as that in [8]) for the full-rank case to obtain corresponding algorithms for the nonfull-rank case.

4. Determination of the predictor. In this section we show that the unique autoregressive series, of [2], giving the linear predictor in the full-rank case, can be used to obtain the predictor in our nonfull-rank case. In fact as we will see, exactly the same formula works in this case as well. We continue to assume that $\mathbf{F}$ is a.c. and the density $\mathbf{f}$ of our stationary stochastic process $\mathbf{X}_n$ satisfies conditions (1.3). Using the notations and results of § 3, we know that

$$\mathbf{f} = \mathbf{U} \begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^*$$

and the density $\mathbf{g}$ satisfies conditions (i) and (ii) of (1.2). Thus, using the technique developed in [2] one can show that

$$\hat{\mathbf{Z}}_\nu = \sum_{k=0}^{\infty} \mathbf{E}_{\nu k} \mathbf{Z}_{-k} \quad \text{in } H^p,$$

where

$$\mathbf{E}_{\nu k} = \sum_{n=0}^{k} \mathbf{C}_{\nu+n} \mathbf{D}_{k-n}$$

with $\mathbf{C}_k$ and $\mathbf{D}_k$ being the kth Fourier coefficients of $\mathbf{\Psi}$ and $\mathbf{\Psi}^{-1}$, respectively. Now one can easily verify that

$$\hat{\mathbf{Y}}_\nu = \begin{bmatrix} \hat{\mathbf{Z}}_\nu \\ \mathbf{0} \end{bmatrix} = \sum_{k=0}^{\infty} \begin{bmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Y}_{-k} \quad \text{in } H^q,$$

and

$$\begin{bmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \sum_{n=0}^{k} \begin{bmatrix} \mathbf{C}_{\nu+n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}_{k-n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.1}$$

Since $\mathbf{Y}_n = \mathbf{U}\mathbf{X}_n$, one can also verify that

$$\hat{\mathbf{X}}_n = \widehat{\mathbf{U}^*\mathbf{Y}}_n = \mathbf{U}^*\hat{\mathbf{Y}}_n.$$

Hence we have

$$\begin{aligned} \hat{\mathbf{X}}_n &= \mathbf{U}^*\left(\sum_{k=0}^{\infty} \begin{vmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{vmatrix} \mathbf{Y}_{-k}\right) \\ &= \sum_{k=0}^{\infty} \left(\mathbf{U}^* \begin{bmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}\right) \mathbf{U}^*\mathbf{Y}_{-k} \quad \text{in } H^q. \end{aligned} \tag{4.2}$$

Letting

$$\mathbf{F}_{\nu k} = \mathbf{U}^* \begin{bmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}, \tag{4.3}$$

we get the following autoregressive series representation for the best linear predictor $\hat{\mathbf{X}}_\nu$:

$$\hat{\mathbf{X}}_\nu = \sum_{k=0}^{\infty} \mathbf{F}_{\nu k}\mathbf{X}_{-k}.$$

Now let us examine the coefficients $\mathbf{F}_{\nu k}$ in (4.3) more carefully. Doing this, we will be able to write $\mathbf{F}_{\nu k}$ in terms of the Fourier coefficients of the generating function $\mathbf{\Phi}$ of our original process $\mathbf{X}_n$ rather than that of the auxiliary process $\mathbf{Z}_n$. From (4.2) we can write

$$\mathbf{F}_{\nu k} = \mathbf{U}^* \begin{bmatrix} \mathbf{E}_{\nu k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}.$$

Now using (4.1), we have

$$\begin{aligned} \mathbf{F}_{\nu k} &= \mathbf{U}^*\left(\sum_{n=0}^{k} \begin{bmatrix} \mathbf{C}_{\nu+n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}_{k-n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)\mathbf{U} \\ &= \sum_{n=0}^{k} \left(\mathbf{U}^* \begin{bmatrix} \mathbf{C}_{\nu+n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}\right)\left(\mathbf{U}^* \begin{bmatrix} \mathbf{D}_{k-n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}\right). \end{aligned}$$

Thus

$$\mathbf{F}_{\nu k} = \sum_{n=0}^{k} \mathbf{M}_{\nu+n}\mathbf{N}_{k-n},$$

with

$$\mathbf{M}_n = \mathbf{U}^* \begin{bmatrix} \mathbf{C}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}, \qquad \mathbf{N}_n = \mathbf{U}^* \begin{bmatrix} \mathbf{D}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}.$$

But by the lemma we have

$$\mathbf{\Phi} = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}, \qquad \mathbf{\Phi}^{\#} = \mathbf{U}^* \begin{bmatrix} \mathbf{\Psi}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}. \tag{4.4}$$

Thus we observe that $\mathbf{M}_n$ and $\mathbf{N}_n$ are exactly the nth Fourier coefficients of $\mathbf{\Phi}$ and $\mathbf{\Phi}^{\#}$, respectively.

Summarizing, we have shown that the best linear predictor $\hat{\mathbf{X}}_\nu$ can be written exactly in the same form obtained in [2] for the full-rank processes, i.e., we have

$$\hat{\mathbf{X}}_\nu = \sum_{k=0}^{\infty} \left(\sum_{n=0}^{k} \mathbf{M}_{\nu+n}\mathbf{D}_{k-n}\right)\mathbf{X}_{-k} \quad \text{in } H^q$$

where $\mathbf{M}_n$ and $\mathbf{N}_n$ are the nth Fourier coefficients of $\mathbf{\Phi}$ and its generalized inverse $\mathbf{\Phi}^{\#}$ (instead of $\mathbf{\Phi}$ and its inverse $\mathbf{\Phi}^{-1}$ in the full-rank case).

REFERENCES

[1] A. MAKAGON AND A. WERON, *Wold-Cramer concordance theorem for interpolation of q-variate stationary processes over locally compact abelian groups*, J. Multivariate Anal., 6 (1976), pp. 123-137.

[2] P. MASANI, *The prediction theory of multivariate stochastic processes*, III, Acta Math., 104 (1960), pp. 141-162.

[3] ———, *Shift invariant spaces and prediction theory*, Acta Math., 107 (1962), pp. 275-290.

[4] ———, *Recent Trends in Multivariate Prediction Theory*, P. R. Krishnaiah, ed., Academic Press, New York, 1966, pp. 351-382.

[5] A. G. MIAMEE AND H. SALEHI, *On the bilateral prediction error matrix of a multivariate stationary stochastic process*, this Journal, 10 (1979), pp. 247-253.

[6] M. POURAHMADI, *A matricial extension of the Helson-Szegö theorem and its application in multivariate prediction*, J. Multivariate Anal., 16 (1985), pp. 265-275.

[7] J. B. ROBERTSON AND M. ROSENBERG, *The decomposition of matrix-valued measures*, Michigan Math. J., 15 (1968), pp. 353-368.

[8] H. SALEHI, *On determination of the optimal factor of a nonnegative matrix-valued function*, Proc. Amer. Math. Soc., 2 (1971), pp. 383-389.

[9] N. WIENER AND P. MASANI, *The prediction theory of multivariate stochastic processes*, I, Acta Math., 98 (1957), pp. 111-150.

[10] ———, *The prediction theory of multivariate stochastic processes*, II, Acta Math., 99 (1958), pp. 93-137.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

003

A STOCHASTIC INTEGRATION FORMULA FOR TWO-PARAMETER WIENER×TWO-PARAMETER WIENER SPACE*

G. W. JOHNSON† AND D. L. SKOUG†

Abstract. We establish a stochastic integration by parts formula in which both the integrator and the integrand are elements of two-parameter Wiener (or Yeh-Wiener) space. We also establish the continuity of the stochastic integral with respect to binary quadratic approximation.

Key words. stochastic integration by parts, Wiener process, two-parameter (or Yeh-Wiener) process

AMS(MOS) subject classifications. Primary 60H05; secondary 28C20

1. Introduction. We prove an integration by parts formula where the integrator and the integrand are independent random functions from two-parameter Wiener (or Yeh-Wiener) space. In light of the recent great interest in Markov random fields among probabilists [15], [26] and mathematical physicists [8], [18], [19], [27], [29], it seems likely that this result and extensions of it will be useful. There are parts formulas in the literature involving two independent one-parameter processes [7, p. 268], but, as far as we can tell, the present result is new.

Our interest in this topic came out of our joint work on the Feynman integral [4] with Kun Chang. In proving that a certain class of functions on one-parameter Wiener space is in a Banach algebra of Feynman integrable functions, we found it necessary to move a Wiener variable from the integrand to the integrator. Ultimately we justified this by proving a parts formula involving two independent Wiener processes, one a one-parameter and the other a two-parameter process [4, Thm. 3.1]. An additional nonstochastic function was also involved. The present result was essentially discovered while we were finishing the earlier paper.

Our proofs here and in [4] are, in some respects, analysis proofs. It is quite likely that alternate proofs making more use of the machinery of stochastic processes can be found which will shed further light on these formulas. The second author is pursuing extensions of these parts formulas in various directions.

Let $C_1[0,1]$ denote Wiener space; that is, the space of continuous functions x on $[0,1]$ such that $x(0)=0$. Let $Q=[0,1]^2$ and let $C_2=C_2(Q)$ denote two-parameter Wiener (or Yeh-Wiener) space; that is, the space of continuous functions f on Q such that $f(0,t)=f(s,0)=0$ for all $(s,t)\in Q$. Let m_1 denote Wiener measure and m_2 denote Yeh-Wiener measure.

We now briefly describe our results, delaying further precise definitions until later. In our main result, Theorem 4.1, we obtain an integration by parts formula relating the stochastic integrals $\int_Q f\,\tilde{d}_2 g$ and $\int_Q g\,\tilde{d}_2 f$, where both f and g are elements of $C_2(Q)$. One version of the corresponding well-known result for stochastic integrals of functions of one variable is given in [7, p. 268].

The key to our parts formula is Theorem 3.1 in which we obtain the continuity with respect to "binary quadratic approximation" (to be defined below) of the stochastic

* Received by the editors November 18, 1985; accepted for publication (in revised form) May 20, 1986. This research was partially supported by National Science Foundation grant DMS-8403197.

† Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588-0323.

integral

$$F(f) := \int_Q g\, \tilde{d}_2 f$$

for each g in $L_2(Q)$. This is an extension of a recent result of Cameron and Storvick [1, Lemma 2.3] in which they obtain the continuity with respect to "binary polygonal approximation" of the stochastic integral $\int_0^1 V(s)\, \tilde{d}x(s)$ for each V in $L_2[0, 1]$.

2. Preliminaries. The concept of bounded variation for a function of two variables is surprisingly complex. Several nonequivalent definitions have appeared in the literature; see for example [5], [6], [9], [10], [17]. (We would like to thank our colleague Professor Gary Meisters for bringing some of these references to our attention.) The paper [5] by Clarkson and Adams is old but still useful in sorting out many of the relationships between the various definitions. Throughout this paper we will use the definition used by Hardy and by Krause (see [5, p. 825] and [9, p. 345]) which we now briefly review.

Let Δ denote the rectangular partition of Q determined by $0 = s_0 < s_1 < \cdots < s_n = 1$ and $0 = t_0 < t_1 < \cdots < t_m = 1$. A function $f(s, t)$ is said to be of bounded variation on Q in the sense of Hardy and Krause provided the following three conditions hold:

(i) There exists a constant K such that for any partition Δ

$$\sum_{i=1}^{n} \sum_{j=1}^{m} |f(s_i, t_j) - f(s_{i-1}, t_j) - f(s_i, t_{j-1}) + f(s_{i-1}, t_{j-1})| \leqq K, \tag{2.1}$$

(ii) $f(s, t)$ is a function of bounded variation in s for each $t \in [0, 1]$,

(iii) $f(s, t)$ is a function of bounded variation in t for each $s \in [0, 1]$.

The total variation of f over Q, Var (f, Q), is defined to be the supremum of the sums in (2.1) over all partitions Δ. It is easy to see [9, p. 345] that conditions (ii) and (iii) can be relaxed to the requirements that $f(s, t)$ is of bounded variation in s for one fixed value of t and is of bounded variation in t for one fixed value of s. It is also easy to see that if f is of bounded variation on Q then the set of discontinuities of f lie on a countable number of vertical and horizontal lines.

For later reference we mention that if a function f satisfies condition (i) above, but not necessarily conditions (ii) and (iii), then f is said to be of bounded variation on Q in the sense of Vitali [5, pp. 824-825].

Next we give a brief discussion of the Riemann-Stieltjes integral $\int_Q g\, df$. Let $f(s, t)$ and $g(s, t)$ be defined and bounded on Q. Let Δ be the rectangular partition of Q given above and let $|\Delta|$, the norm of Δ, be given by

$$|\Delta| = \sup_{i,j} \{[(s_i - s_{i-1})^2 + (t_j - t_{j-1})^2]^{1/2}\}.$$

Then $g(s, t)$ is said to be Riemann-Stieltjes integrable with respect to $f(s, t)$ on Q with Riemann-Stieltjes integral I, if and only if corresponding to any $\varepsilon > 0$ there exists a $\delta > 0$ such that, for any rectangular partition Δ of Q with $|\Delta| < \delta$ and any choice of points (ξ_i, η_j) with $s_{i-1} \leqq \xi_i \leqq s_i$ and $t_{j-1} \leqq \eta_j \leqq t_j$, we have

$$\left| \sum_{i=1}^{n} \sum_{j=1}^{m} g(\xi_i, \eta_j)[f(s_i, t_j) - f(s_{i-1}, t_j) - f(s_i, t_{j-1}) + f(s_{i-1}, t_{j-1})] - I \right| < \varepsilon.$$

The following well-known theorem [9, p. 561] gives a necessary and sufficient condition for the existence of the Riemann-Stieltjes integral $\int_Q g\, df$.

THEOREM. *Let f be of bounded variation (Hardy and Krause) on Q and let g be a bounded function defined on Q. Then a necessary and sufficient condition that g be Riemann-Stieltjes integrable with respect to f on Q is that the set of discontinuities of g have total variation measure zero with respect to f.*

COROLLARY. *If g is continuous on Q and f is of bounded variation (Hardy and Krause) on Q then $\int_Q g\,df$ exists.*

The definition of bounded variation used by Hardy and Krause has the important property that if g is continuous on Q and f is of bounded variation on Q then the Riemann-Stieltjes integrals $\int_Q g\,df$ and $\int_Q f\,dg$ both exist and are related by the usual integration by parts formula. The definition of bounded variation given by Vitali fails to have this property since $f(s, t)$ may be of bounded variation on Q while $f(s, 1)$ is not of bounded variation on $[0, 1]$; for example $f(s, t) = s \sin(1/s)$.

A paper of Yeh [31] has a nice discussion of the n-dimensional Riemann-Stieltjes integral and some of its properties. However his definition is based on the Vitali concept of bounded variation. Of course all of the results he obtains concerning the Riemann-Stieltjes integral are true in our more restrictive setting. In theorems involving Yeh-Wiener space $C_2(Q)$, Yeh then adds conditions (ii) and (iii) as hypotheses; so in the final analysis the setting is essentially the same.

Next we give the definition of the Paley-Wiener-Zygmund (P.W.Z.) integral, a simple type of stochastic integral, for functions of one and two variables.

Let $\{\phi_j\}$ be a complete orthonormal set of functions of bounded variation on $[0, 1]$. For V in $L_2[0, 1]$ let

$$V_n(s) = \sum_{j=1}^{n} (V, \phi_j)\phi_j(s).$$

The P.W.Z. integral for functions of one variable is defined by the formula

$$\int_0^1 V(s)\,\tilde{d}x(s) := \lim_{n\to\infty} \int_0^1 V_n(s)\,dx(s)$$

for all x in $C_1[0, 1]$ for which the limit exists.

Let $\{\phi_j\}$ be a complete orthonormal set of functions of bounded variation on Q. For g in $L_2(Q)$ let

$$g_n(s, t) = \sum_{j=1}^{n} (g, \phi_j)\phi_j(s, t).$$

Then the P.W.Z. integral for functions of two variables is defined by the formula

$$\int_Q g(s, t)\tilde{d}_2 f(s, t) := \lim_{n\to\infty} \int_Q g_n(s, t)\,df(s, t)$$

for all f in $C_2(Q)$ for which the limit exists.

In order to state various properties of the P.W.Z. integral we need the notion of scale-invariant measurability. A subset A of $C_2(Q)$ is said to be scale-invariant measurable provided ρA is Yeh-Wiener measurable for every $\rho > 0$, and a scale-invariant measurable set N is said to be scale-invariant null provided $m_2(\rho N) = 0$ for every $\rho > 0$ [3], [11]. A property that holds except on a scale-invariant null set is said to hold scale-invariant almost everywhere (s-a.e.).

Following are some useful facts about the P.W.Z. integral (we will state them in terms of functions of two variables; similar properties of course hold in the one variable setting):

(2.2) For each g in $L_2(Q)$ the P.W.Z. integral $\int_Q g\,\tilde{d}_2 f$ exists for s-a.e. f in $C_2(Q)$.

(2.3) The P.W.Z. integral $\int_Q g\,\tilde{d}_2 f$ is essentially independent of the complete orthonormal set $\{\phi_j\}$. (For us it will often be convenient to let $\{\phi_j\}$ be the Haar functions on Q.)

(2.4) If g is of bounded variation on Q, then the P.W.Z. integral $\int_Q g\,\tilde{d}_2 f$ is s-a.e. equal to the Riemann-Stieltjes integral $\int_Q g\,df$.

(2.5) The P.W.Z. integral has the usual linearity properties when interpreted properly.

(2.6) The map sending g in $L_2(Q)$ to the function F_g on $C_2(Q)$ given by $F_g(f) := \int_Q g\,\tilde{d}_2 f$ is an isometric isomorphism of $L_2(Q)$ into $L_2(C_2(Q), m_2)$.

(2.7) The sequence $\{\int_Q g_n df\}$, considered as a function of f converges in $L_2(C_2(Q))$ mean to $\int_Q g\,\tilde{d}_2 f$.

Remark. W. J. Park [24] defines the stochastic integral with respect to the two-parameter (actually p-parameter) Wiener process in two ways: (i) as above except using the two variable Haar functions for the complete orthonormal set $\{\phi_j\}$; (ii) following the usual approach of the Itô theory. Park shows that these two approaches are equivalent in the sense that for every g in $L_2(Q)$ they are equal for m_2-a.e. f in $C_2(Q)$. It is easy to see that they are in fact equal for s-a.e. f in $C_2(Q)$. When these facts are combined with (2.3) above, we see that in our setting the P.W.Z. integral essentially agrees with the usual Itô stochastic integral. However, these integrals do not necessarily agree when the integrand contains a random function. For example, in the one variable setting, the Itô integral $\int_0^1 x(t)\,dx(t)$ equals $(x^2(1)-1)/2$ for a.e. $x \in C_1[0,1]$, while the P.W.Z. integral $\int_0^1 x(t)\,\tilde{d}x(t)$ equals $x^2(1)/2$ for a.e. $x \in C_1[0,1]$.

3. Continuity of the P.W.Z. integral with respect to binary quadratic approximation. As mentioned in the introduction, one of the keys to our proof of the parts formula is an extension to the two-parameter P.W.Z. integral of a recent result of Cameron and Storvick [1, Lemma 2.3] concerning the one-parameter P.W.Z. integral. Their result says that for every V in $L_2[0,1]$, $\int_0^1 V(s)\,\tilde{d}x(s)$ is continuous with respect to "binary polygonal approximation" for s-a.e. x in $C_1[0,1]$. The binary polygonal approximators for x are piecewise linear functions which agree with x on a certain set of binary rationals which form a partition of $[0,1]$. In trying to extend this result, it seemed natural to use piecewise linear functions of two variables as approximators for f in $C_2(Q)$. There are technical problems with this and we could not see how to make it work. We have instead used functions which are binary polygonal approximations in each variable when the other variable is fixed but which are quadratic functions of two variables. We will refer to these approximators $[f]_m$ for f as "binary quadratic approximations." We begin this section by defining $[f]_m$ precisely and noting several of its properties.

Let m be a nonnegative integer and consider the division of Q into 2^{2m} squares by means of the partition $0 = s_0 < s_1 = 1/2^m < \cdots < s_i = i/2^m < \cdots < s_{2^m} = 1$ and $0 = t_0 < t_1 = 1/2^m < \cdots < t_j = j/2^m < \cdots < t_{2^m} = 1$. For each f in $C_2(Q)$ we define the mth binary quadratic approximation $[f]_m$ by the formula

$$[f]_m(s,t) := \frac{f(s_i,t_j) - f(s_{i-1},t_j) - f(s_i,t_{j-1}) + f(s_{i-1},t_{j-1})}{(s_i - s_{i-1})(t_j - t_{j-1})}(s - s_{i-1})(t - t_{j-1})$$

$$+\frac{f(s_i,t_{j-1}) - f(s_{i-1},t_{j-1})}{(s_i - s_{i-1})}(s - s_{i-1}) \tag{3.1}$$

$$+\frac{f(s_{i-1}, t_j)-f(s_{i-1}, t_{j-1})}{(t_j - t_{j-1})}(t-t_{j-1})+f(s_{i-1}, t_{j-1})$$

for $(s, t) \in [s_{i-1}, s_i] \times [t_{j-1}, t_j]$, $i, j = 1, 2, \cdots, 2^m$.

We first note the following properties of $[f]_m$:

(3.2) $[f]_0(s, t) = stf(1, 1)$,

(3.3) $[f]_m(s_i, t_j) = f(s_i, t_j)$ at all binary partition points (s_i, t_j),

(3.4) $[f]_m \in C_2(Q)$ for each m and all $f \in C_2(Q)$,

(3.5) $\|[f]_m\|_\infty \leqq \|f\|_\infty$ for each m and all $f \in C_2(Q)$,

(3.6) For each f in $C_2(Q)$, $\|f-[f]_m\|_\infty \to 0$ as $m \to +\infty$.

(3.7) A direct caclulation shows that

$$\operatorname{Var}([f]_m, [s_{i-1}, s_i] \times [t_{j-1}, t_j]) = \int_{t_{j-1}}^{t_j} \int_{s_{i-1}}^{s_i} \left| \frac{\partial^2 [f]_m(s, t)}{\partial s \partial t} \right| ds\, dt$$
$$= |f(s_i, t_j) - f(s_{i-1}, t_j) - f(s_i, t_{j-1}) + f(s_{i-1}, t_{j-1})|$$

for each i, $j = 1, 2, \cdots, 2^m$ (also see [31, p. 413]). Hence for each f in $C_2(Q)$ and $m = 0, 1, 2, \cdots, [f]_m$ is of bounded variation on Q and we have the formula

$$\operatorname{Var}([f]_m, Q) = \int_0^1 \int_0^1 \left| \frac{\partial^2 [f]_m(s, t)}{\partial s\, \partial t} \right| ds\, dt$$
$$= \sum_{i=1}^{2^m} \sum_{j=1}^{2^m} |f(s_i, t_j) - f(s_{i-1}, t_j) - f(s_i, t_{j-1}) + f(s_{i-1}, t_{j-1})|.$$

(3.8) For each $s \in [0, 1]$, $[f]_m$ is a binary polygonal function of t in $C_1[0, 1]$ while, for each $t \in [0, 1]$, $[f]_m$ is a binary polygonal function of s in $C_1[0, 1]$, i.e., it is linear on each interval $[s_{i-1}, s_i]$.

Remark. After reading a preliminary handwritten version of this paper, Chull Park has kindly pointed out to us that he made use of these "binary quadratic approximators" in his paper [22].

Now using our binary quadratic approximators, we make a definition which parallels Cameron and Storvick's definition [1, p. 6] of continuity with respect to binary polygonal approximation.

DEFINITION. A function $F: C_2(Q) \to C$ will be called continuous at $f \in C_2(Q)$ with respect to binary quadratic approximation if $\lim_{m\to\infty} F([f]_m) = F(f)$.

We note that if F is continuous on $C_2(Q)$ then it is certainly continuous with respect to binary quadratic approximation at every f in $C_2(Q)$.

The Haar functions

$$H_0^{(0)}(s) \equiv 1 \quad \text{on } [0, 1]$$

and

$$H_n^{(k)}(s) = \begin{cases} 2^{n/2}, & \dfrac{2k-2}{2^{n+1}} \leqq s < \dfrac{2k-1}{2^{n+1}}, \\ -2^{n/2}, & \dfrac{2k-1}{2^{n+1}} < s \leqq \dfrac{2k}{2^{n+1}}, \\ 0 & \text{elsewhere}, \end{cases}$$

for $n=0,1,2,\cdots$ and $k=1,2,\cdots,2^n$ are a complete orthonormal set on $[0,1]$. Denote this collection by $\mathscr{H}$ and let

$$\mathscr{G}:=\{H_n^{(k)}(s)H_l^{(q)}(t)\colon H_n^{(k)} \text{ and } H_l^{(q)} \text{ are in } \mathscr{H}\}.$$

Then $\mathscr{G}$, the set of Haar functions on Q, is a complete orthonormal set. Furthermore the elements of $\mathscr{G}$ are defined everywhere on Q, are step-functions, and are of bounded variation on Q.

Remark. Let m be a positive integer and let $h(s,t)$ be a step-function on Q with partition points at the binary points (s_i,t_j), $i,j=0,1,\cdots,2^m$. Then $h(s,t)$ is *orthogonal* to the Haar function $H_n^{(k)}(s)H_l^{(q)}(t)\in\mathscr{G}$ whenever $\max\{n,l\}\geqq m$. Hence the orthogonal development of $h(s,t)$ in terms of the Haar functions, namely

$$h(s,t)=\sum_{g\in\mathscr{G}}(h,g)g(s,t)$$

will only involve terms $g(s,t)=H_n^{(k)}(s)H_l^{(q)}(t)$ with $n\leqq m-1$ and $l\leqq m-1$.

Next we give two lemmas which will be used in proving our continuity result.

LEMMA 3.1. *Let $g(s,t)=H_n^{(k)}(s)H_l^{(q)}(t)$ be any element of $\mathscr{G}$ (here of course n, l, k and q are nonnegative integers such that $0\leqq k\leqq 2^n$ and $0\leqq q\leqq 2^l$). Then for each $m>\max\{n,l\}$ and each $f\in C_2(Q)$ we have that*

$$\int_Q g(s,t)\,df(s,t)=\int_Q g(s,t)d[f]_m(s,t). \tag{3.9}$$

Proof. Since $m\geqq n+1$ and $m\geqq l+1$ we can interpret $g(s,t)$ as a step-function on Q with partition points at the binary points (s_i,t_j), $i,j=0,1,2,\cdots,2^m$. Since g is of bounded variation on Q we can apply the ordinary integration by parts formula for two-dimensional Riemann-Stieltjes integrals [9, p. 666] or [31, p. 415] and obtain the formula

$$\begin{aligned}\int_Q g(s,t)d(f(s,t)-[f]_m(s,t))&=g(1,1)(f(1,1)-[f]_m(1,1))\\&\quad-\int_0^1\{f(s,1)-[f]_m(s,1)\}\,dg(s,1)\\&\quad-\int_0^1\{f(1,t)-[f]_m(1,t)\}\,dg(1,t)\\&\quad+\int_Q\{f(s,t)-[f]_m(s,t)\}\,dg(s,t).\end{aligned}\tag{3.10}$$

First we note that each of the first three terms on the right-hand side of (3.10) equals zero since f and $[f]_m$ agree at the partition points (s_i,t_j) and the jumps of the functions $g(s,1)$ and $g(1,t)$ occur only at the partition points $(s_i,1)$ and $(1,t_j)$ respectively.

Next we see that $\int_Q\{f(s,t)-[f]_m(s,t)\}\,dg(s,t)=0$. This follows from the definition of the integral since, for any rectangle $R=[a,b]\times[\alpha,\beta]\subseteq Q$, if R contains no binary partition points (s_i,t_j), then g is constant in either its first or second variable and so

$$g(b,\beta)-g(b,\alpha)-g(a,\beta)+g(a,\alpha)=0.$$

Thus Lemma 3.1 is established.

LEMMA 3.2. *Let $V(s, t) \in L_2(Q)$. Then for $m = 0, 1, 2, \cdots$, the left member below exists for every $f \in C_2(Q)$ and we have*

$$\int_Q V(s, t)\tilde{d}_2[f]_m(s, t) = \int_0^1 \int_0^1 V(s, t)\frac{\partial^2[f]_m(s, t)}{\partial s \partial t}\, ds\, dt. \tag{3.11}$$

In the special case $V(s, t) = H_n^{(k)}(s)H_l^{(q)}(t) \in \mathscr{G}$, both sides of equation (3.11) vanish whenever $m \leqq \max\{n, l\}$.

Proof. Equation (3.11) follows from [21, Thm. 4] since $[f]_m(s, t)$ is absolutely continuous on Q. Next assume that $m \leqq \max\{n, l\}$ and consider the partition of Q with binary partition points (s_i, t_j), $i, j = 0, 1, \cdots, 2^m$. Without loss of generality assume that $n \geqq m$. Then for all k, $\int_{s_{i-1}}^{s_i} H_n^{(k)}(s)\, ds = 0$, and so, on each square $(s_{i-1}, s_i) \times (t_{j-1}, t_j)$, we see that

$$\begin{aligned}\int_{s_{i-1}}^{s_i} \int_{t_{j-1}}^{t_j} & H_l^{(q)}(t)H_n^{(k)}(s)\frac{\partial^2[f]_m(s, t)}{\partial s \partial t}\, dt\, ds \\ &= \frac{f(s_i, t_j) - f(s_{i-1}, t_j) - f(s_i, t_{j-1}) + f(s_{i-1}, t_{j-1})}{(s_i - s_{i-1})(t_j - t_{j-1})} \int_{t_{j-1}}^{t_j} H_l^{(q)}(t)\, dt \int_{s_{i-1}}^{s_i} H_n^{(k)}(s)\, ds \\ &= 0\end{aligned}$$

from which it easily follows that the right-hand side of (3.11) equals zero. Thus Lemma 3.2 is established.

We are now ready to establish the main result in this section.

THEOREM 3.1. *Let $V \in L_2(Q)$ and let*

$$F(f) := \int_Q V(s, t)\tilde{d}_2 f(s, t). \tag{3.12}$$

Then for s-a.e. *$f \in C_2(Q)$, $F(f)$ is continuous with respect to binary quadratic approximation.*

Proof. Assume that the P.W.Z. integral in (3.12) is given in terms of the Haar functions $\mathscr{G}$. Let

$$V_0(s, t) := (H_0^{(0)}H_0^{(0)}, V)H_0^{(0)}(s)H_0^{(0)}(t) = \int_0^1 \int_0^1 V(u, v)\, du\, dv$$

and for $m = 1, 2, 3, \cdots$. Let

$$V_m(s, t) := \sum_{\substack{g \in \mathscr{G} \\ \max\{n, l\} < m}} (g, V)g(s, t), \tag{3.13}$$

where as usual $g(s, t) = H_n^{(k)}(s)H_l^{(q)}(t)$. Next for each $f \in C_2(Q)$ let

$$f_0(s, t) := \int_Q H_0^{(0)}(s)H_0^{(0)}(t)\, df(s, t) = f(1, 1)$$

and for $m = 1, 2, \cdots$, let

$$f_m(s, t) := \sum_{\substack{g \in \mathscr{G} \\ \max\{n,l\} < m}} g(s, t) \int_Q g(u, v)\, df(u, v). \tag{3.14}$$

Note that the sequences $\{V_m\}_{m=0}^{\infty}$ and $\{f_m\}_{m=0}^{\infty}$ are "expansions" in terms of the Haar functions $\mathscr{G}$. It is quite easy to see that the functions V_m and f_m are related by the equation

$$\int_Q V_m(s,t)\,df(s,t)=\int_0^1\int_0^1 V(s,t)f_m(s,t)\,ds\,dt \tag{3.15}$$

for all m.

Next we claim that for each $m=0,1,2,\cdots$

$$\frac{\partial^2[f]_m(s,t)}{\partial s\partial t}=f_m(s,t) \tag{3.16}$$

for almost all $(s,t)\in Q$. This is clear in the case $m=0$ since in that case each side of (3.16) equals $f(1,1)$. We will establish (3.16) by showing that both sides of (3.16) have the same orthogonal expansion in terms of the Haar functions $\mathscr{G}$. First, using (3.14), Lemma 3.1, and Lemma 3.2, it follows that

$$\begin{aligned} f_m(s,t)&=\sum_{\substack{g\in\mathscr{G}\\ \max\{n,l\}<m}} g(s,t)\int_Q g(u,v)\,df(u,v)\\ &=\sum_{\substack{g\in\mathscr{G}\\ \max\{n,l\}<m}} g(s,t)\int_Q g(u,v)d[f]_m(u,v)\\ &=\sum_{\substack{g\in\mathscr{G}\\ \max\{n,l\}<m}} g(s,t)\int_0^1\int_0^1 g(u,v)\frac{\partial^2[f]_m(u,v)}{\partial u\partial v}\,du\,dv\\ &=\int_0^1\int_0^1\frac{\partial^2[f]_m(u,v)}{\partial u\partial v}\Bigg(\sum_{\substack{g\in\mathscr{G}\\ \max\{n,l\}<m}} g(u,v)g(s,t)\Bigg)\,du\,dv. \end{aligned} \tag{3.17}$$

Using (3.17) and the fact that $\mathscr{G}$ is a complete orthonormal set of functions on Q we obtain that

$$\int_0^1\int_0^1 f_m(s,t)h(s,t)\,ds\,dt=\int_0^1\int_0^1\frac{\partial^2[f]_m(u,v)}{\partial u\partial v}h(u,v)\,du\,dv \tag{3.18}$$

for all elements $h(s,t)=H_n^{(k)}(s)H_l^{(q)}(t)$ in $\mathscr{G}$ with $\max\{n,l\}<m$. (Simply multiply both sides of (3.17) by $h(s,t)$ and then integrate with respect to s and t over Q.) Next, using (3.17) we see that

$$\int_0^1\int_0^1 f_m(s,t)h(s,t)\,ds\,dt=0$$

for all elements $h(s,t)=H_n^{(k)}(s)H_l^{(q)}(t)$ in $\mathscr{G}$ with $m\leqq\max\{n,l\}$. Also for all such h, using Lemma 3.2 we see that

$$\int_0^1\int_0^1 h(s,t)\frac{\partial^2[f]_m(s,t)}{\partial s\partial t}\,ds\,dt=0.$$

Thus (3.18) holds for all $h\in\mathscr{G}$ and so both sides of (3.16) have the same orthogonal expansion in terms of the Haar functions on Q and so they are equal almost everywhere on Q.

Finally, using the definition of the P.W.Z. integral, (3.15), (3.16), and Lemma 3.2, we see that for s-a.e. $f\in C_2(Q)$

$$\begin{aligned}F(f)&:=\int_Q V(s,t)\tilde{d}_2f(s,t)\\&=\lim_{m\to\infty}\int_Q V_m(s,t)\,df(s,t)\\&=\lim_{m\to\infty}\int_0^1\int_0^1 V(s,t)f_m(s,t)\,ds\,dt\\&=\lim_{m\to\infty}\int_0^1\int_0^1 V(s,t)\frac{\partial^2[f]_m(s,t)}{\partial s\partial t}\,ds\,dt\\&=\lim_{m\to\infty}\int_Q V(s,t)\,\tilde{d}_2[f]_m(s,t)\\&=\lim_{m\to\infty}F([f]_m),\end{aligned}$$

which establishes Theorem 3.1.

COROLLARY 3.1. *Let $R=[a,b]\times[\alpha,\beta]\subseteq Q$. Let $U(s,t)\in L_2(R)$ and let*

$$H(f):=\int_R U(s,t)\,\tilde{d}_2f(s,t).$$

Then for s-a.e. $f\in C_2(Q)$, $H(f)$ is continuous with respect to binary quadratic approximation.

Proof. Let $V(s,t)=U(s,t)\chi_R(s,t)$. Then $V\in L_2(Q)$ and

$$\begin{aligned}H(f)&:=\int_R U(s,t)\,\tilde{d}_2f(s,t)\\&=\int_Q V(s,t)\,\tilde{d}_2f(s,t)\\&=\lim_{m\to\infty}\int_Q V(s,t)\,\tilde{d}_2[f]_m(s,t)\\&=\lim_{m\to\infty}\int_R U(s,t)\,\tilde{d}_2[f]_m(s,t)\\&=\lim_{m\to\infty}H([f]_m).\end{aligned}$$

4. A stochastic integration by parts formula for Yeh–Wiener × Yeh–Wiener space. To obtain our main result we will use the fact that the function $\|f\|_\infty^2$ is in $L_1(C_1,m_2)$. For the sake of completeness we include an elementary proof of a result somewhat stronger than this (see [16, pp. 159-164] for a related discussion).

LEMMA 4.1. *For $\alpha<\frac{1}{2}$, $\exp\{\alpha\|f\|_\infty^2\}\in L_1(C_2,m_2)$.*

Proof. Let $G(f)=\exp\{\alpha\|f\|_\infty^2\}$, where $\|f\|_\infty=\sup_Q|f(s,t)|$. By [25, p. 30] we know that G is in $L_1(C_2,m_2)$ if and only if

$$\int_0^\infty m_2\{f\in C_2(Q):|Gf)|>\lambda\}\,d\lambda<\infty.$$

Since m_2 is a probability measure it suffices to show that

$$\int_2^\infty m_2\{f \in C_2(Q): |G(f)| > \lambda\}\, d\lambda < \infty.$$

Let $N(\cdot)$ denote the standard normal distribution function. Kiefer has shown [23, p. 455] that $m_2\{f: \sup_Q f(s,t) > \lambda\} \leqq 4N(-\lambda)$. Thus for $\lambda \geqq 2$, letting $\rho \equiv (\ln \lambda/\alpha)^{1/2}$ we see that

$$\begin{aligned}
m_2\{f: |G(f)| > \lambda\} &= m_2\{f: \exp\{\alpha\|f\|_\infty^2\} > \lambda\} \\
&= m_2\{f: \|f\|_\infty > \rho\} \\
&= m_2[\{f: \sup_Q f(s,t) > \rho\} \cup \{f: \inf_Q f(s,t) < -\rho\}] \\
&\leqq m_2\{f: \sup_Q f(s,t) > \rho\} + m_2\{f: \inf_Q f(s,t) < -\rho\} \\
&= 2m_2\{f: \sup_Q f(s,t) > \rho\} \\
&\leqq 8N(-\rho) \\
&= \frac{8}{\sqrt{2\pi}} \int_{-\infty}^{-\rho} \exp\left[\frac{-u^2}{2}\right] du \\
&= \frac{8}{\sqrt{2\pi}} \int_{\rho}^{\infty} \exp\left[\frac{-u^2}{2}\right] du \\
&\leqq \frac{8}{\rho\sqrt{2\pi}} \int_{\rho}^{\infty} u \exp\left[\frac{-u^2}{2}\right] du \\
&= \frac{8}{\rho\sqrt{2\pi}} \exp\left(\frac{-\rho^2}{2}\right) \\
&= 8\left(\frac{\alpha}{2\pi \ln \lambda}\right)^{1/2} \exp\left(-\frac{\ln \lambda}{2\alpha}\right) \\
&= 8\left(\frac{\alpha}{2\pi \ln \lambda}\right)^{1/2} \left(\frac{1}{\lambda}\right)^{1/2\alpha}
\end{aligned}$$

which is in $L_1([2,\infty))$ as a function of λ if and only if $\alpha < \frac{1}{2}$.

COROLLARY 4.1. $\|f\|_\infty^2 \in L_1(C_2, m_2)$.

In our next lemma we establish the convergence of the P.W.Z. integral in the L_2-norm on $C_2(Q) \times C_2(Q)$ with respect to binary quadratic approximation in the first variable.

LEMMA 4.2. *Let* $L: C_2(Q) \times C_2(Q) \to \mathbb{R}$ *be given by*

$$L(f,g) := \int_Q f(s,t)\, \tilde{d}_2 g(s,t). \tag{4.1}$$

Then $L([f]_m, g) \to L(f,g)$ *in* L_2*-norm as* $m \to \infty$, *i.e.,*

$$\lim_{m\to\infty} \int_{C_2 \times C_2} |L(f,g) - L([f]_m, g)|^2 d(m_2 \times m_2)(f,g) = 0. \tag{4.2}$$

Proof. Using (2.6) we see that for $f \in C_2(Q)$ and each $m = 0, 1, 2, \cdots$

$$\left\{\int_{C_2} |L(f,g) - L([f]_m, g)|^2\, dm_2(g)\right\}^{1/2} = \|f - [f]_m\|_2. \tag{4.3}$$

It follows from (4.3) that

$$(4.4)\qquad \int_{C_2} |L(f,g)-L([f]_m,g)|^2\,dm_2(g)=\|f-[f]_m\|_2^2\to 0\quad \text{as } m\to\infty.$$

Also from (4.3) and the Minkowski inequality we get

$$(4.5)\qquad \begin{aligned} \int_{C_2} |L(f,g)-L([f]_m,g)|^2\,dm_2(g) &\leqq [\|f\|_2+\|[f]_m\|_2]^2 \\ &\leqq [\|f\|_\infty+\|[f]_m\|_\infty]^2 \\ &\leqq [\|f\|_\infty+\|f\|_\infty]^2 \\ &=4\|f\|_\infty^2. \end{aligned}$$

Now using (4.4), (4.5), the Fubini Theorem, Corollary 4.1 and the Dominated Convergence Theorem we obtain (4.2).

The following corollary to the proof of Lemma 4.2 gives the L_2-norm convergence on subrectangles of Q.

COROLLARY 4.2. *Let $R=[a,b]\times[\alpha,\beta]\subseteq Q$ and let $L\colon C_2(Q)\times C_2(Q)\to\mathbb{R}$ be given by*

$$(4.6)\qquad L(f,g):=\int_Q \chi_R(s,t)f(s,t)\,\tilde{d}_2 g(s,t).$$

Then

$$\lim_{m\to\infty}\int_{C_2\times C_2} |L(f,g)-L([f]_m,g)|^2 d(m_2\times m_2)(f,g)=0.$$

THEOREM 4.1. *For any rectangle $R=[a,b]\times[\alpha,\beta]\subseteq Q$ and for $m_2\times m_2$-a.e. $(f,g)\in C_2\times C_2$*

$$(4.7)\qquad \begin{aligned} \int_R f(s,t)\,\tilde{d}_2 g(s,t) &= f(b,\beta)g(b,\beta)-f(b,\alpha)g(b,\alpha) \\ &\quad -f(a,\beta)g(a,\beta)+f(a,\alpha)g(a,\alpha) \\ &\quad -\int_a^b [g(s,\beta)\,\tilde{d}f(s,\beta)-g(s,\alpha)\,\tilde{d}f(s,\alpha)] \\ &\quad -\int_\alpha^\beta [g(b,t)\,\tilde{d}f(b,t)-g(a,t)\,\tilde{d}f(a,t)] \\ &\quad +\int_R g(s,t)\,\tilde{d}_2 f(s,t). \end{aligned}$$

In fact, given any (f,g) for which (4.7) *holds and any ρ_1, $\rho_2>0$,* (4.7) *also holds for $(\rho_1 f,\rho_2 g)$. In fact for m_2-a.e.f,* (4.7) *holds for s-a.e.g and for m_2-a.e.g,* (4.7) *holds for s-a.e.f.*

Proof. Let $L\colon C_2\times C_2\to\mathbb{R}$ be defined by the left-hand side of (4.7) (i.e. L is given by (4.6)) for all (f,g) for which the P.W.Z. integral exists, and let $M\colon C_2\times C_2\to\mathbb{R}$ be defined by the right-hand side of (4.7) for all (f,g) for which all the P.W.Z. integrals exist. We first note that for each $f\in C_2(Q)$, $L(f,g)$ exists for s-a.e.$g\in C_2(Q)$ while for each $g \in C_2(Q)$, $M(f, g)$ exists for s-a.e.$f \in C_2(Q)$. To establish Theorem 4.1 it clearly suffices to establish (i), (ii), and (iii) below.

(i) For each $f \in C_2(Q)$ and each $m = 0, 1, 2, \cdots$, $L([f]_m, g) = M([f]_m, g)$ for s-a.e.$g \in C_2(Q)$.

(ii) $L([f]_m, g) \to L(f, g)$ in L_2-norm as $m \to \infty$ (this is simply Corollary 4.2 above).

(iii) For each $g \in C_2(Q)$, $M([f]_m, g) \to M(f, g)$ as $m \to \infty$ for s-a.e.$f \in C_2(Q)$.

Proof of (i). Let $f \in C_2(Q)$ and m be given. Then for s-a.e.$g \in C_2(Q)$, $L([f_m], g)$ exists and satisfies the equation

$$\begin{aligned} L([f]_m, g) &:= \int_R [f]_m(s, t)\, \tilde{d}_2 g(s, t) \\ &= \int_R [f]_m(s, t)\, dg(s, t). \end{aligned} \tag{4.8}$$

But, by the nonstochastic integration by parts formula [31, Thm. 4], we know that for *all* $(f, g) \in C_2 \times C_2$ and all m

$$\begin{aligned} \int_R [f]_m(s, t)\, dg(s, t) = {} & [f]_m(b, \beta) g(b, \beta) - [f]_m(b, \alpha) g(b, \alpha) \\ & - [f]_m(a, \beta) g(a, \beta) + [f]_m(a, \alpha) g(a, \alpha) \\ & - \int_a^b [g(s, \beta) d[f]_m(s, \beta) - g(s, \alpha) d[f]_m(s, \alpha)] \\ & - \int_\alpha^\beta [g(b, t) d[f]_m(b, t) - g(a, t) d[f]_m(a, t)] \\ & + \int_R g(s, t) d[f]_m(s, t). \end{aligned} \tag{4.9}$$

Next by Lemma 3.2 and the corresponding result for P.W.Z. integrals of functions of one variable we see that for *all* $(f, g) \in C_2 \times C_2$ and all m, $M([f]_m, g)$ exists and equals the right-hand side of (4.9) above. Thus, using (4.8) and (4.9), we see that for fixed f and m, $L([f]_m, g) = M([f]_m, g)$ for s-a.e.$g \in C_2(Q)$ and so (i) is established.

Proof of (iii). To establish (iii) it suffices to show that for each $g \in C_2(Q)$, each of the seven terms involved in the definition of $M([f]_m, g)$ converges to the corresponding term in the definition of $M(f, g)$ for s-a.e.$f \in C_2(Q)$. This is easy to see for the first 4 terms. Corollary 3.1 assures that this is true for the 7th term also, i.e.,

$$\lim_{m \to \infty} \int_R g(s, t)\, \tilde{d}_2 [f]_m(s, t) = \int_R g(s, t)\, \tilde{d}_2 f(s, t)$$

for s-a.e.$f \in C_2(Q)$. Next, using a result of Cameron and Storvick [1, Lemma 2.3] together with [28, p. 306], it is quite easy to see that, for example,

$$\lim_{m \to \infty} \int_a^b g(s, \beta) d[f]_m(s, \beta) = \int_a^b g(s, \beta)\, \tilde{d} f(s, \beta)$$

for s-a.e.$f \in C_2(Q)$. Similarly we get the desired convergence for the other integrals involved in the 5th and 6th terms of $M([f]_m, g)$ and $M(f, g)$. Thus (iii) is established, which concludes the proof of Theorem 4.1.

Choosing $R = Q = [0, 1]^2$ in Theorem 4.1 and keeping in mind that f and g vanish on the left-hand and lower edges of Q, we get the following corollary.

COROLLARY 4.3. *For $m_2 \times m_2$-a.e.$(f, g) \in C_2 \times C_2$*

$$\int_{[0,1]^2} f(s,t)\,\tilde{d}_2 g(s,t) = f(1,1)g(1,1) - \int_0^1 g(s,1)\,\tilde{d}f(s,1) - \int_0^1 g(1,t)\,\tilde{d}f(1,t) + \int_{[0,1]^2} g(s,t)\,\tilde{d}_2 f(s,t). \tag{4.10}$$

In addition, given any (f, g) for which (4.10) *holds and any $\rho_1, \rho_2 > 0$,* (4.10) *also holds for $(\rho_1 f, \rho_2 g)$. In fact for m_2-a.e.f,* (4.10) *holds for s-a.e.g and for m_2-a.e.g,* (4.10) *holds for s-a.e.f.*

Next, using Theorem 3.1 and Theorem 4.1, we obtain the following interesting corollary.

COROLLARY 4.4. *For m_2-a.e.$g \in C_2(Q)$, the function $\int_Q f(s,t)\,\tilde{d}_2 g(s,t)$ is continuous with respect to binary quadratic approximation for s-a.e.f in $C_2(Q)$. That is to say, for almost all $g \in C_2(Q)$,*

$$\lim_{m\to\infty} \int_Q [f]_m(s,t)\,\tilde{d}_2 g(s,t) = \int_Q f(s,t)\,\tilde{d}_2 g(s,t)$$

for s-a.e.f in $C_2(Q)$.

Concluding remarks. Using the techniques of § 4 together with Lemma 2.3 of [1, p. 7], one can quite easily obtain an alternate proof of the parts formula for the stochastic integral of functions of one variable [7, p. 268] which follows.

THEOREM 4.1(a). *For $0 \leqq a < b \leqq 1$ and $m_1 \times m_1$-a.e.(x, w) in $C_1[0,1] \times C_1[0,1]$*

$$\int_a^b x(s)\,\tilde{d}w(s) = x(b)w(b) - x(a)w(a) - \int_a^b w(s)\,\tilde{d}x(s). \tag{4.11}$$

In fact for m_1-a.e.x, (4.11) *holds for s-a.e.w and for m_1-a.e.w,* (4.11) *holds for s-a.e.x.*

COROLLARY 4.4(a). *For m_1-a.e.$w \in C_1[0,1]$, the function $H(x) := \int_0^1 x(s)\,\tilde{d}w(s)$ is continuous with respect to binary polygonal approximation for s-a.e.$x \in C_1[0,1]$.*

Proof. Theorem 4.1(a) and Lemma 2.3 of [1, p. 7].

REFERENCES

[1] R. H. CAMERON AND D. A. STORVICK, *A simple definition of the Feynman integral, with applications*, Mem. Amer. Math. Soc., 288, 46 (1983), pp. 1-46.

[2] ———, *A new translation theorem for the analytic Feynman integral*, Rev. Roumaine Math. Pures Appl., 27 (1982), pp. 937-944.

[3] K. S. CHANG, *Scale-invariant measurability in Yeh-Wiener space*, J. Korean Math. Soc., 19 (1982), pp. 61-67.

[4] K.S. CHANG, G. W. JOHNSON AND D. L. SKOUG, *The Feynman integral of quadratic potentials depending on two time variables*, Pacific J. Math., 122 (1986), pp. 11-33.

[5] J. A. CLARKSON AND C. R. ADAMS, *On definitions of bounded variation for functions of two variables*, Trans. Amer. Math. Soc., 35 (1933), pp. 824-854.

[6] ———, *Properties of functions $f(x, y)$ of bounded variation*, Trans. Amer. Math. Soc., 36 (1934), pp. 711-730.

[7] I. I. GIHMAN AND A. V. SKOROKHOD, *Stochastic Differential Equations*, Ergebnisse der Mathematik u. ihrer Grenzgebiete No. 72, Springer-Verlag, Berlin, New York, 1972.

[8] J. GLIMM AND A. JAFFE, *Quantum Physics, A Functional Integral Point of View*, Springer, New York, 1981.

[9] E. W. HOBSON, *The Theory of Functions of a Real Variable and the Theory of Fourier Series, Vol.* I, Dover, New York, 1957.

[10] R. L. JEFFERY, *Functions of bounded variation and non-absolutely convergent integrals in two or more dimensions*, Duke Math. J., 5 (1939), pp. 753-774.

[11] G. W. Johnson and D. L. Skoug, *Scale-invariant measurability in Wiener space*, Pacific J. Math., 83 (1979), pp. 157-176.

[12] ———, *Notes on the Feynman integral*, I, Pacific J. Math., 93 (1981), pp. 313-324.

[13] ———, *Notes on the Feynman integral*, II, J. Funct. Anal., 41 (1981), pp. 277-289.

[14] ———, *Notes on the Feynman integral*, III: *The Schrödinger equation*, Pacific J. Math., 105 (1983), pp. 321-358.

[15] R. Kindermann and J. L. Snell, *Markov random fields and their applications*, Contemporary Mathematics Series, Vol. 1, Amer. Math. Soc., Providence, RI, 1980.

[16] Hui-Hsiung Kuo, *Gaussian measures in Banach spaces*, Lecture Notes in Math., 463, Springer, Berlin, New York, 1975.

[17] M. S. MacPhail, *Functions of bounded variation in two variables*, Duke Math. J., 8 (1941), pp. 215-222.

[18] E. Nelson, *The construction of quantum fields from Markov fields*, J. Funct. Anal., 12 (1973), pp. 97-112.

[19] ———, *The free Markov field*, J. Funct. Anal., 12 (1973), pp. 211-217.

[20] R. E. A. C. Paley, N. Wiener and A. Zygmund, *Notes on random functions*, Math. Z., 37 (1933), pp. 647-688.

[21] C. Park, *A generalized Paley-Wiener-Zygmund integral and its applications*, Proc. Amer. Math. Soc., 23 (1969), pp. 388-400.

[22] ———, *On Fredholm transformations in Yeh-Wiener space*, Pacific J. Math., 40 (1972), pp. 173-195.

[23] C. Park and D. L. Skoug, *Distribution estimates of barrier-crossing probabilities of the Yeh-Wiener process*, Pacific J. Math., 78 (1978), pp. 455-466.

[24] W. J. Park, *A multi-parameter Gaussian process*, Ann. Math. Statist., 41 (1970), pp. 1582-1595.

[25] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* II: *Fourier Analysis, Self-adjointness*, Academic Press, New York, 1975.

[26] Y. A. Rozanov, *Markov Random Fields* (translated by C. M. Elson), Springer, New York, 1981.

[27] B. Simon, *The* $P(\phi)_2$ *Euclidean* (*Quantum*) *Field Theory*, Princeton Series in Physics, Princeton Univ. Press, Princeton, NJ, 1974.

[28] D. L. Skoug, *Converses to measurability theorems for Yeh-Wiener space*, Proc. Amer. Math. Soc., 57 (1976), pp. 304-310.

[29] G. Velo and A. Wightman, eds., *Constructive Quantum Field Theory*, Lecture Notes in Phys., 25, Springer, Berlin, New York, 1973.

[30] J. Yeh, *Wiener measure in a space of functions of two variables*, Trans. Amer. Math. Soc., 95 (1960), pp. 433-450.

[31] ———, *Cameron-Martin translation theorems in the Wiener space of functions of two variables*, Trans. Amer. Math. Soc., 107 (1963), pp. 409-420.

[32] ———, *Orthogonal development of functionals and related theorems in the Wiener spaces of two variables*, Pacific J. Math., 13 (1963), pp. 1427-1436.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

004

SOME APPROXIMATION FORMULA FOR STOCHASTIC EIGENVALUES*

DAVID C. BARNES†

Abstract. We consider some random eigenvalue problems of the form $L(\cdot)=\lambda M(\cdot)$, where $L(\cdot)$ and $M(\cdot)$ may be ordinary or partial differential operators which depend on a (perhaps multi-dimensional) random variable ω. We generalize some formulas due to Boyce [*Probabilistic Methods in Applied Mathematics*, Academic Press, New York, 1968, pp. 1-73] which give estimates for the eigenvalues of the form $\lambda(\omega)=\lambda^*+K(\omega)+O(\|\omega\|^2)$. Here $\lambda(\omega)$ is an eigenvalue corresponding to the random variable ω while λ^* is an eigenvalue corresponding to some approximating deterministic problem. The term $K(\omega)$ will be $O(\|\omega\|)$, although it may not be linear in ω. These formulas may be used with very general boundary conditions, including those which contain random coefficients and those which may be nonlinear and nonhomogeneous. The boundary conditions also may contain the eigenvalue parameter λ and they need not be self-adjoint. We will first give the theory for ordinary equations, then generalize to partial differential equations, first in a deterministic domain, then in a random domain.

Key words. stochastic eigenvalues, approximation

1. Introduction.

Consider the eigenvalue problem

$$L(y)=\lambda g y, \quad L(y)=-(fy')'-qy, \quad 0<x<l. \tag{1}$$

We assume the coefficients f, g and q depend on x as well as a random variable $\omega=(\omega_1, \omega_2, \cdots, \omega_N)$, taken from a sample space $\Omega=\Omega_1\times\Omega_2\times\cdots\times\Omega_N$ where, for each $x\in[0, l]$, probability measures $\mu_i(\cdot)$ are defined on Ω_i. The ω_i may or may not be independent. When appropriate boundary conditions are given, and with some mild restrictions on the coefficients, the problem will have eigenvalues $\lambda(\omega)$ which will then be random variables defined on Ω. In this work, we will be concerned with finding approximations of the form $\lambda(\omega)=\lambda^*+K(\omega)+O(\|\omega\|^2)$. Here, λ^* is an eigenvalue of an approximating deterministic problem

$$L^*(y^*)=\lambda^* g^* y^*, \qquad L^*(y^*)=-(f^*y^{*\prime})'-q^*y^*. \tag{2}$$

We will first develop the theory for (1), then generalize it to problems of the form $L(y)=\lambda M(y)$, where $L(\cdot)$ and $M(\cdot)$ are (possibly partial) differential operators having random coefficients. Finally, we will give a modest treatment of the simple equation $\nabla^2 u+\lambda u=0$ with $u=0$ on $\partial\mathcal{R}$, where $\mathcal{R}$ is a random domain. Throughout we will use the notation $\Delta(\cdot)=(\cdot)-(\cdot)^*$, and *-ed quantities will all be deterministic; so, for example, $\Delta L=L-L^*$, $\Delta\lambda=\lambda-\lambda^*$ and so on. We denote the mean of a random variable by $\langle\cdot\rangle$ and also use the notations

$$\|\omega\|^2=\max_{i,j}\{|\omega_i\omega_j|\}, \qquad (u, v)=\int_0^l u(x)v(x)\,dx.$$

2. The second order equation.

Methods used by D. C. Barnes [3] can be used to prove the following theorem.

THEOREM 1. *Let y, λ and y^*, λ^* be eigenpairs for* (1), (2). *Define a random variable $J(\omega)$ by*

$$J(\omega)=\int_0^l [\lambda^* g y^{*2}+q y^{*2}-f(y^{*\prime 2})]\,dx. \tag{3}$$

* Received by the editors March 28, 1985; accepted for publication April 7, 1986.

† Mathematics Department, Washington State University, Pullman, Washington 99164-2930.

Define also a boundary term BT and a random variable $K(\omega)$ by

$$BT = f^*y^{*\prime}y - fy'y^* - f^*y^{*\prime}y^*|_{x=0}^{x=l}, \qquad K(\omega) = \lambda^* + BT - J(\omega). \tag{4}$$

Suppose that y^* is normalized so that

$$\int_0^l g^*y^{*2}\,dx = 1. \tag{5}$$

Then

$$\lambda(\omega) = K(\omega) + O_2 = \lambda^* + BT - J(\omega) + O_2 \tag{6}$$

where the term O_2 is defined by

$$O_2 = \int_0^l [\Delta\lambda\,\Delta y g^* y^* + \lambda^*\Delta g\Delta y y^* + \Delta q \Delta y y^* - \Delta f \Delta y' y^{*\prime}]\,dx. \tag{7}$$

In the context of [3], ω was simply a real parameter and (6) was used to develop some variational properties of the eigenvalues λ. However, (6) shows that the functional $K(\omega)$ is tangent to $\lambda(\omega)$ at $\omega = 0$ and thus will provide a good approximation formula when $\|\omega\|$ is small. Now we need to make O_2 small by selecting the base problem (2). One good way to do this is to take

$$f^* = \langle f \rangle, \qquad g^* = \langle g \rangle, \qquad q^* = \langle q \rangle. \tag{8}$$

As an example, we will take the simple boundary conditions, $y(0) = y(l) = 0$, and use (6) to approximate $\langle \lambda \rangle$. We see that

$$\langle \lambda(\omega) \rangle = \lambda^* - \int_0^l [\lambda^*\langle g \rangle y^{*2} + \langle q \rangle y^{*2} - \langle f \rangle (y^{*\prime 2})]\,dx + O_2. \tag{9}$$

Now multiply (2) by y^* and integrate. This shows that the integral term in (9) vanishes, so $\langle \lambda(\omega) \rangle = \lambda^* + \langle O_2 \rangle$. This estimate has been given by Boyce [4]. Generally, in the case of homogeneous, linear, self-adjoint, and deterministic boundary conditions, and in the choice of base problem (8), the result (9) is well known.

It is, however, not at all necessary that the boundary conditions be homogeneous, linear, self-adjoint or deterministic. Equation (6) holds in any case. The boundary conditions might even involve the eigenvalue parameter λ.

There are only two conditions which are necessary in order to use the approximation (6) effectively. First of all BT will, in general, involve the values of y at $x = 0$ and at $x = l$. So we must be able to use the boundary conditions to eliminate, or at least approximate, the terms in BT which involve y. The second requirement is that the eigenvalue problem must be well posed when considered as a function of ω. That is, the eigenvalues, the eigenfunctions and their derivatives must depend continuously on ω. This will insure that the error term O_2 is small. Other than these requirements, the boundary conditions are quite arbitrary.

Suppose we are given constants a, b and random variables ω_i taken from sample spaces (Ω_i, μ_i) with $\omega = (\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$. Consider the boundary conditions

$$y(0) + (a + \omega_2) f(0, \omega_1) y'(0) = 0, \qquad (b + \omega_3) y(l) + f(l, \omega_1) y'(l) = 0. \tag{10}$$

We select the base problem by letting $\omega = 0$, so that

$$\begin{gathered} f^*(x) = f(x, 0), \qquad g^*(x) = g(x, 0), \qquad q^*(x) = q(x, 0), \\ y^*(0) + af^*(0) y^{*\prime}(0) = 0, \qquad by^*(l) + f^*(l) y^{*\prime}(l) = 0. \end{gathered} \tag{11}$$

Using (10) and (11) in (4), it follows that

$$BT=(by^{*2}+\omega_3 yy^*)|_{x=l}-(af^{*2}y^{*\prime 2}-\omega_2 ff^*y^{*\prime}y')|_{x=0}.$$

We now use the approximations $y=y^*+O(\omega)$ and $y'=y^{*\prime}+O(\omega)$ to eliminate the terms involving y and obtain the first order approximation

$$BT=y^{*2}(b+\omega_3)|_{x=l}-(a-\omega_2)f^{*2}y^{*\prime 2}|_{x=0}+O_2.$$

Since we have eliminated the terms in BT which involve y, we may substitute this formula in (6) to obtain first order approximation to $\lambda(\omega)$ which can be easily computed.

As a second example, consider a slender column subject to a compressive load λ which may cause it to buckle. The critical buckling load is determined by the smallest nonzero eigenvalue of an equation of the form (1). The eigenfunction $y(x)$ represents the bending moment of the column in the buckled state. If the load λ is applied exactly on the center line of the column, then the boundary conditions will be $y^*(0)=y^*(l)=0$. Suppose, however, that the load at $x=l$ is applied at a small, random distance ω away from the center line. If the column is pinned at both ends, then the boundary conditions are [2], [6]

$$y(0)=0,\quad y(l)=\lambda\omega,\qquad y^*(0)=0,\quad y^*(l)=0. \tag{12}$$

We suppose that the shape of the column is deterministic, so that the coefficients f, g, q do not depend on ω. The eigenfunction y^* is uniquely determined by (2) and (5) up to a factor of ± 1. Suppose first that $\omega>0$. We then take y^* to be the eigenfunction satisfying $y^*>0$ for $0<x<l$ so that $y^{*\prime}(l)<0$. However, if $\omega<0$, then we will take y^* to satisfy $y^*<0$ so that $y^{*\prime}(l)>0$. This will insure, at least for small ω, that y and y^* will have no zeros on $0<x<l$ so that Δy and $\Delta y'$ will be small. It also insures that the approximation formulas will be symmetric about $\omega=0$, as is the eigenvalue itself, $\lambda(-\omega)=\lambda(\omega)$.

Since the problem (12) is not homogeneous, the eigenfunction y cannot be normalized at will. We still, however, use the relation

$$\int_0^l gy^2\,dx=1, \tag{13}$$

in order to force the approximations of y to y^*. The three conditions (12), (13) will serve to determine the two constants in the general solution of (1) as well as λ. These conditions, together with the above sign conventions, will determine the eigenfunction y uniquely.

Using (4), we find $BT\approx\lambda(\omega)f^*(l)y^{*\prime}(l)\omega$. Since the choice of y^* depends on sign(ω), we recast BT into the general form (which works for either y^*) $BT\approx -\lambda(\omega)f^*(l)|y^{*\prime}(l)\omega|$. Putting this into (6) and solving the resulting equation for $\lambda(\omega)$ yields the first order approximation

$$\lambda(\omega)=\frac{\lambda^*-J(\omega)+O_2}{1+f^*(l)|y^{*\prime}(l)\omega|}=\frac{\lambda^*-J(\omega)}{1+f^*(l)|y^{*\prime}(l)\omega|}+O_2. \tag{14}$$

The form of this equation, and physical intuition, suggests that $\lambda(\omega)\leqq\lambda^*$ but this has not been proved.

Consider an example of (14). Suppose that the column is uniform. That is, $f=g=1$, $q=0$, and use nondimensional coordinates, so that

$$y''+\lambda y=0,\quad y(0)=0,\quad y(1)=\lambda\omega,\quad\text{and}\quad y^*=\pm\sqrt{2}\sin\pi x,\quad \lambda^*=\pi^2.$$

Solving for y and $\lambda(\omega)$, we find, for given ω, that $\lambda(\omega)$ is the smallest positive root of the equation

$$(15)\qquad t^3\omega^2(2t-\sin t)=4\sin^2 t,\quad t=\sqrt{\lambda}\quad\text{and}\quad y=2\sqrt{\frac{t}{(2t-\sin 2t)}}\sin tx.$$

Using (14) we find that

$$(16)\qquad \lambda(\omega)=\frac{\lambda^*}{1+\sqrt{2}\pi|\omega|}+O_2.$$

We will now check the accuracy of this approximation. Rather than attempt to solve (15) for $\lambda(\omega)$, it is more convenient to use the inverse functions for $|\omega|$ as a function of $t=\sqrt{\lambda}$. Solving for $|\omega|$, using both (15) and (16) we find

$$|\omega|=\frac{2\sin t}{\sqrt{t^3(2t-\sin 2t)}},\qquad |\omega|=\frac{\pi^2-t^2}{\sqrt{2}\pi t^2}+O_2.$$

Now one can easily verify that the two right-hand sides are tangent to each other at the point $t=\pi$, $\omega=0$. Thus the approximation (16) will be good for small ω.

The sign conventions used in this example essentially amounted to decomposing the sample space Ω into its positive and negative parts, Ω^+ and Ω^-, and then using two distinct approximation formulas on each part. Such decompositions can be useful in other situations as well.

Suppose that either the coefficient functions f, g, q, or the boundary conditions are not continuous in ω. In such an event, the eigenvalue problem may not be well posed and the methods used here would not apply. Suppose, however, that we can decompose Ω into a disjoint union,

$$(17)\qquad \Omega=\bigcup_{i=1}^{N}\Omega_i,\qquad \Omega_i\cap\Omega_j=\varnothing\quad\text{for } i\neq j$$

in such a way that the problem is well posed on each Ω_i. We then select approximating base problems for each Ω_i,

$$(18)\qquad (f_i^*y_i^{*\prime})'+(\lambda_i^*g_i^*+q_i^*)y_i^*=0,\qquad U_1^{(i)}(y_i^*)=U_2^{(i)}(y_i^*)=0$$

where, for each i, we might select a fixed value of $\omega_i\in\Omega_i$ and define the approximating problems by

$$f_i^*(x)=f(x,\omega_i),\qquad g_i^*(x)=g(x,\omega_i),\qquad q_i^*(x)=q(x,\omega_i).$$

Now construct the piecewise O_2 approximation to $\lambda(\omega)$ using

$$(19)\quad \lambda(\omega)=\lambda_i^*+BT_i-J_i(\omega)+O_{2,i},\quad \omega\in\Omega_j,\qquad J_i(\omega)=\int_0^l \lambda_i^*gy_i^*+qy_i^*-fy_i^{*\prime 2}\,dx.$$

The error terms $O_{2,i}$ will be $O((\omega-\omega_i)^2)$ on Ω_i. Even though such a piecewise approximation may not be continuous on all Ω, it will still be a global first order approximation to $\lambda(\omega)$.

This also suggests numerical procedures for high accuracy computations. Suppose, for example, that we need to compute $\langle\lambda\rangle$. We could then decompose Ω as in (17) so that $\mu(\Omega_i)=1/N$. Then solve the N base problems (18) for y_i^* and take the mean in (19) to obtain

$$(20)\qquad \langle\lambda\rangle=\sum_{i=1}^{N}\int_{\Omega_i}\langle\lambda_i^*+BT_i-J_i(\omega)\rangle\,d\mu+O\left(\frac{1}{N}\right).$$

Since each error term in (19) is $O(1/N^2)$, the error term in (20) will be $O(1/N)$. The obvious thing to do at this point, is to use (20) together with a Romberg style extrapolation scheme to compute more accurate values for $\langle\lambda\rangle$.

3. More general problems. These methods will also work on more general problems of the form

$$L(y)=\lambda M(y), \qquad U_p(y,\omega,\lambda)=0, \quad p=1,2,\cdots,2m, \tag{21}$$
$$L(y)=\sum_{i=0}^{m}(-1)^i(f_iy^{(i)})^{(i)}, \qquad M(y)=\sum_{j=0}^{m'}(-1)^j(g_jy^{(j)})^{(j)}.$$

Here, f_i and g_j are random functions, and we will assume that $m>m'\geqq 0$. The boundary conditions may be quite arbitrary, being subject to the same conditions outlined in the second order case.

Consider any deterministic problem which approximates (21),

$$L^*(y^*)=\lambda^*M^*(y^*), \qquad U_p^*(y^*,\lambda^*)=0, \quad p=1,2,\cdots,2m, \tag{22}$$
$$L^*(y^*)=\sum_{i=0}^{m}(-1)^i(f_i^*y^{*(i)})^{(i)}, \qquad M^*(y^*)=\sum_{i=0}^{m'}(-1)^j(g_j^*y^{*(j)})^{(j)}.$$

We have the following theorem.

THEOREM 2. *Let y, λ and y^*, λ^* be eigenpairs for* (21) *and* (22). *Define a random variable $J(\omega)$ by*

$$J(\omega)=\int_0^l \sum_{j=0}^{m'} g_j(y^{*(j)})^2-\sum_{i=0}^{m} f_i(y^{*(i)})^2\,dx. \tag{23}$$

Define boundary terms $BT1$, $BT2$ and $BT3$ by the following equations:

$$(L^*(y),y^*)=BT1+(y,L^*(y^*)), \tag{24}$$

$$(M^*(y),y^*)=BT2+(y,M^*(y^*)), \tag{25}$$

$$(L(y^*),y^*)-\lambda^*(M(y^*),y^*)=BT3-J(\omega). \tag{26}$$

Suppose that y^* is normalized so that

$$(M^*(y^*),y^*)=1, \qquad (L^*(y^*),y^*)=\lambda^*. \tag{27}$$

Then

$$\lambda(\omega)=\lambda^*+BT1-\lambda^*BT2+(L(y^*),y^*)-\lambda^*(M(y^*),y^*)+O_2, \tag{28}$$

$$\lambda(\omega)=\lambda^*+BT1-\lambda^*BT2+BT3-J(\omega)+O_2. \tag{29}$$

The error term O_2 is given by

$$O_2=(\Delta L(\Delta y)-\Delta\lambda M^*(\Delta y)-\Delta\lambda\,\Delta M(y^*)-\lambda\,\Delta M(\Delta y),y^*) \tag{30}$$

where $\Delta(\cdot)=(\cdot)-(\cdot)^*$.

The proof consists of a rather straightforward, but lengthy, computation. First multiply (22) by y and (21) by y^*. Then subtract the equations and integrate to find

$$(L(y),y^*)-(L^*(y^*),y)=\lambda(M(y),y^*)-\lambda^*(M^*(y^*),y).$$

We now substitute the Δ notation into this equation, giving

$$((L^*+\Delta L)(y^*+\Delta y),y^*)-(L^*(y^*),y^*+\Delta y)$$
$$=(\lambda^*+\Delta\lambda)((M^*+\Delta M)(y^*+\Delta y),y^*)-\lambda^*(M^*(y^*),y^*+\Delta y).$$

Now multiply out these expressions, collecting all terms of second order into O_2 as given by (30). Next, use the relations $\Delta(\cdot) = (\cdot) - (\cdot)^*$ to eliminate all of the remaining terms which involve Δ. Finally, use (24), (25) and (27) to simplify the formula. After some manipulation, (28) follows. Finally, (26) implies (29).

In the important special case $m = 2$ and $m' = 1$, we find that

$$BT1 = (f_2^* y'')' y^* - (f_2^* y^{*''})' y + f_2^* y^{*''} y' - f_2^* y'' y^{*'} + f_1^* y y^{*'} - f_1^* y' y^* \Big|_{x=0}^{x=l}, \tag{31}$$

$$BT2 = g_1^* y^{*'} y - g_1^* y^* y' \Big|_{x=0}^{x=l}, \tag{32}$$

$$BT3 = (f_2 y^{*''})' y^* - f_2 y^{*'} y^{*''} - f_1 y^{*'} y^* + \lambda^* g_1 y^{*'} y^* \Big|_{x=0}^{x=l}. \tag{33}$$

As an example, consider a uniform slender column subject to an axial compressive load λ which may cause it to buckle. Suppose that it is pinned at each end and that it is supported on an elastic foundation which provides, at each point x, a random restoring force $F(x, \omega_1)y$, which is directly proportional to the displacement y. The critical buckling load is determined [6] by the first eigenvalue of the system:

$$y'''' + F(x, \omega_1) y = \lambda(-y''), \qquad y(0) = y''(0) = y(l) = y''(l) = 0.$$

Take the base problem to correspond to $\omega_1 = 0$, so that

$$y^{*''''} + F^*(x) y^* = \lambda^*[-y^{*''}], \qquad F^*(x) = F(x, 0). \tag{34}$$

Using (29) we see that

$$\lambda(\omega) = \lambda^* - \int_0^l \lambda^* (y^{*'})^2 - F(x, \omega_1) y^{*2} + (y^{*''})^2 \, dx + O_2.$$

Multiplying (34) by y^* and integrating shows that

$$\lambda(\omega) = \lambda^* - \int_0^l (F(x, \omega_1) - F^*(x)) y^{*2} \, dx + O_2.$$

We can also deal with random boundary conditions. Suppose, for example, that

$$y'(0) + (a + \omega_2) y''(0) = 0, \qquad (b + \omega_3) y'(l) + y''(l) = 0, \qquad y(0) = 0, \quad y(l) = 0.$$

Using (24)-(26) and the approximations $y = y^* + O(\|\omega\|)$, we soon find that

$$BT1 - \lambda^* BT2 + BT3 = (b + \omega_3) y^{*'2}(l) - (a + \omega_2) y^{*''2}(0) + O_2.$$

This approximation can now be used in (29) to obtain an approximation for $\lambda(\omega)$.

A second example, which has been used to study vibrations of a helicopter roter (see [1] and the references listed there) is given by

$$\begin{gathered} y'''' - ((\tfrac{1}{2})\alpha^2(1 - x^2 + \gamma^2) y')' = \lambda y, \qquad 0 < x < 1, \\ y(0) = y'(0) - a y''(0) = y''(1) = 0, \qquad y'''(1) = (\tfrac{1}{2})\gamma^2 \alpha^2 y'(1) - (\tfrac{1}{2})\lambda \gamma^2 y(1). \end{gathered} \tag{35}$$

We suppose that α and γ are fixed but that a is a random variable, $a = a^* + \omega$, with $\omega \in \Omega$. Letting $\omega = 0$ correspond to the base problem, we use (31)-(33) to find that $BT2 = 0$ and that

$$BT1 = (\lambda^* - \lambda)(\tfrac{1}{2})\gamma^2 y^{*2}(1) - \omega y^{*''2}(0) + O_2, \qquad BT3 = -(\tfrac{1}{2})\lambda^* \gamma^2 y^{*2}(1).$$

Substituting this into (29) and solving the resulting equation for λ yields the first order approximation

$$\lambda(\omega) = \frac{\lambda^* - \omega y^{*''2}(0) - J(\omega) + O_2}{1 + \frac{1}{2}\gamma^2 y^{*2}(1)}. \tag{36}$$

Now $J(\omega)$ is actually independent of ω so that $J(\omega)=J(0)$. Letting $\omega=0$ in (26) shows that $J(\omega)=BT3=-\frac{1}{2}\lambda^*\gamma^2y^{*2}(1)$. Substituting into (36) we obtain

$$(37) \qquad \lambda(\omega)=\lambda^*-\omega\frac{y^{*\prime 2}(0)}{1+\frac{1}{2}\gamma^2y^{*2}(1)}+O(\omega^2).$$

It follows from this equation that, at $\omega=0$, $d\lambda(\omega)/d\omega<0$. Therefore, $\lambda(\omega)$ is a decreasing function of ω. Thus, we see that λ^* is a decreasing function of a^*.

With more involved calculations, we could allow the coefficients in (35) to be random also. We will not pursue that idea here.

It is interesting to note that Theorem 1 is not, exactly, a special case of Theorem 2. To see this, let $m=1$, $m'=0$ in Theorem 2, and take

$$L(y)=-(fy')'-qy, \qquad M(y)=gy,$$

$$L^*(y^*)=-(f^*y^{*\prime})-q^*y^*, \qquad M^*(y^*)=g^*y^*.$$

Computing the boundary term given in Theorem 2 we find

$$(38) \qquad BT1-\lambda^*BT2+BT3=f^*y^{*\prime}y-f^*y'y^*-fy^{*\prime}y^*\Big|_{x=0}^{x=l}.$$

However, computing the boundary term BT given in Theorem 1 yields a different result. But taking the difference between the two boundary terms and doing a direct computation shows that the difference $=y^*\Delta f\Delta y'|_0^l$. Since this term is of second order, we see that either boundary term could be used to get a first order approximation to $\lambda(\omega)$. This discrepancy arises from the fact that in the work [3] the O_2 terms were collected after the integration by parts whereas in Theorem 2 the O_2 terms were collected before the integration by parts. It seems difficult, at this point, to decide which method should give a better result. Such questions would have to be dealt with by a more careful analysis of the O_2 terms.

4. Partial differential equations. These methods generalize easily to partial differential equations in a deterministic domain $\mathcal{R}$. Using Green's Theorem in place of integration by parts shows that Theorem 2 can be used when $L(\cdot)$ and $M(\cdot)$ are partial differential operators. Consider the special case of the two-dimensional equation

$$(fu_x)_x+(fu_y)_y+(\lambda g+q)u=0, \quad \text{in } \mathcal{R} \text{ with } u+(a+\omega)u_{\vec{\eta}}=0 \text{ on } \partial\mathcal{R}.$$

Here, f, g, q, and ω are random but a is a deterministic function of s, arc length on $\partial\mathcal{R}$. For now, $\mathcal{R}$ is a deterministic domain. Using (29), we see that

$$\lambda=\lambda^*+\int_{\partial\mathcal{R}}(a+\omega)\left(\frac{\partial u^*}{\partial\vec{\eta}}\right)^2 ds+\iint_{\mathcal{R}}\lambda^*gu^{*2}+qu^{*2}-f(u_x^{*2}+u_y^{*2})\,dA+O_2$$

where

$$\iint_{\mathcal{R}} g^*u^{*2}\,dA=1.$$

If l, the length of the interval in a one-dimensional problem, was a random function of some $\omega\in\Omega$, then a simple change of variable $s=x/l$ would give a new problem on the fixed domain $0\leqq s\leqq 1$ with random coefficients. Theorem 2 would then apply. In the case of partial differential equations, such a change of variables is much more difficult. One can, however, appeal to the Hadamard variational formula to accomplish much the same kind of manipulation.

Following Garabedian [5, Chap. 15], we consider a fixed two-dimensional domain $\mathcal{R}^*$ and let $\mathcal{R}(\omega)$ be the region whose boundary, $\partial\mathcal{R}(\omega)$, is obtained by shifting $\partial\mathcal{R}^*$ an "infinitesimal" distance $\delta\vec{\eta}=\rho(s,\omega)\vec{\eta}$ along its inner normal $\vec{\eta}$, so that $\mathcal{R}(0)=\mathcal{R}^*$. Here, s is arc length on $\partial\mathcal{R}$ and ω is a random variable. We assume that $\partial\rho/\partial\omega$ exists near $\omega=0$.

Let $\lambda_1(\mathcal{R}(\omega))$ denote the first eigenvalue of $\nabla^2u+\lambda u=0$, with $u=0$ on $\partial\mathcal{R}(\omega)$, and let u_1 and u_1^* be the eigenfunctions corresponding to $\mathcal{R}$ and $\mathcal{R}^*$. It follows from the Hadamard variational formula [5, Chap. 15] that

$$\Delta\lambda_1=\lambda_1(\mathcal{R}(\omega))-\lambda_1(\mathcal{R}^*)=\delta\lambda_1+O(\omega^2),\qquad \delta\lambda_1=\int_{\partial\mathcal{R}^*}\rho(s,\omega)\left(\frac{\partial u_1^*}{\partial\vec{\eta}}\right)^2ds.$$

Thus we obtain the first order approximation formula

$$\lambda_1(\mathcal{R}(\omega))=\lambda_1(\mathcal{R}^*)+\int_{\partial\mathcal{R}^*}\rho(s,\omega)\left(\frac{\partial u_1^*}{\partial\vec{\eta}}\right)^2ds+O(\omega^2). \tag{39}$$

As an example of this, consider a triangular domain $\mathcal{R}^*$ bounded by $x=0$, $y=0$, and $x+y=\pi$. Suppose the sides $x=0$, $y=0$ are fixed but that the diagonal side has a random variation ω. Thus $\rho(s,\omega)$ is zero on $x=0$, $y=0$ but $\rho(s,\omega)=\omega$ on $x+y=\pi$, and $\mathcal{R}(\omega)$ is the domain bounded by the lines $x=0$, $y=0$, and $x+y=\pi-\sqrt{2}\omega$. The minus sign is due to the inward directed normal. There is a problem at the corner points since $\partial\mathcal{R}^*$ is not smooth there and ρ is not continuous. However, there do exist analytic approximations to $\partial\mathcal{R}^*$, $\partial\mathcal{R}(\omega)$ and ρ, and the error introduced using them will be $O(\omega^2)$.

Solving for u_1^*, λ_1^*, we find that $u_1^*=2(\sin x\sin 2y+\sin y\sin 2x)$ and that $\lambda_1^*=5$. Now the change of variables $x=\bar{x}(1-\pi\omega/\sqrt{2})$, $y=\bar{y}(1-\pi\omega/\sqrt{2})$ shows that

$$\lambda_1(\mathcal{R}(\omega))=\frac{\lambda_1^*}{(1-\pi\omega/\sqrt{2})^2}. \tag{40}$$

Using u_1^* in (39), we find, after some calculation, that

$$\lambda_1(\mathcal{R}(\omega))=\lambda_1^*+\omega\lambda_1^*2^{3/2}/\pi+O(\omega^2). \tag{41}$$

One can now easily verify that (40) and (41) agree up to second order terms.

REFERENCES

[1] H. J. AHN, *On random transverse vibrations of a rotating beam with a tip mass: Method of integral equations*, Quart. J. Mech. Appl. Math., 36 (1983), pp. 97-109.

[2] D. C. BARNES, *Buckling of columns and rearrangements of functions*, Quart. Appl. Math., 41 (1983), pp. 169-180.

[3] ———, *Extremal problems for eigenvalue functionals*, this Journal, 16 (1985), pp. 1284-1294.

[4] W. E. BOYCE, *Random eigenvalue problems*, in Probabilistic Methods in Applied Mathematics, Vol. 1, A. T. Barucha-Reid, ed., Academic Press, New York, 1968, pp. 1-73.

[5] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.

[6] S. P. TIMOSHENKO AND J. N. GERE, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

EIGENVALUES ON A DOMAIN WITH DISCRETE ROTATIONAL SYMMETRY*

BONITA HART DRISCOLL†

Abstract. The spectrum of the Laplace operator on a bounded domain or manifold consists of only isolated eigenvalues of finite multiplicity. In problems with a high degree of symmetry it is necessary that these multiplicities are large. The generic structure of the eigenspaces of the Laplace operator with no symmetry constraints is that the eigenvalues are simple.

We consider the structure of the eigenfunctions of the Laplace operator in a planar domain under deformations that preserve symmetry under a discrete rotation. There are two types of eigenfunctions: symmetric and asymmetric.

The main theorem shows that generically, symmetric eigenfunctions are simple and asymmetric eigenfunctions are of multiplicity two. This is a partial proof of the conjecture of V. I. Arnol'd concerning the codimension of large multiplicity eigenspaces in the space of domains preserving symmetry.

Key words. symmetric eigenfunctions, asymmetric eigenfunctions, generic

AMS (MOS) subject classifications. Primary 58G25; secondary 35P05

Introduction. The spectrum of the Laplace operator on a bounded domain or manifold consists of only isolated eigenvalues of finite multiplicity. If the domain has symmetry, for example, and is invariant under the action of a cyclic group, the fact that the Laplace operator commutes with the group action forces multiplicities in the eigenvalues. The minimal polynomial for the group generator splits the space into subspaces. The subspaces determine different types of eigenfunctions and determine the multiplicities of the eigenvalues.

Let D be the unit disk in R^2 and $\Delta=(\partial/\partial x)^2+(\partial/\partial y)^2$; then what follows is the main result of this paper.

MAIN THEOREM. *Let $\Delta U+\lambda\rho U=0$ with $U=0$ on the boundary and ρ invariant under the Z_p action $\alpha z=e^{2\pi i/p}z$; then the set of $\rho\in C^k$, $k>2$ such that the following*:

(a) *Symmetric eigenspaces (those corresponding to rotation invariant eigenfunctions) are one-dimensional*;

(b) *Asymmetric eigenspaces (those corresponding to irreducible quadratic factors of the minimal polynomial for the group) are two-dimensional*;

(c) *No symmetric eigenvalue equals an asymmetric eigenvalue*;

is residual in C^k, i.e., is a countable intersection of open dense sets in a Banach space.

As a corollary we obtain: For a residual set of surfaces of the conformal type of the disk having Z_3 symmetry, the eigenspaces algebraically have minimal dimension. Note that by "residual" we mean that the measure in the conformal factor ρ is residual.

The main theorem is actually Theorem 2.8, which establishes for $p=2$, 3, 4 that the eigenspaces generically have the lowest possible dimension. We have not included the details of the case for p even corresponding to $f(\alpha z)=-f(z)$, since it is virtually identical to the symmetric case, $f(\alpha z)=f(z)$. For $p\geqq 5$ we have only partial results. In addition to the results in the main theorem, we need to show that eigenspaces corresponding to different conjugate pairs of complex roots generically have different eigenvalues. This poses interesting algebraic questions. The computations reduce to algebraic identities between eigenfunctions. For $p<5$, these identities cannot be

* Received by the editors October 2, 1985; accepted for publication (in revised form) January 13, 1986.

† Department of Mathematical Sciences, Loyola University of Chicago, Chicago, Illinois 60626.

satisfied by orthogonal eigenfunctions, hence assuming that they are leads to a contradiction. However, for $p \geqq 5$, it is possible that there are some rational relations between eigenfunctions, i.e., these identities may be satisfied in some cases. Results in Eskin, Ralston and Trubowitz [9] suggest that problems will arise for certain algebraic manifolds.

The results in this paper originate in a conjecture by Arnol'd [6]. He has conjectured that the map from membranes into bilinear forms is transverse to the various strata of the varieties of bilinear forms with multiple eigenvalues.

From this hypothesis, the results of this paper would follow. One would also be able to determine the codimension of the degeneracies, i.e., the codimension of the submanifolds in the space of membranes which correspond to higher than necessary eigenvalue multiplicities. K. Uhlenbeck has used transversality in the problem on a domain without symmetry [20], [21]. Clearly her technique could be used in the problem with symmetry on the space of symmetric eigenfunctions, but there are inherent difficulties on the space of asymmetric eigenfunctions. We use a perturbation argument similar to that used by Albert [2], [3]. The results are easily extended to a larger class of elliptic operators on an n-dimensional manifold: in particular to the Schrödinger operator. We have chosen to restrict the argument to R^2 since this case seems to contain the basic phenomena. In particular the extension to noncommutative or continuous groups appears to present serious problems.

1. Preliminary results. In what follows Ω will be a bounded simply connected nonempty set in the plane, bounded by a curve of class $C^{k+\beta}$, $k \geqq 3$, and invariant under rotation by $\alpha = 2\pi/p$ where p is an integer. Recall that $C^{k+\beta}$ functions are continuous with continuous derivatives and kth derivative that is Lipschitz with Lipschitz constant β, $0 < \beta < 1$. $D = \{x \mid x \in R^2, \|x\| \leqq 1\}$. The Sobolev space of functions on the unit disk with distributional derivatives through order k that are p integrable is denoted $H_k^p(D)$ and, for $p = 2$, $H_k(D)$. $H_{k,0}^p(D) \subset H_k^p(D)$ is the closure of smooth functions with compact support in the interior of D. We will use generic to mean residual, i.e., a countable intersection of open dense sets.

The first topic we consider is reducing the eigenvalue problem for the Laplace operator on a domain of the kind considered to one on a simpler set, namely, the unit disk, D. To do this we use an equivariant form of the Riemann mapping theorem. Consider a domain $\Omega \subset R^2$ to be a subset of the complex plane by $x = (x, y) \to (x + iy) = z$.

LEMMA 1.1. *For a domain Ω as described, Ω invariant under rotation by α, $0 < \alpha \leqq 2\pi$ and $p\alpha = 2\pi$ for p an integer, there is a conformal map $g: D \to \Omega$ such that $g \in C^{k+\beta}(D)$ and g commutes with rotation by α.*

Proof. Pick a point w_0 on the boundary of Ω and z_0 on the boundary of the unit disk. Choose the unique conformal map g^{-1} which takes Ω to D with

$$g^{-1}(e^{k\alpha i} w_0) = e^{k\alpha i} z_0, \qquad k = 0, 1, 2.$$

Let $h(z) = e^{\alpha i} z$ and $h^{-1}(w) = e^{-\alpha i} w$; then

$$h^{-1} \circ g \circ h(e^{k\alpha i} z_0) = e^{k\alpha i} w_0, \qquad k = 0, 1, 2.$$

Therefore, by uniqueness, since $h^{-1} \circ g \circ h$ maps $D \to \Omega$ and fixes the same points as g, $h^{-1} \circ g \circ h = g$ hence $g(e^{\alpha i} z) = e^{\alpha i} g(z)$.

In the coordinates of the disk used as conformal coordinates for Ω, we have

$$\Delta_\Omega = \rho^{-1} \left[\left(\frac{\partial}{\partial x} \right)^2 + \left(\frac{\partial}{\partial y} \right)^2 \right]$$

where $\rho=|g'(z)|^2$. Because g commutes with the rotation α,

$$\rho(e^{\alpha i}z)=|g'(e^{\alpha i}z)\,e^{\alpha i}|^2=|g'(z)|^2.$$

THEOREM 1.2. *The problem $\Delta_\Omega U+\lambda U=0$, for Ω invariant under rotation by $\alpha=2\pi/p$, is equivalent to $\Delta_D U+\lambda\rho U=0$ for $\rho=|g'(z)|^2$ and ρ invariant under rotation by α.*

Rather than consider only ρ's that arise from the conformal maps described, we consider all $\rho\in C^k(D)$, $k>2$, $\rho(x)\neq 0$ for all x, and ρ invariant under rotation by α. The problem we consider is the problem of the Laplace-Beltrami operator on C^k surfaces of the conformal type of the disk.

THEOREM 1.3. *Let $\rho: D\to R^+$, and a metric on D be given by $g_{ij}=\rho\delta_{ij}$. This puts the set of $\rho>0$, $\rho\in C^k(D)$, $k>2$ in one-to-one correspondence with the diffeomorphism classes of C^k surfaces of the conformal type of the disk. The Laplace-Beltrami operator in this parameterization appears as $\rho^{-1}\Delta$.*

Proof. See [13, pp. 366-367] and [10, pp. 88-101].

The rotational symmetry of the domain Ω under α induces well-known properties on the Hilbert space $H_{1,0}=H$ and the eigenvalues of $\rho^{-1}\Delta$. Let O be the rotation matrix

$$O=\begin{bmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{bmatrix},\qquad \alpha=2\pi/p.$$

For $f\in H$, define $O^*(f(x))=f(O(x))$. Note that $O^{*p}=I$ so we have an irreducible representation of the Z_p action on H,

$$Z_p=\{I, O^*, O^{*2}\cdots O^{*p-1}\}.$$

Clearly O^* is orthogonal on $H_{1,0}(D)$ and $H_0(D)$. The minimal polynomial for O^* is t^p-1 which has as possible factors $t-1$, $t+1$ and $t^2+c(k)t+1$. Let

$$\begin{aligned}&H_1=\{f\colon (O^*-1)f=0\},\\ &H_{-1}=\{f\colon (O^*+1)f=0\},\\ &H_{c(k)}=\{f\colon ((O^*)^2+c(k)O^*+1)f=0,\ k\in\mathcal{I}\}\end{aligned}$$

where $\mathcal{I}=\{k\colon k=1,\cdots,l,\ l=(p-1)/2$ when p is odd or $l=(p-2)/2$ when p is even$\}$. We call elements of H_1 symmetric functions and elements of $H_{c(k)}$ asymmetric functions.

The following are well-known facts commonly used in mathematics and physics.

PROPOSITION 1.4. *The equation $\Delta U+\lambda\rho U=0$ has countably many real eigenvalues that are bounded from below by 0 and have no finite accumulation point. The eigenvectors for $\rho^{-1}\Delta$ form a complete ρ-orthonormal basis for H, and the eigenspaces are all finite dimensional.*

The Hilbert space H is split into invariant subspaces by the irreducible actions of Z_p, namely

$$H=H_1\oplus H_{-1}\oplus\sum_{c(k)\in\mathcal{I}}H_{c(k)}.$$

Since O^* and $\rho^{-1}\Delta$ commute, it follows that there is a ρ-orthonormal basis of eigenvectors for $\rho^{-1}\Delta$ that splits into a basis for H_1, H_{-1} and $H_{c(k)}$.

We will proceed with some technical lemmas. The main result is the reduction of the perturbation problem to a computation on the finite dimensional eigenspace. The technique used relies heavily on Rellich's lemma which says that $H_{1,0}(D)$ is relatively compact in $H_0(D)$ [11, p. 31]. In what follows we use $\int_D f$ to mean $\int_D\int f(x)\,dx$ with respect to Lebesgue measure in R^2. We include the first lemma, although it is well known, because the method of proof is typical of the technique used.

LEMMA 1.5. (Upper semi-continuity of the dimension of the eigenspaces). *Let $\rho(t_i, x) = \rho_i$ for $t_i \in [0, 1]$ where ρ_i converges to $\rho = \rho(0, x)$ in C^0. Let $E_i = \{U \mid U \in H_{1,0}, \Delta U + \lambda_i \rho_i U = 0, \lambda_i$ is the mth eigenvalue$\}$, where m is fixed and $E = \{U \mid U \in H_{1,0}, \Delta U + \lambda \rho U = 0$ and λ is the mth eigenvalue$\}$. If E_i is l-dimensional for all i and $\{\lambda_i\}$ is bounded, then dimension $(E) \geqq l$.*

This is a well-known result. However we need a construction which occurs in the proof. We will construct a sequence of eigenspaces E_{ik}, which converges to a subspace of the eigenspace E in the sense that the orthonormal basis we chose for $E_{i,k}$ converges to a subbasis of E in $H = H_{1,0}(D)$.

Proof. For U in eigenfunction, by Green's formula we have:

$$\|U\|_{H_1}^2 = \int_D (U_x^2 + U_y^2 + U^2) = \int_D [(-\Delta U)U + U^2] = \int (\rho\lambda U^2 + U^2).$$

Hence,

$$\|U_i\|_{H_1}^2 = \int_D (\rho_i\lambda_i + 1) U_i^2.$$

Since $\rho_i \to \rho$ in C^0 and $\{\lambda_i\}$ is bounded, there is a constant K such that

$$\|\rho_i\lambda_i\|_\infty \leqq K.$$

Therefore

$$\|U_i\|_{H_1}^2 \leqq \int_D (K+1)U_i^2 = (K+1)\|U_i\|_{H_0}^2.$$

In addition we have,

$$\Delta U_i + \lambda_i\rho_i U_i = 0 \Rightarrow \Delta U_i = -\lambda_i\rho_i U_i.$$

So by regularity theory [11, p. 68],

$$\|U_i\|_{H_2}^2 \leqq C(\|\lambda_i\rho_i U_i\|_{H_0}^2 + \|U_i\|_{H_1}^2)$$
$$\leqq K'\|U_i\|_{H_0}^2, \qquad K' = C(2K+1).$$

Since $\{\lambda_i\}$ is a bounded sequence in R, there is a convergent subsequence. Call it $\{\lambda_i\}$, and let $\{E_i\}$ be the corresponding sequence of eigenspaces. Choose a ρ_i-orthonormal basis $\{U_{ik}\}$, $1 \leqq k \leqq l$ for E_i. Consider the sequence $\{U_{i1}\}$.

$$\|U_{i1}\|_{H_2}^2 \leqq K'\|U_{i1}\|_{H_0}^2 = K'$$

from the estimate above. By Rellich's lemma, there is a convergent subsequence, call it $\{U_{i1}\}$, such that $\{U_{i1}\} \to U_1$ in $H = H_{1,0}$. Let $\{E_{i1}\}$ be the subsequence of $\{E_i\}$ associated with the subsequence $\{U_{i1}\}$; i.e., U_{i1} is the first basis vector of E_{i1}.

We proceed by induction. Let $r \leqq l$ and assume we have sequence $\{U_{ik}\} \to U_k$ in H, $1 \leqq k \leqq r-1$ and a sequence of eigenspaces $\{E_{ir-1}\}$ where

$$\{U_{ik} \mid 1 \leqq k \leqq r-1\} \subset E_{ir-1} \quad \text{for all } i.$$

Consider the sequence $\{U_{ir}\}$ where $U_{ir} \in E_{ir-1}$. Since each E_{ir-1} is l-dimensional and $r \leqq l$, this sequence is nonempty.

From the earlier estimate,

$$\|U_{ir}\|_{H_2}^2 \leqq K'\|U_{ir}\|_{H_0}^2 = K'.$$

So by Rellich's Lemma, there is a convergent subsequence $U_{ir} \to U_r$ in H. Let $\{E_{ir}\}$ be the subsequence of $\{E_{ir-1}\}$ associated with the subsequence U_{ir}, i.e., $U_{ir} \in E_{ir}$. Choose subsequences of $\{U_{ik}\}$, $1 \leqq k \leqq r-1$ by requiring that elements of $\{U_{ik}\}$ $1 \leqq k \leqq r-1$ be elements of E_{ir}. Call the subsequences $\{U_{ik}\}$ $1 \leqq k \leqq r-1$. We now have that $\{U_{ik}\} \to U_k$ $1 \leqq k \leqq r$ and $U_{ik} \in E_{i,r}$ to complete the induction.

We must show that $\{U_k \mid 1 \leqq k \leqq l\} \subset E$, and is orthogonal in E.

(i) Show $\Delta U_k + \lambda \rho U_k = 0$, i.e., $U_k \in E$. Recall that $\{U_{ik}\} \to U_k$ in H and $\rho_i \to \rho$ uniformly in x, where $\rho \in C^k(D)$, $k \geqq 2$ and $\lambda_i \to \lambda$. By a regularity theorem, a weak solution in H is a classical solution [11, p. 67].

$$\Delta U_{ik} + \lambda_i \rho_i U_{ik} = 0,$$

so

$$\int_D (\Delta U_{ik} + \lambda_i \rho_i U_{ik}) \phi = 0$$

and

$$\int_D \Delta U_{ik} \phi = -\int_D \lambda_i \rho_i U_{ik} \phi \quad \text{for all } \phi \in C_0^\infty(D).$$

Clearly

$$\int_D \lambda_i \rho_i U_{ik} \phi \to \int_D \lambda \rho U_k \phi$$

and

$$\int_D U_{ik} \Delta \phi \to \int_D U_k \Delta \phi;$$

then

$$\int_D (\Delta \phi + \lambda \rho \phi) U_k = 0 \quad \text{for all } \phi \in C^\infty(D)$$

and U_k is a weak solution and hence a classical solution of $\Delta U_k + \lambda \rho U_k = 0$; i.e., $U_k \in E$.

(ii) Show $(U_k \mid U_j)_\rho = 0$ for $k \neq j$.

$$U_{ik} \to U_k, \qquad U_{ij} \to U_j,$$
$$(U_{ik} \mid U_{ij})_{\rho_i} = 0 \quad \text{for all } i \text{ and } k \neq j.$$

Therefore if $k \neq j$,

$$\lim_{i \to \infty} (U_{ik} \mid U_{ij})_{\rho_i} = (U_k \mid U_j)_\rho = 0.$$

By the same reasoning

$$\lim_{i \to \infty} (U_{ik} \mid U_{ik})_{\rho_i} = (U_k \mid U_k)_\rho = 1.$$

Hence $\dim E \geqq l$ since E contains l orthogonal vectors.

THEOREM 1.6. *Let* $\rho_i = \rho(t_i, x)$, $\lambda_i = \lambda(t_i)$ *where* $\lim_{t_i \to 0} \rho_i = \rho(0, x)$. *Let* $\partial \rho / \partial t(0, x) \in C^0(D)$ *and* E_i (*the mth eigenspace*) *be l-dimensional for all i and l fixed. Then* $\dim E \geqq l$ *and equality implies that there exists a number* $d\lambda/dt$ *such that for all* $U, V \in E_{\lambda\rho}$

$$\lambda \int_D \frac{\partial \rho}{\partial t}(0, x) UV = -\frac{d\lambda}{dt} \int_D \rho VU.$$

Proof. Let $U_1, \cdots, U_l$ be the orthonormal basis for $E_{\lambda,\rho}$ constructed in Lemma 1.5. There are sequences $\{U_{ik}\} \to^H U_k$, $k=1, \cdots, l$ where $U_{ik} \in E_i$ and $H = H_{1,0}(D)$. Let $V_i = \sum_{k=1}^l a_k U_{ik}$. Then $V_i \in E_i$ and $V_i \to_H \sum_{k=1}^l a_k U_k = V \in E$;

$$\Delta U + \lambda \rho U = 0$$

so

$$\Delta U + \lambda_i \rho_i U + \lambda_i(\rho - \rho_i)U + \rho(\lambda - \lambda_i)U = 0.$$

The image is ρ_i orthogonal to the eigenspace E_i so $(\Delta U + \lambda_i \rho_i U \mid U_i) = 0$;

$$(\Delta U + \lambda_i \rho_i U + \lambda_i(\rho - \rho_i)U + \rho(\lambda - \lambda_i)U \mid U_i) = 0$$

so

$$\lambda_i\left(\left(\frac{\rho - \rho_i}{t_i}\right)U \mid U_i\right) + \left(\frac{\lambda - \lambda_i}{t_i}\right)(\rho U \mid U_i) = 0.$$

For $V_i \in E_i$,

$$V_i = \sum_{k=1}^{l} a_k U_{ik},$$

$$\lambda_i\left(\left(\frac{\rho - \rho_i}{t_i}\right)U \mid V_i\right) + \left(\frac{\lambda - \lambda_i}{t_i}\right)(\rho U \mid V_i) = 0. \tag{1.6}$$

Look at

$$\lim_{i\to\infty} \lambda_i\left(\left(\frac{\rho - \rho_i}{t_i}\right)U \mid V_i\right) = \lim_{i=\infty} \sum_{k=1}^{l} a_k \lambda_i\left(\left(\frac{\rho - \rho_i}{t_i}\right)U \mid U_{ik}\right).$$

Since $\lambda_i \to \lambda$, $(\rho - \rho_i)/t_i \to d\rho/dt(0)$ uniformly in x, and $U_{ik} \to^H U_k$;

$$\lim_{i\to\infty} \sum_{k=1}^{l} a_k \lambda_i\left(\left(\frac{\rho - \rho_i}{t_i}\right)U \mid U_{ik}\right) = \sum_{k=1}^{l} a_k \lambda\left(\frac{\partial \rho}{\partial t}(0)U \mid U_k\right)$$

$$= \lambda\left(\frac{\partial \rho}{\partial t}(0)U \mid V\right).$$

Also

$$\lim_{i\to\infty} (\rho U \mid V_i) = (\rho U \mid V)$$

since $V_i \to^H V$.

From (1.6), with $U = V \neq 0$, $\lim_{i\to\infty} (\lambda - \lambda_i)/t_i$ exists, since

$$\lim_{i\to\infty} \frac{\lambda - \lambda_i}{t_i} = \frac{-\lambda(\partial\rho/\partial t(0)U \mid U)}{(U \mid U)_\rho} = \frac{d\lambda}{dt}.$$

We have

$$\lambda\left(\frac{\partial \rho}{\partial t}(0)U \mid V\right) = -\frac{d\lambda}{dt}(\rho U \mid V)$$

for all $U, V \in E$.

COROLLARY 1.7. *Let $\rho(t,\ ) \in C^k(D)$, $k \geqq 2$ and assume $\partial\rho/\partial t(t,\ ) \in C^k(D)$. Let $q = \partial\rho/\partial t(0,\ ) \in C^k(D)$. If there exists a U, $V \in E$ and an eigenspace for $\Delta U + \lambda \rho U = 0$ of dimension l, such that $(U \mid V)_\rho = 0$ and $(U \mid V)_q \neq 0$, then there exists an $\varepsilon > 0$ such that* dim $E_t < l$ *for $0 < t < \varepsilon$, where E_t is the eigenspace corresponding to $\Delta U + \lambda \rho(t,\ )U = 0$, $\lambda_i \to \lambda$.*

Proof. Suppose not. Apply the theorem to $E_{t_i} = E_i$ for $t_i \to 0$ and E_i of dimension l. Then $\lambda(U|V)_q = -d\lambda/dt(U|V)_\rho = 0$, which is a contradiction.

2. Main theorem. In this section we prove the main theorem. The technique used is similar to that of Albert [2], [3]. We define sequences of nested sets. Let

$$P = \{\rho \in C^k(D) \mid O^*\rho = \rho \text{ and } \rho > 0\},$$

$$S_n = \{\rho \in P \mid \text{the first } n \text{ symmetric eigenvalues of } \rho^{-1}\Delta \text{ are of multiplicity } 1\},$$

$$A_n(k) = \{\rho \in P \mid \text{the first } 2n \text{ asymmetric eigenvalues corresponding to eigenfunctions which satisfy } O^{*2} + C(k)O^* + 1 = 0 \text{ have multiplicity } 2\},$$

$$T_n = \{\rho = P \mid \rho \in S_n \cap A_n(k) \text{ and none of the first } n \text{ symmetric eigenvalues is equal to any of the first } 2n \text{ asymmetric eigenvalues}\},$$

$$S_0 = A_0(k) = T_0 = P.$$

Then

$$S_0 \supset S_1 \supset S_2 \supset \cdots, \qquad S = \bigcap_{n=1}^{\infty} S_n,$$

$$A_0(k) \supset A_1(k) \supset A_2(k) \supset \cdots, \qquad A(k) = \bigcap_{n=1}^{\infty} A_n(k),$$

$$T_0 \supset T_1 \supset T_2 \supset \cdots, \qquad T = \bigcap_{n=1}^{\infty} T_n.$$

We show that S_n is open in P and $A_n(k)$ is open in P, and T_n is open in P. The proof of openness is an easy consequence of Lemma 1.5. The proof of density is more complicated and constitutes the rest of the section. We apply the criterion we have developed in Corollary 1.7 to show that S_n is dense in S_{n-1}, $A_n(k)$ is dense in $A_{n-1}(k)$ and T_n is dense in T_{n-1}. The proof consists in showing that the multiplicity of the eigenspaces can be reduced by at least one (two in the asymmetric case), and of using a series of perturbations as required. The proof for eigenvalues corresponding to eigenfunctions satisfying $O^*f = -f$ is identical to the symmetric case, so for simplicity it is not included.

We will first prove openness. In what follows, we define simple eigenvalue to mean: for symmetric eigenvalues the multiplicity is 1 and for asymmetric eigenvalues the multiplicity is 2.

THEOREM 2.1. *S_n, $A_n(k)$ and T_n are open in P in the C^k topology.*

Proof. We shall first consider T_n.

For $\rho \in T_n$, we must show that there is an ε-neighborhood of ρ such that for all q with $\|\rho - q\|_k < \varepsilon$, the first n eigenspaces of q are simple. We proceed by contradiction.

Assume that for every $t > 0$ there is a q_t such that $\|\rho - q_t\|_k < t$ and the dimension of at least one of the first n eigenspaces of q_t is not simple. Since n is finite, we may assume that the mth eigenspace is not simple for each t. Let λ be the mth eigenvalue for ρ and λ_t the mth eigenvalue for q_t. There is a sequence of $q_t \to \rho$, with dimension of the mth eigenspace $\geqq 2$ if λ is a symmetric eigenvalue or $\geqq 3$ if λ is a asymmetric eigenvalue. Since $q_t \to \rho$ and D is compact, it is easily shown that $\{\lambda_t\}$ is bounded. Hence by the proof of Lemma 1.5, we have that the dimension of the mth eigenspace of ρ is not simple, which is a contradiction. Restriction to H_1 and H_{c_k} gives the lemma for S_n and $A_n(k)$.

We will now prove the perturbation lemmas, which easily imply density.

LEMMA 2.2. *Let E_0, the nth eigenspace for ρ_0, be l-dimensional where $l>1$ and $E_0 \subset H_1$. Then there is a ρ, C^k close to ρ_0, such that the nth eigenspace of ρ has dimension $<l$.*

Proof. The proof is by contradiction. If false by Theorem 1.6 and Corollary 1.7,

$$\lambda \int_D \frac{\partial \rho}{\partial t}(0,\ )UV + \frac{d\lambda}{dt}\int_D \rho UV = 0$$

for all U, $V \in E$ and all $\rho \in P$. Then for U, V such that $(U \mid V)_\rho = 0$ we have $\int_D \rho UV = 0$ and so

$$\lambda \int_D \frac{\partial \rho}{\partial t}(0,\ )UV = 0, \qquad \lambda \neq 0.$$

Choose $\rho(t,\ ) = \rho_0 + t\cdot S$ where $S \in P$. Then $\partial\rho/\partial t(0,\ ) = S$, and S is invariant under O^*. Since U and V are symmetric eigenfunctions, U and V are invariant, under O^*, so by change of variable,

$$\int_D S\cdot U\cdot V = p\int_{D/p} SUV = 0.$$

But S is arbitrary on D/p, so by the Fundamental Lemma of the Calculus of Variations, $UV=0$. Notice that $L(U) = \Delta U + \lambda\rho_0 U$ has the weak continuation property [12]. This means that if U is a solution of $L(U)=0$ and U vanishes on an open set, U is identically zero. Since U and V are C^2 and $UV=0$, either U or V must vanish on an open set which, as we just showed, implies U or V is identically zero. This is a contradiction.

LEMMA 2.3. *Let E_0, the nth eigenspace for ρ_0, be 2l-dimensional where $l>1$ and $E_0 \subset H_{c(k)}$. Then there is a ρ, C^k close to ρ_0, such that the nth eigenspace of ρ has dimension $<2l$.*

Proof. The proof is by contradiction. As in Lemma 2.2, if the theorem is false, then for $(u \mid v)_\rho = 0$ and $\rho(t,\ ) = \rho_0 + tS$, $S \in P$, we have $\lambda \int_D Suv = 0$, $\lambda \neq 0$.

By change of variable, recalling S is invariant under O^*, we have

$$\int_D Suv = \sum_{i=0}^{p-1}\int_{D/p} SO^{*i}(uv) = 0.$$

But S is arbitrary on D/p, so

$$\sum_{i=0}^{p-1} O^{*i}(uv) = 0. \tag{2.3}$$

Let E_u = space spanned by $\{u, O^*u\}$ where $u \in E \subset H_{c(k)}$. Notice that E_u is two-dimensional over the reals since $O^*u \neq ku$ for all $k \in R$. For $u \in E$, choose $v \in E$ so that $(u \mid v)_\rho = 0$ and $(O^*u \mid v)_\rho = 0$. This is possible since $\dim E \geqq 4$. We notice that $(u \mid O^*v)_\rho = 0$. This is true since $(O^*)^T = (O^*)^{-1}$, $(U + O^*u \mid v)_\rho = 0$, and $-O^{*-1}u = O^*u + u$. Therefore $-(O^{*-1}u \mid v)_\rho = 0 = (u \mid O^*v)$. We will show $uv=0$ for u, v as above by looking at the problem in the complex plane. Let $(x,y) \to z \in C$ by $(x,y) \to (x+iy)$, and let $v \notin E_u$ as above.

$$u, v \in H_{c(k)} \Rightarrow [(O^*)^2 + c(k)O^* + 1]u = 0 \quad \text{and}$$

$$[(O^*)^2 + c(k)O^* + 1]v = 0.$$

Then

$$(O^*-\lambda)(O^*-\bar{\lambda})u=0, \quad \lambda=e^{2\pi ki/p}, \quad k\neq p \text{ and } k\neq p/2 \text{ if } p \text{ is even}$$

and

$$(O^*-\lambda)(O^*-\bar{\lambda})v=0.$$

Define $U(z)=O^*u(z)-\bar{\lambda}u(z)$, and $V(z)=O^*v(z)-\bar{\lambda}v(z)$. Then $(O^*-\lambda)V(z)=(O^*-\lambda)(O^*-\bar{\lambda})v(z)=0$, so $(O^*-\lambda)V(z)=0$ or $O^*V(z)=\lambda V(z)$. Also, $O^*U(z)=\lambda U(z)$, so $O^*(U\bar{V})=\lambda\bar{\lambda}U\bar{V}=U\bar{V}$, and hence

$$(O^*)^k(U\bar{V})=O^*(U\bar{V})=U\bar{V}, \qquad k=1,\cdots,p.$$

Therefore

$$\begin{aligned} U\bar{V}=\frac{1}{p}\sum_{k=0}^{p-1}(O^*)^k(U\bar{V})&=\frac{1}{p}\sum_{k=0}^{p-1}[O^{*^{k+1}}u(z)-\bar{\lambda}O^{*^k}u(z)][(O^*)^{k+1}v(z)-\lambda O^{*^k}v(z)]\\ &=\frac{1}{p}\sum_{k=0}^{p-1}[O^{*^{k+1}}u(z)O^{*^{k+1}}v(z)+O^{*^k}u(z)O^{*^k}v(z)\\ &\qquad -\bar{\lambda}O^{*^k}u(z)O^{*^{k+1}}v(z)-\lambda O^{*^k}v(z)O^{*^{k+1}}u(z)]. \end{aligned}$$

By (2.3)

$$\sum_{k=0}^{p-1}O^{*^k}(uv)=0.$$

Since O^* is linear,

$$O^*\sum_{k=0}^{p-1}O^{*^k}(uv)=\sum_{k=0}^{p-1}O^{*^{k+1}}(uv)=0$$

if $(u|v)_\rho=0$. Recall that $(u|O^*v)_\rho=0$ and $(O^*u|v)_\rho=0$. Therefore

$$\sum_{k=0}^{p-1}O^{*^{k+1}}u(z)O^{*^{k+1}}v(z)=\sum_{k=0}^{p-1}O^{*^{k+1}}[u(z)v(z)]=0,$$

$$\sum_{k=0}^{p-1}O^{*^k}u(z)O^{*^k}v(z)=\sum_{k=0}^{p-1}O^{*^k}(u(z)v(z))=0$$

and

$$\sum_{k=0}^{p-1}O^{*^k}u(z)O^{*^{k+1}}v(z)=\sum_{k=0}^{p-1}O^{*^k}(u(z)O^*v(z))=0,$$

by (2.3) with O^*v replacing v. Similarly,

$$\sum_{k=0}^{p-1}O^{*^k}v(z)O^{*^{k+1}}u(z)=0.$$

Therefore $U\bar{V}=0$ so either U or $\bar{V}$ is zero on an open set. Hence Im $U=\sin(2\pi k/p)u$ and Im $V=\sin(2\pi k/p)v$ are zero on an open set, so by the weak continuation property u and v are identically zero.

LEMMA 2.4. *Let E_0, the nth eigenspace for ρ_0, be l-dimensional $l\geqq 3$, where $E_0\cap H_{c(k)}\neq\emptyset$ and $E_0\cap H_1\neq\emptyset$. Then there is a ρ, arbitrarily C^k close to ρ_0, such that the nth eigenspace of ρ has dimension less than l.*

Proof. As in Lemmas 2.2 and 2.3, we will use proof by contradiction. We will assume the theorem is false for $U \in E_0 \cap H_{c(k)}$ and $V \in E_0 \cap H_1$. Let $(U|U)_\rho = 1$ and $(V|V)_\rho = 1$. Notice that $H_1 \perp H_{c(k)}$, for all ρ, so we will consider $U+V$, $U-V$.

$$(U+V|U-V)_\rho = (U|U)_\rho - (V|V)_\rho = 0.$$

As before, assume the lemma is false, and by Theorem 1.6 and Corollary 1.7, we have for all ρ and for $(U+V|U-V)_\rho = 0$ where $U \in H_{c(k)}$ and $V \in H_1$,

$$\int_D \frac{\partial \rho}{\partial t}(0, \ )(U^2 - V^2) = 0.$$

For $\rho(t, \ ) = \rho + ts$, $s \in P$. Since $O^*s = s$ and $O^*V = V$, by change of variable,

$$0 = \int_D s(U^2 - V^2) = \int_{D/p} s \sum_{k=0}^{p-1} O^{*k}(U^2 - V^2).$$

So since s is arbitrary,

$$\sum_{k=0}^{p-1} O^{*k}U^2 - pV^2 = 0.$$

Therefore,

$$V^2 = \frac{1}{p} \sum_{k=0}^{p-1} O^{*k}U^2.$$

Since V is an eigenfunction for $\rho^{-1}\Delta$, there is a connected component G of the set in D on which V is positive bounded by nodal curves of V [8, p. 451]. Consider $\Delta V + \lambda \rho V = 0$ on G, $V = 0$ on ∂G. V is an eigenfunction for this domain G, and since $V > 0$ on G, λ must be the first eigenvalue [8, p. 451]. The first eigenvalue is simple. Notice that $V = 0$ on the boundary of G, gives $V^2 = 0$, so

$$0 = \frac{1}{p} \sum_{k=0}^{p-1} O^{*k}U^2 \quad \text{on } \partial G.$$

Each summand is positive, so $U^2 = 0$ on ∂G which says $U = 0$ on ∂G. Therefore U is an eigenfunction for $\Delta V + \lambda \rho V = 0$ on G and since λ is simple this says $U = V$ which is a contradiction since $(U|V)_\rho = 0$.

We shall now prove that S_n is dense in S_{n-1}, $A_n(k)$ is dense in $A_{n-1}(k)$ and T_n is dense in T_{n-1}. We have shown in Lemmas 2.2-2.4 that if the multiplicities are not minimal we can reduce the multiplicity of the eigenspaces at least one in the symmetric case, at least two in the asymmetric case and at least one if asymmetric and symmetric coincide. We shall show the density, by using a sequence of perturbations to lower the multiplicity of the nth eigenspace one step at a time. By the openness one can make small enough perturbations so none of the multiplicities of any other eigenspace is increased.

THEOREM 2.5. *S_n is dense in S_{n-1} in the C^k topology.*

Proof. Let $\rho \in S_{n-1}$ and let E_0, the nth eigenspace of ρ, have dimension $l > 1$. Since S_{n-1} is open in P in the C_k topology (Theorem 2.1), There is a δ so that if $\|\rho - q\|_k < \delta$ then $q \in S_{n-1}$. By Lemma 2.2, there is an ε_1, $\varepsilon_1 < \delta$ and a ρ_1 with $\|\rho - \rho_1\|_k < \varepsilon_1/l$, where the nth eigenspace E_1 of ρ_1 has dimension less than or equal to $l-1$. Notice that $\rho_1 \in S_{n-1}$ by construction. We proceed by induction. In what follows $l > 2$. For $r \leqq l$ assume that we have ρ_{r-1} with $\|\rho_{r-2} - \rho_{r-1}\|_k < \varepsilon_{r-1}/l$ where $\rho_{r-1} \in S_{n-1}$ and the dimension of the nth eigenspace E_{r-1} of ρ_{r-1} is less than or equal to $l-(r-1)$.

By Lemma 2.2, there is an ε_r, $\varepsilon_r < \delta$ and a ρ_r so that $\|\rho_{r-1} - \rho_r\|_k < \varepsilon_r / l$ and the dimension of the nth eigenspace E_r of ρ_r is less than or equal to $[l-(r-1)]-1 = l-r$.

We must show $\|\rho - \rho_r\|_k < \delta$. Let $\rho = \rho_0$ and look at

$$\|\rho - \rho_r\|_k = \left\| \sum_{j=0}^{r-1} (\rho_j - \rho_{j+1}) \right\|_k \leqq \sum_{j=0}^{r-1} \|\rho_j - \rho_{j+1}\|_k,$$

$$\|\rho - \rho_r\|_k \leqq \sum_{j=0}^{r-1} \frac{1}{l} \varepsilon_{j+1} \leqq \sum_{j=0}^{r-1} \frac{1}{l} \delta \leqq \delta,$$

since $r \leqq l$.

Therefore $\rho_r \in S_{n-1}$ and the dimension of E_r is less than $l-r$. Clearly this process will terminate only if $\dim(E_j) = 1$ and that must happen before $j = l$.

THEOREM 2.6. *$A_n(k)$ is dense in $A_{n-1}(k)$ in the C^k topology.*

Proof. Use Lemma 2.3 and the same construction as Theorem 2.5.

THEOREM 2.7. *T_n is dense in T_{n-1}.*

Proof. Recall that

$$T_{n-1} = \{\rho \mid \rho \in S_{n-1} \cap A_{n-1}(k) \text{ such that the first } (n-1) \text{ symmetric eigenvalues are all different from the first } 2(n-1) \text{ asymmetric eigenvalues}\}.$$

For $\rho \in T_{n-1}$, the nth eigenspace E_n is at most 3-dimensional since $\rho \in S_{n-1} \cap A_{n-1}(k)$. Since T_{n-1} is open in P, by Theorem 2.1 there is a neighborhood of ρ in C^k so that for q in that neighborhood, $q \in T_{n-1}$. Using Lemma 2.4, pick ρ_ε so that $\rho_\varepsilon \in T_{n-1}$ and the dimension of the nth eigenspace of $\rho_\varepsilon < 3$. This completes the proof, since for ρ_ε the nth asymmetric eigenvalue must be different from the nth symmetric one.

THEOREM 2.8. *T_n is dense and open in P and $T = \bigcap_{n=1}^{\infty} T_n$ is residual in P.*

Proof. T_n is open in P from Theorem 2.1. T_n is dense in P from Lemma 2.4.

3. Some examples.

Example 1. The simplest example of a domain with Z_3 symmetry is an equilateral triangle. Let D be an equilateral triangle of side 1, and consider the problem $\Delta u + \lambda u = 0$ in D with $u = 0$ on the boundary of D. The eigenvalues of Δ on

$$D = \{(x, y) : 0 < y < x\sqrt{3}, y < \sqrt{3}(1-x)\} \tag{3.1}$$

are the numbers

$$\lambda_{mn} = \left(\frac{16\pi^2}{27}\right)(m^2 + n^2 - mn), \qquad m, n = 0, \pm 1, \tag{3.2}$$

with the following conditions:

$$\begin{aligned} &\text{(A)} \quad m+n \text{ is a multiple of } 3,\\ &\text{(B)} \quad m \neq 2n,\\ &\text{(C)} \quad n \neq 2m. \end{aligned} \tag{3.3}$$

The multiplicity of λ_{mn} is $(1/6) \times$ the number of times it appears in the lattice (3.2). The eigenfunctions are of the form

$$f(x, y) = \sum_{(m,n)} \pm \exp\left(\frac{2\pi i}{3}\right)\left(nx + \frac{(2n-m)y}{\sqrt{3}}\right). \tag{3.4}$$

In this sum (m, n) range over $\mathscr{L} \subseteq Z^2$, where $|\mathscr{L}| = 6$ and ± is determined by the following

$$\begin{tikzcd}
 & (m, n) \arrow[dr] & \\
(n-m, n) \arrow[ur] & & (m, m-n) \\
(n-m, -m) \arrow[u] & & (-n, m-n) \arrow[dl] \\
 & (-n, -m) \arrow[ul] &
\end{tikzcd} \tag{3.5}$$

Each transition induces a change of sign in the (m, n) entry of (3.4). Each pair (m, n) that appear in (3.5) must satisfy (B) and (C).

These results were obtained by M. Pinsky [15]. He also shows that a given eigenvalue is either symmetric or asymmetric, i.e., either all of its eigenfunctions are symmetric or all of them are asymmetric. He includes a formula for the multiplicity of the eigenvalues which shows that the dimensions of the eigenspaces become arbitrarily large for both symmetric and asymmetric eigenvalues. Hence this example does not exhibit generic behavior.

Example 2. Domains with Z_5 symmetry illustrate what is necessary to complete the proof for Z_p. The minimal polynomial for Z_5 is

$$\begin{aligned} t^5 - 1 &= (t-1)\left(t^2 - 2t\left(\frac{\sqrt{5}-1}{4}\right) + 1\right)\left(t^2 + 2t\left(\frac{\sqrt{5}+1}{4}\right) + 1\right) \\ &= (t-1)P_1(t)P_2(t). \end{aligned}$$

Rotation through $\alpha = 2\pi/5$ gives an irreducible representation for the group Z_5.

Since $\rho^{-1}\Delta$ commutes with rotation, the 3 invariant subspaces for Z_5 are invariant subspaces for $\rho^{-1}\Delta$. Hence the eigenfunction for $\rho^{-1}\Delta$ must satisfy one of these three equations:

(i) $(O^* - 1)f = 0,$

(ii) $P_1(O^*)f = 0,$

(iii) $P_2(O^*)f = 0.$

It remains to be shown that if an eigenvalue has some eigenfunctions satisfying (ii) and others satisfying (iii) (hence the eigenspace has dimension $\geqq 4$), one can perturb ρ so as to lower the dimension of the eigenspaces. This is the analogue of Lemma 2.4. From this it would follow that generically a given eigenvalue has eigenfunctions that satisfy exactly 1 of the equations (i), (ii) or (iii). This would complete the proof that the eigenspaces for a domain with Z_5 symmetry (in fact Z_p symmetry) have algebraically minimal dimension.

Example 3. Consider the Dirichlet problem for the Laplace operator on the unit disk in R^2, that is, $\Delta U + \lambda U = 0$ where $U(1, \theta) = 0$. It can be shown that the eigenvalues are squares of the zeros of the Bessel functions J_n. Let $k_{n,m}$ be the mth zero of the nth Bessel function J_n. Then

$$J_n(k_{n,m}r) = \frac{(k_{n,m}r)^n}{2^n n!}\left\{1 - \frac{(k_{n,m}r)^2}{2(2n+2)} + \frac{(k_{n,m}r)^4}{2 \cdot 4(2n+2)(2n+4)} \cdots\right\}.$$

The eigenfunctions are of the form $J_n(k_{n,m}r)(\alpha \cos n\theta + \beta \cos n\theta)$ where α and β are arbitrary [8, pp. 302-305]. For n and k positive integers, the functions $J_n(z)$ and $J_{n+k}(z)$ have no common zeros other than the origin [22, pp. 484-485]. Therefore except for

$n=0$, all of the eigenvalues have multiplicity 2, since $J_n(k_{n,m}r)\cos n\theta$ and $J_n(k_{n,m}r)\sin n\theta$ are independent eigenfunctions.

The symmetric eigenfunctions occur for $n\equiv 0 \pmod 3$ and the asymmetric eigenfunctions for $n\equiv S \pmod 3$, $S=1$ or 2. It is clear that approximately $\frac{1}{3}$ of the eigenfunctions are symmetric. Notice that a given eigenvalue is either symmetric or asymmetric.

We do not see generic behavior in this example for a Z_3 action since the symmetric eigenvalues have multiplicity two. However, the disk is invariant under S^1 and for this group the example is generic.

REFERENCES

[1] T. Aubin, *Nonlinear Analysis on Manifolds, Monge-Ampere Equations*, Springer-Verlag, Berlin-New York-Heidelberg, 1982.

[2] J. H. Albert, *Genericity of simple eigenvalues for elliptic PDE's*, Proc. Amer. Math. Soc., 48 (1975), pp. 413-418.

[3] ———, *Nodal and critical sets for eigenfunctions of elliptic operators*, Proc. Sympos. Pure Math., Vol. 23, Amer. Math. Soc., Providence, R.I., 1973, pp. 71-78.

[4] Lars V. Ahlfors, *Complex Analysis*, second edition, McGraw-Hill, New York, 1966.

[5] N. Aronszajn, *A unique continuation theorem for solution of elliptic partial differential equations or inequalities of the second order*, J. Math. Pures Appl., 36 (1957), pp. 235-247.

[6] V. I. Arnol'd, *Modes and quasimodes*, Funct. Anal. Appl., 6 (1972), pp. 94-101 (translation).

[7] David D. Bleecker and Leslie C. Wilson, *Splitting the spectrum of a Riemannian manifold*, this Journal, 11 (1980), pp. 813-818.

[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Wiley Interscience, New York, 1953.

[9] G. Eskin, J. Ralston and E. Trubowitz, *On Isospectral periodic potentials in R^n*, Comm. Pure Appl. Math., 37 (1984), pp. 715-753.

[10] Luther P. Eisenhart, *A Treatise on the Differential Geometry of Curves and Surfaces*, Ginn and Company, New York, 1909.

[11] Avner Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, Inc., Chicago, 1969.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, second edition, Springer-Verlag, Berlin-New York-Heidelberg, 1983.

[13] Charles B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Die Grundlehren der Mathematischen Wissenschaften, Vol. 130, Springer-Verlag, New York, 1966.

[14] S. Ozawa, *The eigenvalues of the Laplacian and perturbation of boundary condition*, Proc. Japan Acad. Ser. A Math. Sci., 55 (1979), pp. 121-124.

[15] M. Pinsky, *The eigenvalues of an equilateral triangle*, this Journal, 11 (1980), pp. 819-827.

[16] ———, *Completeness of the eigenfunctions of the equilateral triangle*, this Journal, 16 (1985), pp. 848-851.

[17] Michael Reed and Barry Simon, *Methods of Modern Mathematical Physics*, I: *Functional Analysis*, Academic Press, New York, 1972.

[18] M. Tanikawa, *The spectrum of the Laplacian and smooth deformation of the Riemannian metric*, Proc. Japan Acad. Ser. A Math. Sci., 55 (1979), pp. 125-127.

[19] ———, *The spectrum of the Laplacian of Z_3-invariant domains*, Proc. Japan Acad. Ser. A Math. Sci., 57 (1981), pp. 13-18.

[20] K. Uhlenbeck, *Eigenfunctions of Laplace operators*, Bull. Amer. Math. Soc., 78 (1972), pp. 1073-1076.

[21] ———, *Generic properties of eigenfunctions*, Amer. J. Math., 98 (1976), pp. 1059-1078.

[22] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, second edition, Cambridge University Press, Cambridge, 1958.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

THE EXACT NUMBER OF SOLUTIONS OF FINITE DIFFERENCE APPROXIMATIONS OF SOME NONLINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS*

HENNING WIEBERS†

Abstract. We consider finite dimensional nonlinear eigenvalue problems of the type $Az=\lambda F(z)$ where A is a matrix, $(F(z))_j=f(z_j), j=1,\cdots,n$. Such systems arise as discretizations of a corresponding boundary value problem. The number of solutions of both equations at a fixed $\lambda\in\mathbb{R}_+$ may differ significantly, i.e., spurious solutions occur. We give an upper bound on the number of solutions at a fixed λ in case of the nonlinearities $f(z)=\exp(z), f(z)=\sin z$, for polynomial and related nonlinearities. To this end we extend the equation into $\mathbb{C}^n$ and use the Brouwer degree for complex analytic mappings. A homotopy argument yields that, roughly speaking, the number of solutions of the finite system is limited by the number of the solutions of the uncoupled system $z=\lambda F(z)$.

Key words. discretizations, nonlinear eigenvalue problems, spurious solutions

AMS(MOS) subject classifications. 32A10, 34B15, 35J25, 65L10, 65N10

0. Introduction. We consider finite dimensional nonlinear eigenvalue problems of the type

$$(0.1)_\lambda \qquad Az=\lambda F(z),$$

where $\lambda\in\mathbb{R}_+$, $A=(a_{jk})$ is an (n,n) matrix which satisfies

$$a_{jk}\in\mathbb{R},\quad a_{jj}>0,\quad a_{jk}\leqq 0\quad \text{if } j\neq k,$$

$$(0.2)\qquad a_{jj}\geqq\sum_{\substack{k=1\\k\neq j}}^{n}|a_{jk}|\quad\text{with strict inequality for at least one } j\in\{1,\cdots,n\},$$

A is irreducible.

$F:\mathbb{R}^n\to\mathbb{R}^n$ is assumed to be diagonally nonlinear, i.e.,

$$(0.3)\qquad F(z)=(f(z_1),\cdots,f(z_n))^t\quad\text{with } f:\mathbb{R}\to\mathbb{R}.$$

The properties typically arise in the discretization of nonlinear boundary value problems (bvp). Various authors [3], [4], [12] have observed that the solution sets of $(0.1)_\lambda$ and the bvp may differ significantly, i.e., the number of the solution branches of $(0.1)_\lambda$ increases with the number n of meshpoints and $(0.1)_\lambda$ admits extra (spurious) solutions which do not converge to solutions of the bvp as $n\to\infty$.

In this paper we give an upper bound of the number of solutions of $(0.1)_\lambda$ for the examples $f(z)=\exp(z), f(z)=\sin z$, for polynomial and related nonlinearities, which have also been studied in [3], [4], [9], [12].

Let us consider the bvp

$$(0.4)_\lambda\qquad -u''=\lambda f(u)\quad\text{in }(0,1),\qquad u(0)=u(1)=0$$

with $f(z)=\exp(z)$. It is well known [10] that the solution set of $(0.4)_\lambda$ is given by the solid branch in Fig. 1, whereas $(0.1)_\lambda$ with a simple discretization matrix A admits extra solutions (the dotted branches). It is shown in [3] that $(0.1)_\lambda$ admits at least $2n$ solutions if λ is small enough. We will show that there are at most 2^n solutions of $(0.1)_\lambda$ for each $\lambda>0$.

* Received by the editors May 15, 1985; accepted for publication (in revised form) January 13, 1986.

† Institut für Angewandte Mathematik, Universität Hamburg, Bundesstraße 55, D-2000 Hamburg 13, Federal Republic of Germany.

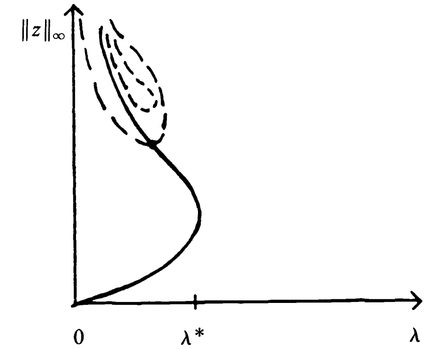


FIG. 1

Another type of spurious solution is obtained by the discretization of $(0.4)_\lambda$ with $f(z) = \sin z$. Equation $(0.4)_\lambda$ admits at most one positive solution, the solid branch in Fig. 2, whereas $(0.1)_\lambda$ possesses an increasing number of solutions as $\lambda \to \infty$ [12], the dotted branches. It is shown in [12] that there are at least $(2 \cdot 3 + 1)^n$ solutions of $(0.1)_\lambda$ with a maximum norm less than 3π, if λ is sufficiently large. We state that there are at most 7^n such solutions for each λ.

A third type of spurious solution occurs in the case $f(z) = (p-z)/1+(p-z)+(p-z)^2$ with $p > 0$ arbitrary which has been considered in [9].

The solution set of $(0.4)_\lambda$ has to be a one-dimensional manifold (cf. [10]), the S-curve in Fig. 3, whereas numerical calculations in [9] show that $(0.1)_\lambda$ with $n = 2$

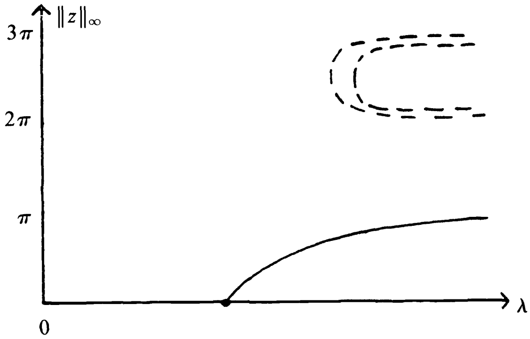


FIG. 2

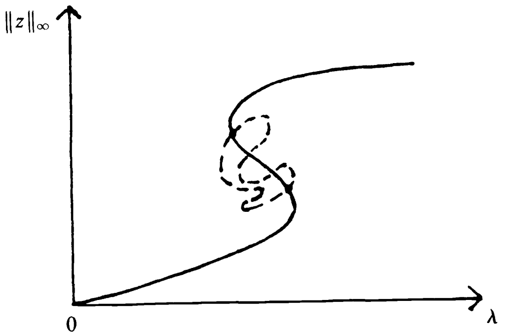


FIG. 3

meshpoints admits 3^2 solutions for specific values of λ and p, the dotted branch. The number of solutions increases with n. Our consideration yields that there are at most 3^n solutions of $(0.1)_\lambda$ for a fixed λ. In the last example the system $(0.1)_\lambda$ is more important than the bvp, because $(0.1)_\lambda$ describes a chemical reaction in an assemblage of n biological cells (cf. [9]). z_j denotes the concentration of a substrate S in the cell C_j which is coupled with its neighbors by diffusion through a membrane. From this point of view all solutions of $(0.1)_\lambda$ are physically relevant.

To obtain our results, two ideas are used. First: relax the coupling between the cells, i.e., we consider the homotopy

$$(0.5)_\lambda \qquad H_\lambda(t,z) := ((1-t)I + tA)z - \lambda F(z) = 0,$$

where $t \in [0, 1]$ and I denotes the identity. If $t=1$ we obtain $(0.1)_\lambda$, the fully coupled state, whereas $t=0$ is the uncoupled state, i.e., a system of n independent equations. In the latter case the number of solutions of $(0.5)_\lambda$ is given by $r(\lambda)^n$, where $r(\lambda)$ denotes the number of solutions of the scalar equation

$$(0.6)_\lambda \qquad \lambda f(z) = z, \qquad z \in \mathbb{R}.$$

Now we choose a solution z_0 of $H_\lambda(0, z)=0$ at a fixed λ and follow the solution curve $(t, z(t))$ of $H_\lambda(t, z)=0$ with $z(0)=z_0$ in the direction of $t \geqq 0$. In § 1 this is examined numerically in the case $f(z)=\exp(z)$.

In the above examples we are able to show that all solutions of $H_\lambda(1, z)=0$ are induced by the solutions of $H_\lambda(0, z)=0$, i.e., the number is limited by $r(\lambda)^n$.

This will be done in §§ 3, 4 and 5. To this end we use a second idea: extend $(0.5)_\lambda$ to a complex analytic mapping in $\mathbb{C}^n$ for each (t, λ). Now degree arguments for complex analytic mappings are used to examine the number of solutions of $(0.1)_\lambda$ in $\mathbb{C}^n$. The special properties of the Brouwer degree for those mappings are listed in § 3.

1. Numerical results for $f(z)=\exp(z)$.

Here we examine a slight modification of the homotopy $(0.5)_\lambda$ with $f(z)=\exp(z)$ and the simplest discretization matrix A of $-u''$. We consider

$$(1.1)_\lambda \qquad \tilde{H}_\lambda(t,z) = ((1-t)\mu I + tA)z - \lambda F(z) = 0,$$

with the lowest eigenvalue μ of A. This seems to be more promising, since the solution sets of $(0.1)_\lambda$ and of $\lambda f(z)=\mu z$ are strongly related. Define $\lambda^* := \sup\{\lambda | (0.1)_\lambda \text{ is solvable}\}$, then the results of [15] yield

$$(1.2) \qquad \lambda_0^*/\mu \|A^{-1}\|_\infty \leqq \lambda^* \leqq \lambda_0^*,$$

where $\lambda_0^* = \mu \exp(1)^{-1}$. Observe that there is no solution of $\tilde{H}_\lambda(0, z)=0$, if $\lambda > \lambda_0^*$. If $\lambda < \lambda_0^*$ there are two solutions of $\lambda f(z)=\mu z$; we denote the minimal one by $+$ and the maximal one by $*$.

We have examined $(1.1)_\lambda$ numerically for $n=3$ and $n=4$ at fixed values of λ. If $n=3$, there are 2^3 solutions of $\tilde{H}_\lambda(0, z)=0$ which we denote in form $z_1 = (+,+,+)$, $z_2(+,*,+), \cdots, z_8=(*,*,*)$. Now we used a path following method as described in [11] to follow the solution curve of $(1.1)_\lambda$ through $(0, z_j)$ in direction of $t>0$.

If $\lambda \in (\lambda^*, \lambda_0^*)$, we cannot reach $t=1$ because of (1.2). Results for various values of λ are shown schematically in Fig. 4. If λ is decreased further, the branch in Fig. 4 connecting $(*,*,*)$ with $(*,+,*)$ will reach $t=1$. Thus there are at least 2^3 solutions of $(1.1)_\lambda$ at $t=1$ if λ is chosen sufficiently small. Fig. 5 shows the case $n=4$, where the 2^4 solutions of $\tilde{H}(0, z)=0$ are denoted in the same manner.

Quite the same results are obtained in the simple case $n=2$. Observe that there is a significant difference between $n=4$ and $n=3$. In the first case bifurcation occurs,

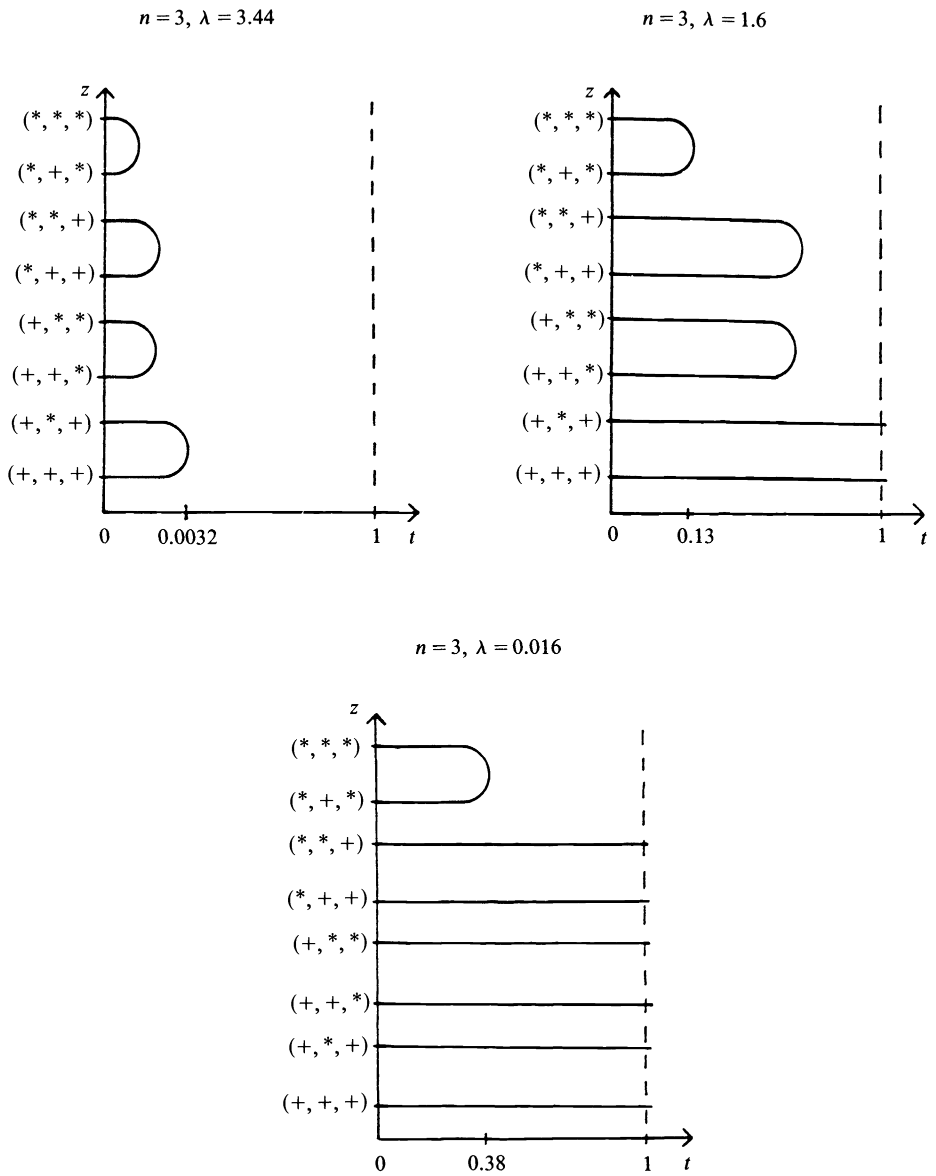


FIG. 4

but not in the latter. We expect diagrams analogous to Figs. 4 and 5 and for the example $f(z) = \sin z$. In this case all branches should reach $t = 1$, if λ is chosen sufficiently large.

With the above computational results the question arises whether all solutions of $(0.1)_\lambda$ are induced by the solutions of the uncoupled system. To answer this question we rewrite $(0.1)_\lambda$ as an equation in $\mathbb{C}^n$ for each $\lambda \in \mathbb{R}_+$, where $F: \mathbb{C}^n \to \mathbb{C}^n$ is an extension of $F: \mathbb{R}^n \to \mathbb{R}^n$. If $f(z) = \exp(z)$ or $f(z) = \sin z$ there is a well-defined extension such that F is a complex analytic mapping. In order to examine $(0.1)_\lambda$ in $\mathbb{C}^n$, we have to deal with the Brouwer degree for complex analytic mappings. In the next section we develop its properties for the reader's convenience.

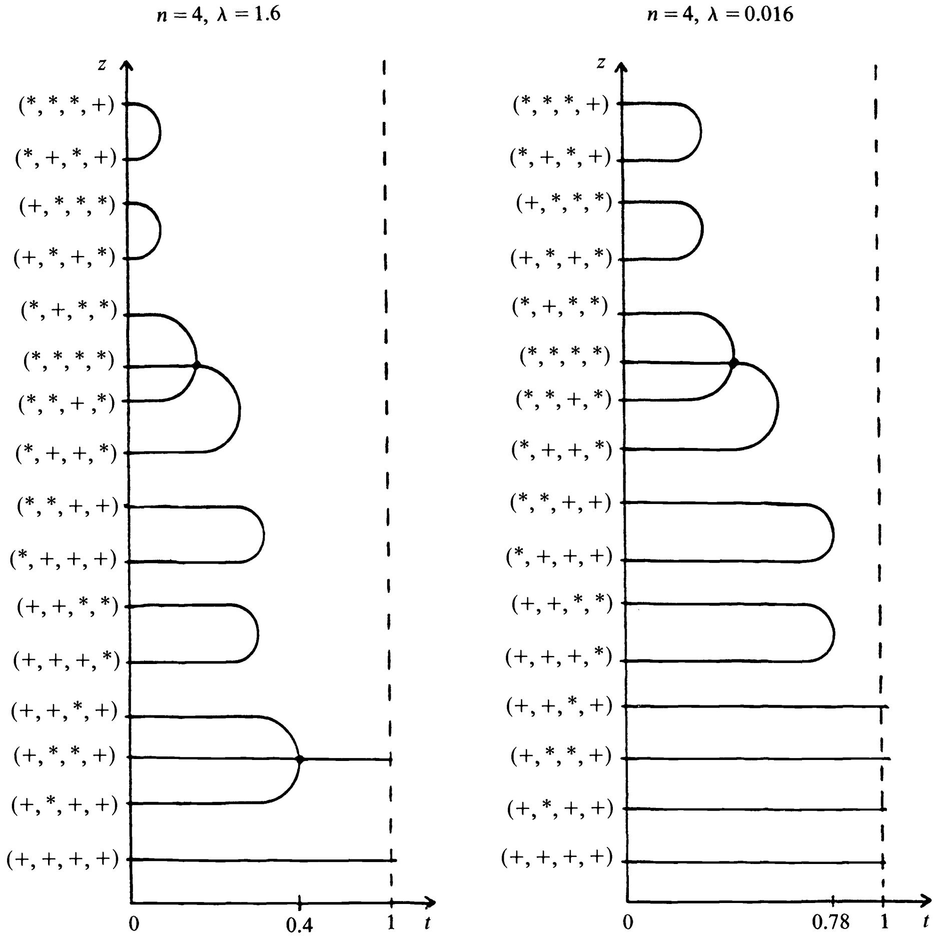


FIG. 5

2. The Brouwer degree of complex analytic mappings. Let G be a complex analytic mapping of an open subset V of $\mathbb{C}^n$ into $\mathbb{C}^n$ and let U be a bounded open subset of V with $\bar{U} \subset V$. By ignoring the complex structure, $G = (g_1 + ih_1, \cdots, g_n + ih_n)$, $g_j, h_j : \mathbb{C}^n \to \mathbb{R}$ can be interpreted as a mapping $(g_1, \cdots, g_n, h_1, \cdots, h_n)$ from $V \subset \mathbb{R}^{2n}$ into $\mathbb{R}^{2n}$. Let w be a point of $\mathbb{R}^{2n} \backslash G(\partial U)$, then the Brouwer degree $\deg(G, U, w)$ is well defined and the usual properties hold (cf. [8] for a detailed discussion).

If the analyticity of G is taken into account these properties can be strengthened. A detailed description of the following considerations can be found in [5], [7], [14].

Denote by $G'(z)$ the complex Jacobian matrix of G at $z = (z_1, \cdots, z_n) = (x_1 + iy_1, \cdots, x_n + iy_n) \in \mathbb{C}^n$ and by $J(G)(z)$ the real Jacobian of $G: V \to \mathbb{R}^{2n}$ at $\tilde{z} = (x_1, \cdots, x_n, y_1, \cdots, y_n)$, then we have

$$\text{Det}\,(J(G)(\tilde{z})) = |\text{Det}\,(G'(z))|^2, \tag{2.1}$$

i.e., the determinant of the real Jacobian is always nonnegative (cf. [5]). Using the identity theorem, it is shown in [5] that for each connected open subset V_1 of V the set

$$S_\varepsilon := \{\tilde{z} \in V_1 \,|\, \text{Det}\,(J(G_\varepsilon)(\tilde{z})) = 0\}$$

of singular points of $G_\varepsilon(z) := G(z) + \varepsilon z$ is nowhere dense in V_1, if $\varepsilon \geqq 0$ is chosen sufficiently small. A theorem of Remmert and Stein [13] yields that the number of solutions of $G(z) = w$ with $z \in U$ is finite. These three properties are responsible for the fact that the following statements hold (cf. [5], [14]).

PROPOSITION 2.1. *Let G, V, U, w be as above. Then*

(2.2) $\deg(G, U, w) \geqq 0$,

(2.3) $\deg(G, U, w) > 0$ *if and only if w lies in $G(U)$,*

(2.4) $\deg(G, U, w) = 1$ *if and only if for the component C of $\mathbb{C}^n \backslash G(\partial U)$ which contains w, G is a one-to-one bicontinuous map of $G^{-1}(C) \cap U$ onto C,*

(2.5) $\deg(G, U, w) \geqq 2$ *if and only if either there is more than one solution in U of $G(z) = w$, or the unique solution is a singular point of G.*

In the following we choose $w = 0$ for convenience. Since there are only finitely many solutions $z_1, \cdots, z_m$ of $G(z) = 0$ in U, the local degree $\deg(G, z_j)$ is well defined for each solution z_j. By the additivity of the degree and (2.3) we have

$$\deg(G, U, 0) = \sum_{j=1}^{m} \deg(G, z_j) \geqq m. \tag{2.6}$$

Thus the degree is an upper bound for the number of solutions.

If G is a function of $U \subset \mathbb{C}$ into $\mathbb{C}$, the degree equals the winding number of $G(\partial U)$ with respect to zero (cf. [8]). In this case it is well known that the local degree $\deg(G, z_j)$ equals the multiplicity of the solution z_j, i.e., the lowest number k at which $G^{(k)}(z_j) \neq 0$. An analogous characterisation of the local degree is valid in general. Since G is analytic, we expand G into the Taylor series at a solution z_j. Without loss of generality we assume that $z_j = 0$.

$$G_j(z) = \sum_{k=1}^{\infty} P_j^{(k)}(z_1, \cdots, z_n), \qquad j = 1, \cdots, n,$$

where $P_j^{(k)}$ is a polynomial in $z_1, \cdots, z_n$ and homogeneous of degree k. Suppose $P_j^{(k_j)}$ is the polynomial of lowest degree in the jth series which is not identically zero. Then it is shown in [6] that

$$\deg(G, 0) = \prod_{j=1}^{n} k_j,$$

if the local degree of the map $(P_1^{(k_1)}, P_2^{(k_2)}, \cdots, P_n^{(k_n)}): \mathbb{C}^n \to \mathbb{C}^n$ is well defined, i.e., $z = 0$ is an isolated zero of this map. Otherwise $\deg(G, 0) \geqq \prod_{j=1}^{n} k_j$ (cf. [6]). Thus it is appropriate to define the multiplicity of a solution of $G(z) = 0$ by its local degree. In this respect (2.6) means that the degree $\deg(G, U, 0)$ equals the number of solutions of $G(z) = 0$, each counted according to its multiplicity.

The considerations above and the homotopy invariance of the degree yield Proposition 2.2 which is basic for our considerations.

PROPOSITION 2.2. *Let V and U be as above. Suppose $H(t, \cdot): V \to \mathbb{C}^n$ is complex analytic for each $t \in [0, 1]$ and continuous in t. Suppose $0 \notin H(t, \cdot)(\partial U)$ for each $t \in [0, 1]$. Then $\deg(H(t, \cdot), U, 0)$ is independent of t, i.e., the number of solutions (counted according to their multiplicities) of $H(t, z) = 0$ in U is finite and independent of t.*

The situation is illustrated in Fig. 6. It is shown in [1], where different methods are used to examine the solution set of $H(t, z) = 0$, that there are indeed solution curves connecting $t = 0$ and $t = 1$.

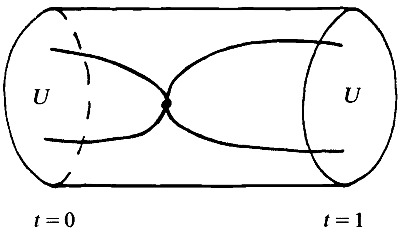


FIG. 6

Now we suppose that the domain U satisfies $U=U^*:=\{z\in\mathbb{C}^n \mid z^*\in U\}$, where z^* denotes the conjugate of z. The mappings we consider fulfill the additional property

$$(2.7)\qquad G(z^*)=(G(z))^*$$

for all $z\in U$. Especially we have $G|_{U\cap\mathbb{R}^n}\subset\mathbb{R}^n$, i.e., G is an extension of a real mapping. Suppose $0\notin G(\partial U)$. Let $\{z_1,\cdots,z_k\}\subset\mathbb{R}^n$ be the set of real solutions of $G(z)=0$. There exists an open subset $W\subset U$ such that

$$(2.8)\qquad W\cap(G^{-1}(0)\cap U)=\{z_1,\cdots,z_k\},\qquad 0\notin G(\partial W).$$

$U\setminus\bar{W}$ is an open set such that $0\notin G(\partial(U\setminus\bar{W}))$. By the additivity of degree we have

$$(2.9)\qquad \deg(G,U,0)=\deg(G,W,0)+\deg(G,U\setminus\bar{W},0).$$

PROPOSITION 2.3. $\deg(G,U\setminus\bar{W},0)$ *is a nonnegative even number.*

Proof. If $G(z)=0$, it follows from (2.7) that $G(z^*)=0$. Denote by $z_1,\cdots,z_l,z_1^*,\cdots,z_l^*$ the solutions of $G(z)=0$ in $U\setminus\bar{W}$, then we have

$$\deg(G,U\setminus\bar{W},0)=\sum_{j=1}^{l}\deg(G,z_j)+\sum_{j=1}^{l}\deg(G,z_j^*).$$

It remains to show that $\deg(G,z_j)=\deg(G,z_j^*)$. This has been done in [7, Proof of Lemma 2]. □

If G satisfies (2.7), it follows from (2.9) and Proposition 2.3 that the number of real zeros of G is even if and only if $\deg(G,U,0)$ is even.

3. The case $f(z)=\exp(z)$.

We choose $F(z):=(\exp(z_1),\cdots,\exp(z_n))$ and examine $(0.1)_\lambda$ in $\mathbb{C}^n$, i.e.,

$$(3.1)_\lambda\qquad Az=\lambda F(z),\qquad \lambda\in\mathbb{R}_+,\quad z\in\mathbb{C}^n.$$

We will apply Proposition 2.2 to the homotopy $(0.5)_\lambda$. Clearly $H_\lambda(t,\cdot):\mathbb{C}^n\to\mathbb{C}^n$ is complex analytic for each (λ,t). To compute $\deg(H_\lambda(0,\cdot),U,0)$ on an appropriate domain U, we deal with the one-dimensional equation

$$(3.2)_\lambda\qquad g_\lambda(z):=z-\lambda\exp(z)=0,\qquad z\in\mathbb{C},\quad \lambda\in\mathbb{R}_+$$

on the domain $U_r:=\{z=x+iy\in\mathbb{C} \mid |x|<r, y\in(-\pi,+\pi)\}$.

PROPOSITION 3.1. *Let* $\lambda>0$ *be arbitrary. Then* $\deg(g_\lambda,U_r,0)=2$, *if* r *is sufficiently large.*

Proof. Observe that $\exp(z)\in\mathbb{R}$, if $z=(x\pm i\pi)$ and that the solution set of $(3.2)_\lambda$ is equibounded with respect to $\lambda\in[\tilde{\lambda},\hat{\lambda}]$ with $\tilde{\lambda},\hat{\lambda}>0$ arbitrary. Thus $\deg(g_\lambda,U_r,0)$ is well defined and independent of $\lambda>0$, if r is chosen sufficiently large. An elementary calculation shows that each solution $z\in U_r$ of $(3.2)_\lambda$ is real and nonsingular, if

$\lambda < \exp(1)^{-1}$. Since there are exactly two solutions of $(3.2)_\lambda$, if $\lambda < \exp(1)^{-1}$ we conclude $\deg(g_\lambda, U_r, 0) = 2$ for those λ. □

Fig. 7 exhibits the real and the imaginary part of the solution set of $(3.2)_\lambda$ in U_r.

Remark 3.2. There exists an infinite number of solutions of $(3.2)_\lambda$ in $\mathbb{C}$ for each λ. Elementary calculations yield that there are exactly two solutions $z = x \pm iy$ with $y \in (k\pi, (k+2)\pi)$ for each number $k = 1, 3, 5, \cdots$.

Now we define $U := U_r^n \subset \mathbb{C}^n$. Choose r sufficiently large, then, for a fixed λ, all solutions of $H_\lambda(0, z) = 0$ are within U. Thus

$$\deg(H_\lambda(0, \cdot), U, 0) = 2^n \tag{3.3}$$

by Proposition 3.1 and the Cartesian product formula for the degree. In the following we denote by $\operatorname{Re}(z) := (x_1, \cdots, x_n)$, $\operatorname{Im}(z) := (y_1, \cdots, y_n)$ the real and the imaginary part of $z = (y_1 + iy_1, \cdots, x_n + iy_n) \in \mathbb{C}^n$.

THEOREM 3.3. *Let $\lambda > 0$ be fixed. Equation $(3.1)_\lambda$ has exactly 2^n solutions (counted according to their multiplicities) with $\operatorname{Im}(z) \in (-\pi, \pi)^n$. Thus, with λ^* defined as in* § 1 *there are at most 2^n distinct real solutions of $(3.1)_\lambda$ for each $\lambda < \lambda^*$. At $\lambda = \lambda^*$ there is exactly one real solution.*

Proof. The assertion follows from Proposition 2.2 and (3.3), if we show that for each $t \in [0, 1]$ there are no solutions of $H_\lambda(t, z) = 0$ on ∂U, if r is sufficiently large. The last assertion can be deduced from the results in [2, Chap. 5].

Suppose there exists $(t, z) \in [0, 1] \times \partial U$ with $H_\lambda(t, z) = 0$ such that $z_j = x_j + i\pi$ for one $j \in \{1, \cdots, n\}$. Then $(\lambda F(z))_j \in \mathbb{R}$ and therefore

$$((1-t) + ta_{jj})\pi + t \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} y_k = 0. \tag{3.4}$$

Since $y_k \in [-\pi, \pi]$ and since A satisfies (0.2), it follows that $y_k = \pi$ in (3.4) for each k with $a_{jk} \neq 0$. Since A is irreducible, (3.4) must hold for each $j \in \{1, \cdots, n\}$ with $y_k = \pi$. But this contradicts the second property of A in (0.2). Using the same arguments we show that there could be no solution of $H_\lambda(t, z) = 0$ with $z_j = x_j - i\pi$ for $j \in \{1, \cdots, n\}$.

Suppose there is a sequence $(t, z(t))$ with $H_\lambda(t, z(t)) = 0$ and $\operatorname{Im}(z(t)) \in [-\pi, \pi]^n$ such that $\|z(t)\|_\infty := \max |z_j(t)|$ becomes unbounded. Then there is a $j_0 \in \{1, \cdots, n\}$

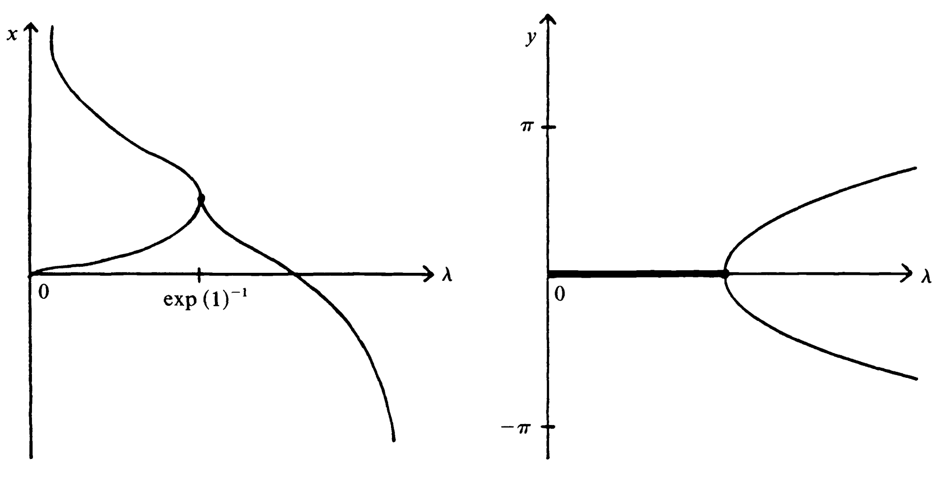


FIG. 7

such that $|x_{j_0}(t)|\to\infty$. Suppose $x_{j_0}(t)\to\infty$. Then we have

$$\lambda=\frac{\|((1-t)I+tA)z(t)\|_\infty}{\|F(z(t))\|_\infty}\leqq\frac{\|(1-t)I+tA\|_\infty(x_{j_0}^2+\pi^2)^{1/2}}{\exp(x_{j_0}(t))}\to 0.$$

This is a contradiction, since λ is fixed.

Suppose $x_{j_0}(t)\to-\infty$. $(1-t)I+tA$ satisfies (0.2), thus the inverse exists and is equibounded with respect to t, i.e., $\|((1-t)I+tA)^{-1}\|_\infty\leqq c$ for all $t\in[0,1]$. Therefore

$$\lambda\geqq\frac{\|z(t)\|_\infty}{c\|F(z(t))\|_\infty}\to\infty, \tag{3.5}$$

since $\|F(z(t))\|_\infty$ is bounded. Formula (3.5) contradicts the choice of λ. □

COROLLARY 3.4. *Suppose $\lambda\leqq\lambda^*$. The number m of real solutions* (*counted according to their multiplicities*) $z\in\mathbb{R}^n$ *of* $(3.1)_\lambda$ *is even.*

Proof. Choose $U\subset\mathbb{C}^n$ as above and an open subset $W\subset U$ as in (2.8) which only contains the real solutions of $H_\lambda(1,z)=0$. Then we deduce

$$2^n=m+\deg(H_\lambda(1,\cdot),U\backslash\bar{W},0)$$

as in (2.9) and m must be even by Proposition 2.3. □

If we change λ in $(3.1)_\lambda$, then we know from Corollary 3.4 that the real solutions appear and disappear pairwise, as is shown in Figs. 1–3.

Remark 3.5. Using Remark 3.2 and the same arguments as in the proof of Theorem 3.3, we conclude that there are exactly $(2k)^n$ solutions of $(3.1)_\lambda$ in the domain $\{z\in\mathbb{C}^n\,|\,\mathrm{Im}(z)\in[-k\pi,k\pi]^n\}$ for each odd number k.

If $(3.1)_\lambda$ is the discretization of the differential equation $(0.4)_\lambda$, one asks for the number of symmetric solutions of $(3.1)_\lambda$, since all solutions of $(0.4)_\lambda$ are symmetric with respect to $\frac{1}{2}$ (cf. [10]). Suppose that the meshpoints are chosen equidistant. Then $(3.1)_\lambda$ reduces to an equation in $\mathbb{R}^{n/2}(\mathbb{C}^{n/2})$, if n is even and to an equation in $\mathbb{R}^{(n+1)/2}(\mathbb{C}^{(n+1)/2})$, if n is odd. Using the same arguments as above, we conclude that $(3.1)_\lambda$ has exactly $2^{n/2}$ symmetric solutions (counted according to their multiplicities) in the former case and $2^{(n+1)/2}$ in the latter. More generally, let L be an (n,n) permutation matrix which commutes with A, i.e., $A\circ L=L\circ A$. Denote by $S\subset\mathbb{R}^n$ the subspace which satisfies $L|_S=I$, by $\tilde{S}$ its complexification, and by $m:=\dim(S)$.

THEOREM 3.6. *Let $\lambda>0$ be fixed. Then there are exactly 2^m solutions of* $(3.1)_\lambda$ (*counted according to their multiplicities*) *in $\tilde{S}$ with* $\mathrm{Im}(z)\in(-\pi,\pi)^n$. *Thus there are at most 2^m* (*real*) *solutions of* $(3.1)_\lambda$ *in S for each $\lambda<\lambda^*$. At $\lambda=\lambda^*$ there is exactly one solution in S.*

Proof. By definition of $F(\cdot)$, L commutes with $F(\cdot)$ and therefore with $H_\lambda(t,\cdot)$ for each $\lambda>0$ and $t\in[0,1]$. Thus $H_\lambda(t,\tilde{S})\subset\tilde{S}$. There exist 2^m zeros of $H_\lambda(0,\cdot)$ in $\tilde{S}\cap U$ with U as above. Since there are no solutions of $H_\lambda(t,z)=0$ on ∂U for each $t\in[0,1]$, the first assertion follows from Proposition 2.2 as in the proof of Theorem 3.3. The last assertion follows again from the results in [2, Chap. 5]. □

With an appropriate modification, Corollary 3.4 is also valid in the above case. All results of this section remain valid, if we consider $F(z)=(c_1\exp(z_1)\cdots c_n\exp(z_n))$ with positive constants c_j.

4. The case $f(z)=\sin z$. We now study $(0.1)_\lambda$ with $F(z)=(\sin z_1,\cdots,\sin z_2)$ in $\mathbb{C}^n$. Analogously to § 3 we use the homotopy $(0.5)_\lambda$ and Proposition 2.2. To this end we have to examine the scalar equation

$$g_\lambda(z):=z-\lambda\sin z=0,\qquad z\in\mathbb{C},\quad\lambda\in\mathbb{R}_+. \tag*{$(4.1)_\lambda$}$$

Since $\sin z = \sin x \cosh y + i \cos x \sinh y$, there are no solutions of $(4.1)_\lambda$ with $z = \pm k\pi + iy$, $y \in \mathbb{R}$, $k \in \mathbb{N}$. For a fixed k we define $U_r := \{z \in \mathbb{C} | x \in (-k\pi, k\pi), |y| < r\}$. Let $\tilde{\lambda} > 0$ be arbitrary then we can choose $r > 0$ such that for each $\lambda \in [\tilde{\lambda}, \infty)$ all solutions of $(4.1)_\lambda$ with real part in $[-k\pi, k\pi]$, are contained in U_r. If λ is sufficiently large, all solutions in U_r become real and nonsingular. Thus for those λ we have

$$(4.2) \qquad \deg(g_\lambda, U_r, 0) = \begin{cases} 2k-1 & \text{if } k \text{ is even,} \\ 2k+1 & \text{if } k \text{ is odd.} \end{cases}$$

By the homotopy invariance of the degree, (4.2) holds for each $\lambda > 0$, if r is sufficiently large.

Now we consider the homotopy $(0.5)_\lambda$ for a fixed $\lambda > 0$. Clearly $H_\lambda(t, \cdot)$ is complex analytic for each (λ, t). Define $U := U_r^n \subset \mathbb{C}^n$. If r is sufficiently large, we obtain

$$(4.3) \qquad \deg(H_\lambda(0, \cdot), U, 0) = \begin{cases} (2k-1)^n & \text{if } k \text{ is even,} \\ (2k+1)^n & \text{if } k \text{ is odd,} \end{cases}$$

by (4.2). Now we can prove the following theorem.

THEOREM 4.1. *Let $\lambda > 0$ be fixed. The nonlinear system*

$$(4.4)_\lambda \qquad Az = \lambda F(z), \qquad z \in \mathbb{C}^n$$

with $F(z) = (\sin z_1, \cdots, \sin z_n)$ has exactly $(2k+1)^n$ solutions (counted according to their multiplicities) with $\operatorname{Re}(z) \in [-k\pi, k\pi]^n$, if $k \in \mathbb{N}$ is odd and exactly $(2k-1)^n$ solutions, if k is even.

Proof. Using (4.3) the assertion follows from Proposition 2.2 if we show that for each $t \in [0, 1]$, there are no solutions of $H_\lambda(t, z) = 0$ on ∂U, if r is sufficiently large. Using the definition of $\sin z$ and the property (0.2) of the matrix A we can deduce this as in the proof of Theorem 3.3. □

Remark 4.2. With slight reformulations the considerations in § 3 concerning the number of symmetric solutions and the appearance of real solutions remain valid in case of the above nonlinearity.

Using an argument given in [12], we can show that all solutions of $(4.4)_\lambda$ become real if λ increases. Let $\tilde{z} \in \mathbb{C}^n$ be a zero of $F(\cdot)$, i.e., $\tilde{z} = (k_1\pi, \cdots, k_n\pi)$, $k_j \in \mathbb{N}$. Consider $F(\cdot)$ as a mapping from $\mathbb{R}^n$ into itself. Dividing $(4.4)_\lambda$ by λ and using the homotopy invariance property, we compute the local degree

$$\deg(A/\lambda - F, \tilde{z}) = \deg(F, \tilde{z}) = \pm 1$$

if λ is sufficiently large. Thus for those λ there exists a solution $z \in \mathbb{R}^n$ of $(4.4)_\lambda$ in a small neighborhood of $\tilde{z}$. This consideration together with Theorem 4.1 yields the following corollary.

COROLLARY 4.3. *For each $k \in \mathbb{N}$ there exists a value $\lambda(k)$ such that for each $\lambda > \lambda(k)$ the system $(4.4)_\lambda$ admits exactly $(2k+1)^n$ different (real) solutions in $[-k\pi, k\pi]^n$ if k is odd and exactly $(2k-1)^n$ solutions if k is even.*

5. Polynomial and related nonlinearities. First we consider $(0.1)_\lambda$ with $F(z) = (p_1(z_1), \cdots, p_n(z_n))$, where $p_j: \mathbb{C} \to \mathbb{C}, j \in \{1, \cdots, n\}$ are polynomials of degree $m_j > 1$. We use the ideas of the preceding sections to prove the following theorem.

THEOREM 5.1. *Let $\lambda > 0$ be fixed and let $F(\cdot)$ be as above. Then there are exactly $\prod_{j=1}^n m_j$ solutions (counted according to their multiplicities) of $(0.1)_\lambda$ in $\mathbb{C}^n$.*

Proof. We use the homotopy $(0.5)_\lambda$. Each component of $H_\lambda(t, \cdot)$ is a polynomial $P_j: \mathbb{C}^n \to \mathbb{C}$ of the form $P_j(z) = p_j(z_j) + q_j(z)$, where q_j is a polynomial of degree one. Therefore it is easy to see that there exists a ball $B_r \subset \mathbb{C}$ with center 0 and radius r

such that for all $t \in [0,1]$ the solutions of $H_\lambda(t,z)=0$ are contained in B_r^n. Since $\deg(p, B_r, 0)$ of a polynomial $p:\mathbb{C}\to\mathbb{C}$ equals the degree of the polynomial we obtain

$$\deg(H_\lambda(0,\cdot), B_r^n, 0) = \prod_{j=1}^{n} m_j.$$

Now the assertion follows from Proposition 2.2. □

Now we consider the third example of § 0, i.e., $(0.1)_\lambda$ with $F(z) = (f(z_1), \cdots, f(z_n))$, where

$$f(z) = \frac{(p-z)}{1+(p-z)+(p-z)^2}, \qquad z \in \mathbb{R}, \quad p > 0.$$

If $n=2$, at most 3^2 solutions occur (cf. (Fig. 3). The following theorem shows that this is true in general.

THEOREM 5.2. *Let $\lambda > 0, p > 0$ be fixed and $F(\cdot)$ be as above. Then there are 3^n solutions (counted according to their multiplicities) of $(0.1)_\lambda$ in $\mathbb{C}^n$.*

Proof. We consider the homotopy $(0.5)_\lambda$ in $\mathbb{C}^n$. But $H_\lambda(t,\cdot)$ is not complex analytic on $\mathbb{C}^n$. Thus for each $j \in \{1, \cdots, n\}$ we multiply the jth row of $(0.1)_\lambda$ by $(1+(p-z_j)+(p-z_j)^2)$ and obtain a mapping $\tilde{H}_\lambda(t,\cdot):\mathbb{C}^n \to \mathbb{C}^n$ which is complex analytic for each (λ, t). The solution sets of $H_\lambda(t,z)=0$ and $\tilde{H}_\lambda(t,z)=0$ coincide. There exists a ball $B_r \subset \mathbb{C}$ with center 0 and radius r such that for each $t \in [0,1]$ all solutions of $\tilde{H}_\lambda(t,z)=0$ are contained in $B_r^n \subset \mathbb{C}^n$. Otherwise there would exist a sequence $(t, z(t)) \subset \mathbb{C}^n$ of solutions of $\tilde{H}_\lambda(t,z)=0$ with $|z_{j_0}(t)| \to \infty$ for a $j_0 \in \{1, \cdots, n\}$ and $|z_{j_0}(t)| \geqq |z_j(t)|$ for all $j \in \{1, \cdots, n\}$. Consider the j_0th row of $\tilde{H}_\lambda(t,z)=0$,

$$\begin{aligned}|((1-t)+ta_{j_0j_0})z_{j_0}(1+(p-z_{j_0})+(p-z_{j_0})^2)\\ +t\sum_{\substack{k=1\\k\neq j_0}}^{n} a_{j_0k}z_k(1+(p-z_{j_0})+(p-z_{j_0})^2)-\lambda(p-z_{j_0})|=0.\end{aligned}$$

Since the matrix A satisfies (0.2), the expression on the left converges to infinity, a contradiction.

Denote by $\tilde{H}_\lambda^j(\cdot)$ the jth row of $\tilde{H}_\lambda(0,\cdot)$. $\tilde{H}_\lambda^j:\mathbb{C}\to\mathbb{C}$ is a polynomial of degree three. Thus by Proposition 2.2 and the Cartesian product formula we obtain

$$\deg(\tilde{H}_\lambda(1,\cdot), B_r^n, 0) = \prod_{j=1}^{n} \deg(\tilde{H}_\lambda^j, B_r, 0) = 3^n.$$

This proves the assertion. □

Observe that $\tilde{H}_\lambda^j(z)=0$, $z \in \mathbb{C}$ is equivalent to $\lambda f(z)=0$ with $f(\cdot)$ as above. Thus also in this example the number of solutions of $(0.1)_\lambda$ is limited by the number of solutions of the uncoupled state. The considerations in § 3 about the number of symmetric solutions also holds in this case.

REFERENCES

[1] E. ALLGOWER, *Bifurcations arising in the calculation of critical points via homotopy methods*, in Numerical methods for bifurcation problems, Internat. Schriftenreihe Numer. Math., 70, T. Küpper, H. D. Mittelmann, H. Weber, eds., Birkhäuser, Basel-Boston-Stuttgart, 1984, pp. 15-28.

[2] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620-709.

[3] W. J. BEYN AND J. LORENZ, *Spurious solutions for discrete superlinear boundary value problems*, Computing, 28 (1982), pp. 43-51.

[4] E. BOHL, *On the bifurcation diagram of discrete analogues for ordinary bifurcation problems*, Math. Methods Appl. Sci., 1 (1979), pp. 566-571.

[5] F. E. BROWDER, *Nonlinear mappings of analytic type in Banach spaces*, Math. Ann., 185 (1970), pp. 259-278.

[6] J. CRONIN, *Analytic functional mappings*, Ann. of Math., 58 (1953), pp. 175-181.

[7] ———, *Topological degree and the number of solutions of equations*, Duke Math. J., 38 (171), pp. 531-538.

[8] K. DEIMLING, *Nichtlineare Gleichungen und Abbildungsgrade*, Springer-Verlag, Berlin-Heidelberg-New York, 1974.

[9] J. P. KERNEVEZ, *Enzyme mathematics*, in Studies in Mathematics and its Applications 10, J. L. Lions, G. Papanicolaou, R. T. Rockafellar, eds., North-Holland, Amsterdam-New-York-Oxford, 1980.

[10] T. LAETSCH, *The number of solutions of a nonlinear two point boundary value problem*, Indiana Univ. Math. J., 20 (1970), pp. 1-13.

[11] R. MENZEL AND H. SCHWETLICK, *Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen*, Numer. Math., 30 (1978), pp. 65-79.

[12] H. O. PEITGEN, D. SAUPE AND K. SCHMITT, *Nonlinear elliptic boundary value problems versus their finite difference approximations: Numerically irrelevant solutions*, J. Reine Angew. Math., 322 (1981), pp. 74-117.

[13] R. REMMERT AND K. STEIN, *Über die wesentlichen Singularitäten analytischer Mengen*, Math. Ann., 126 (1953), pp. 263-306.

[14] J. T. SCHWARTZ, *Compact analytic mappings of B-spaces and a theorem of Jane Cronin*, Comm. Pure Appl. Math., 16 (1963), pp. 253-260.

[15] H. VOSS AND B. WERNER, *Ein Quotienten Einschliessungssatz für kritischen Parameter nichtlinearer Randwertaufgaben*, Internat. Schriftenreihe Numer. Math., 49 (1979), pp. 147-158.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

007

A FAST, ACCURATE ALGORITHM FOR THE ISOMETRIC MAPPING OF A DEVELOPABLE SURFACE*

JOHN C. CLEMENTS† AND L. J. LEON†‡

Abstract. This work is concerned with the derivation of a fast, accurate algorithm for the isometric mapping of a developable surface onto the plane $\mathcal{M}: \vec{r} \rightarrow \vec{R}$. The algorithm is based on the relationship between the ruling lines $\vec{r}$ generating the developable surface $\vec{S}(s, t)$ and one additional geodesic $\vec{g}(s)$ constructed within the surface as the solution of the geodesic curvature equation $\det(\vec{g}'\vec{g}''\vec{n}) = 0$, where $\vec{n}$ is the unit normal to the surface $\vec{S}$ at $\vec{g}$ and $\vec{R}$ is the image of $\vec{r}$ in the plane. Since $\vec{g}$ as well as the ruling lines reduce to straight lines in the plane, the isometric mapping procedure is defined in terms of the ruling line lengths, the arclength along $\vec{g}$ and the angles of intersection of $\vec{r}$ and $\vec{g}$.

Key words. developable surface, isometric mapping

AMS(MOS) subject classifications. 65, 53

1. Introduction. A surface $\vec{S}$ in $\mathbb{R}^3$ is called a ruled surface if it contains a one-parameter family of straight lines called generators or ruling lines $\vec{r}$, which can be chosen as coordinate curves on the surface. A developable surface is a ruled surface defined by nonintersecting generators which has the same tangent plane at all points of each generator. We shall be concerned here with developable surfaces which can be represented in the form (Fig. 1)

$$\vec{S}(s, t) = \vec{f}(s) + t\vec{r}(s), \qquad \vec{f}'(s) \times \vec{r}(s) \neq 0, \quad s \in [a, b], \quad t \in [0, 1] \tag{1.1}$$

where $\vec{f}$ and $\vec{r}$ are twice continuously differentiable vector functions on $[a, b]$.

The term "developable" refers to the property that by a succession of small rotations about each of the generating lines the surface can be laid flat or developed onto a plane without stretching or tearing. That is, it can be mapped isometrically onto a subset of $\mathbb{R}^2$. Conversely, a plane surface material can be shaped into a developable surface with only simple unidirectional bending along the generating lines.

Currently, an important mathematical problem in computer-aided design and manufacture involves determining whether a given design surface is developable and if so, the precise dimensions of a plane surface material which will produce that surface ([3], [5], [8]-[10]). The use of developable hull forms in shipbuilding offers significant advantages in terms of lower cost and faster and simpler construction techniques ([3], [10]). Developable surfaces are also involved in many other industries such as aircraft manufacture, where they are utilized in the fabrication of airfoil and fuselage sections. The motivation for this work is based on an industry requirement ([3]) for a procedure which would permit the very accurate pre-cutting of steel plate sections greater than 50 feet in length to be used in the construction of developable hull steel ships. Previous approaches to hull plate expansion have involved the use of circular arcs generated from offset data ([2]) or the calculation of very fine surface envelope

* Received by the editors November 29, 1983; accepted for publication (in revised form) August 14, 1986.

† Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H8.

‡ The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A5338.

plane sections ([3]) to define the mapping. The first of these approaches makes no use of the geodesic structure of developable surfaces and is a rough approximation at best. The second is computationally complex with the potential for serious error propagation problems. Neither has the capability for error control.

In what follows, a fast, accurate algorithm is derived for the isometric mapping of developable surfaces onto the plane. This simple algorithm is based on the relationship between the ruling lines $\vec{r}$ generating the developable surface $\vec{S}$ and one additional geodesic $\vec{g}(s)$ constructed within the surface. The algorithm defines a numerical procedure $\mathcal{M}$ for mapping the ruling lines $\vec{r}$ of $\vec{S}$ onto the corresponding plane coordinate lines $\vec{R}$ of the surface developed onto the plane $\mathcal{M}: \vec{r} \to \vec{R}$. Accuracy control is achieved through the application of a variable stepsize differential equation solving routine.

2. Preliminary definitions and results. Let $\mathcal{C}^k[a, b]$ denote the linear space

$$\mathcal{C}^k[a, b] \equiv \{f(s) \,|\, f \text{ is } k \text{ times continuously differentiable for all } s \in [a, b]\}.$$

C will denote a curve in $\mathbb{R}^3$:

$$\begin{aligned} C: \vec{f}(s) &= f_1(s)\vec{i} + f_2(s)\vec{j} + f_3(s)\vec{k} \\ &= (f_1(s), f_2(s), f_3(s))^T, \qquad s \in [a, b], \end{aligned}$$

with components $f_i(s)$, $i = 1, 2, 3$, and Euclidean norm $|\vec{f}(s)| \equiv (f_1^2(s) + f_2^2(s) + f_3^2(s))^{1/2}$. Here T denotes the usual vector transpose and $\circ$ and $\times$ the scalar and vector products respectively.

A ruled surface can be thought of as the surface $\vec{S}$ generated by the continuous motion of a straight line along a curve C (Fig. 1). It will be assumed here ([7]) that $\vec{S}$ has the representation (1.1), where:

(2.1)
- (i) $f_i(s) \in \mathcal{C}^2[a, b]$, $i = 1, 2, 3$,
- (ii) $\vec{r}(s) = (r_1(s), r_2(s), r_3(s))^T$ with $r_i(s) \in \mathcal{C}^2[a, b]$, $i = 1, 2, 3$,
- (iii) each point of $\vec{S}$ corresponds to only one ordered pair (s, t),
- (iv) $t \in [0, 1]$ is the directed distance along $\vec{r}(s)$ from $\vec{f}(s)$.

The ruled surface (1.1) is developable if and only if ([7])

$$\det(\vec{f}'\vec{r}\vec{r}') = 0 \quad \text{for all } s \in [a, b] \tag{2.2}$$

where $' \equiv d/ds$. This is equivalent to the requirement that the tangent planes at every point on a given ruling line must coincide or that the normals must be parallel (Fig. 1).

Let $g(s) = (g_1(s), g_2(s), g_3(s))^T$ be a curve in R^3 with $g_i(s) \in \mathcal{C}^2[a, b]$, $i = 1, 2, 3$. The curvature κ at $\vec{g}(s_0)$ on C: $\vec{g}(s)$, $a \leqq s \leqq b$ is given by

$$\kappa(s_0) = |\vec{K}(s_0)| = |\dot{\vec{T}}(s_0)| = |\vec{T}'(s_0)|/|\vec{g}'(s_0)|$$

where $\vec{T}(s_0) = \vec{g}'(s_0)/|\vec{g}'(s_0)|$. Thus the curvature vector $\vec{K}$ has the same or opposite direction as the principal unit normal $\vec{N}(s_0) = \vec{T}'(s_0)/|\vec{T}'(s_0)|$ to C at $\vec{g}(s_0)$ when $|\vec{T}'(s_0)| > 0$. Here $\cdot \equiv d/d\sigma$ denotes differentiation with respect to the arclength parameter σ. The geodesic curvature $\kappa_{\vec{g}}$ of a curve C: $\vec{g}(s)$ in a surface $\vec{S}$ is given by

$$\kappa_{\vec{g}}(s_0) = \det(\dot{\vec{g}}\ddot{\vec{g}}\vec{n})$$

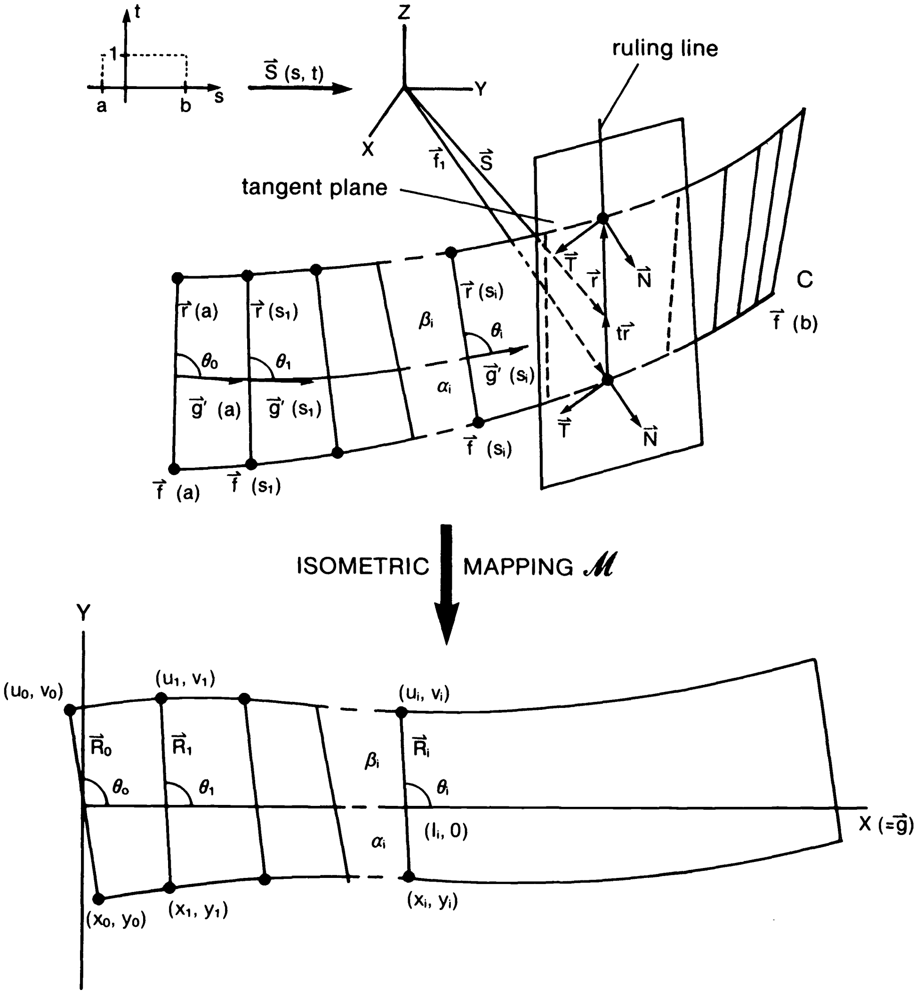


FIG. 1. $\mathcal{M}: \vec{r} \to \vec{R}$.

where $\vec{n}$ is the unit normal to the surface $\vec{S}$ at $\vec{g}(s_0)$. Two important properties of geodesics $\vec{r}$ in a surface $\vec{S}$ are ([8]):

(i) $\vec{g}$ on $\vec{S}$ is a geodesic if and only if

$$\det(\vec{g}'\vec{g}''\vec{n}) = \vec{g}'' \circ (\vec{g}' \times \vec{n}) = 0; \tag{2.3}$$

(ii) If $\vec{S}$ is the developable surface given by (1.1), a geodesic $\vec{g}$ in $\vec{S}$ joining any two points of $\vec{S}$ not on the same generator can be represented in the form

$$\vec{g}(s) = \vec{f}(s) + t^*(s)\vec{r}(s), \qquad t^*(s) \in \mathscr{C}^2[a, b]. \tag{2.4}$$

Equation (2.3) is equivalent to the requirement that the plane of curvature of $\vec{g}$ (provided $\vec{N} \neq 0$) is orthogonal to the tangent plane to $\vec{S}$ (i.e. coincides with $\vec{N}$) at every point of $\vec{g}$ in $\vec{S}$. Equations (2.3) and (2.4) also ensure that $\vec{r} \circ (\vec{g}' \times \vec{n}) \neq 0$ everywhere on $\vec{g}$ in $\vec{S}$.

3. The mapping $\mathscr{M}$: $\vec{r}\to\vec{R}$. Equation (2.3) gives

$$\vec{g}'(s)=\vec{f}'(s)+t^{*\prime}(s)\vec{r}(s)+t^*(s)\vec{r}'(s),$$
$$g''(s)=\vec{f}''(s)+t^{*\prime\prime}(s)\vec{r}(s)+2t^{*\prime}(s)\vec{r}'(s)+t^*(s)\vec{r}''(s)$$

for $s\in[a,b]$. Since $\vec{g}(s)$ must have geodesic curvature zero, it follows from (2.3) that $t^*(s)$ must satisfy the second order nonlinear ordinary differential equation

$$\begin{aligned}t^{*\prime\prime}(s)&=-[\vec{f}''\circ(\vec{g}'\times\vec{n})+t^*\vec{r}''\circ(\vec{g}'\times\vec{n})+2t^{*\prime}\vec{r}'\circ(\vec{g}'\times\vec{n})]/\vec{r}\circ(\vec{g}'\times\vec{n})\\&=F(t^*(s),t^{*\prime}(s))\end{aligned}\tag{3.1}$$

at each s in $[a,b]$, where $\vec{n}(s)=\vec{r}(s)\times\vec{T}(s)$ is the normal to the developable surface $\vec{S}$ at $\vec{g}(s)$ and $\vec{T}(s)=\vec{f}'(s)/|\vec{f}'(s)|$. The equivalent first order system is given by

$$\begin{aligned}u_1'(s)&=u_2(s),\\u_2'(s)&=F(u_1(s),u_2(s)),\end{aligned}\qquad a\leqq s\leqq b,\tag{3.2}$$

for u_1 and u_2 on $[a,b]$, where $u_1(s)=t^*(s)$ and $u_2(s)=t^{*\prime}(s)$. The numerical solution of system (3.2) requires some starting values at $s=a$ and involves the computation of $\vec{f}$, $\vec{f}'$, $\vec{f}''$, $\vec{r}$, $\vec{r}'$ and $\vec{r}''$. For simplicity, the starting values employed here will be $u_1(0)=t^*(a)=.5$ and $u_2(a)=t^{*\prime}(a)=0$. Since the denominator of $\partial F/\partial u_1$ and $\partial F/\partial u_2$ is $[\vec{r}\circ(\vec{g}'\times\vec{n})]^2$, these terms are bounded away from zero except near an edge of regression and F satisfies a Lipschitz condition [4]. If $\vec{r}\circ(\vec{g}'\times\vec{n})$ becomes small, it is only necessary to restart the solution of (3.2) with a new set of initial conditions. Thus, $t^*(s)$ and $t^{*\prime}(s)$ can be solved numerically to within a given specified accuracy ε_{t^*} using a standard variable stepsize differential equation solving routine.

Let $P_N=\{a=s_0<s_1<s_2<\cdots<s_N=b\}$ be any partition of the parameter interval $[a,b]$ and let $\vec{h}(s)$ be defined by $\vec{h}(s)=\vec{S}(s,1)=\vec{f}(s)+\vec{r}(s)$, $s\in[a,b]$. The isometric mapping of the surface $\vec{S}(s,t)$ in (1.1), or more precisely the isometric mapping of the geodesics in $\vec{S}$, is accomplished as follows (Fig. 1). $\vec{g}(a)=\vec{f}(a)+(\frac{1}{2})\vec{r}(a)$ is mapped to the origin of the xy-plane, and computing

$$\begin{aligned}&\vec{r}(a)=\vec{h}(a)-\vec{f}(a),\\&\alpha_0=\beta_0=|\vec{r}(a)|/2,\\&\vec{g}'(a)=\vec{f}'(a)+t^{*\prime}(a)\vec{r}(a)+t^*(a)\vec{r}'(a),\\&\theta_0=\cos^{-1}((\vec{g}'(a)\circ\vec{r}(a))/(|\vec{g}'(a)||\vec{r}(a)|))\end{aligned}\tag{3.3}$$

gives

$$(x_0,y_0)=(-\alpha_0\cos\theta_0,-\alpha_0\sin\theta_0),\qquad(u_0,v_0)=(\beta_0\cos\theta_0,\beta_0\sin\theta_0),$$

and $\vec{R}_0=(u_0-x_0,v_0-y_0)$ is the isometric image of $\vec{r}_0=\vec{r}(a)$. More important, since $\vec{g}(s)$ is a geodesic in the developable surface $\vec{S}$, its image under the mapping must be a straight line in the plane and, by our choice of θ_0, it must be that subinterval of the positive x-axis from zero to the point

$$I_N=\int_{s_0}^{s_N}|\vec{g}'(s)|\,ds.$$

That is, the coordinate in the xy-plane of the intercept $\vec{g}(s_i)$ of the ruling line $\vec{r}(s_i)$

and the geodesic $\vec{g}(s)$ is always given by $(I_i, 0)$, where

$$
\begin{aligned}
I_i &= \int_{s_0}^{s_i} |\vec{g}\,'(s)|\, ds \\
&= \int_{s_0}^{s_i} \sqrt{(f_1' + t^* r_1' + t^{*\prime} r_1)^2 + (f_2' + t^* r_2' + t^{*\prime} r_2)^2 + (f_3' + t^* r_3' + t^{*\prime} r_3)^2}\, ds
\end{aligned}
\tag{3.4}
$$

for every $i = 1, \cdots, N$. To determine the image of $\vec{r}(s_1)$, we compute I_1 and

$$
\begin{aligned}
&\alpha_1 = |\vec{g}(s_1) - \vec{f}(s_1)| = t^*(s_1)|\vec{r}(s_1)|, \\
&\beta_1 = |\vec{g}(s_1) - \vec{h}(s_1)|, \\
&\theta_1 = \cos^{-1}((\vec{g}\,'(s_1) \circ \vec{r}(s_1))/(|\vec{g}\,'(s_1)|\,|\vec{r}(s_1)|))
\end{aligned}
\tag{3.5}
$$

where for simplicity it is assumed that $\vec{g}$ always stays within the surface $\vec{S}$ (that is, that $0 \leqq t^*(s) \leqq 1$ for all s in $[a, b]$). Then

$$
\begin{aligned}
&(x_1, y_1) = (I_1 - \alpha_1 \cos\theta_1, -\alpha_1 \sin\theta_1), \\
&(u_1, v_1) = (I_1 + \beta_1 \cos\theta_1, \beta_1 \sin\theta_1), \\
&\vec{R}_1 = (u_1 - x_1, v_1 - y_1)
\end{aligned}
\tag{3.6}
$$

and that portion of $\vec{S}$ bounded by $\vec{r}(a)$ and $\vec{r}(s_1)$ has been mapped isometrically onto the portion bounded by $\vec{R}_0$ and $\vec{R}_1$ in the plane. Replacing the subscript 1 by 2 in (3.5) and (3.6) and repeating the operation successively for $i = 2, \cdots, N$ completes the isometric mapping of the surface.

The important calculation in this mapping procedure is the numerical evaluation of I_i in (3.4). Consequently, it will be assumed here that a standard quadrature rule is to be used which is to satisfy a user-specified maximum error tolerance ε_I. This in turn induces a refinement $P_{\tilde{N}} = \{a = s_0 = \tau_0 < \tau_1 < \cdots < \tau_{M_1} = s_1 < \cdots < \tau_{M_N} = s_N = b\}$ of P_N and imposes a maximum error tolerance ε_{t^*} on the evaluation of t^* and $t^{*\prime}$ at each point of $P_{\tilde{N}}$.

4. The algorithm. The following algorithm assumes that the geodesic $\vec{g}(s)$ does not encounter an edge of regression of the developable surface and hence that the system of differential equations (3.2) is solvable everywhere on $[a, b]$. One method for ensuring this is to restart the mapping procedure with new initial values at that point on $\vec{g}(s)$ where it crosses one of the boundary curves $\vec{f}$ and $\vec{h}$—that is, whenever $t^*(\tau_j) > 1$ or $t^*(\tau_j) < 0$ for some j.

ALGORITHM. ($\mathcal{M} : \vec{r} \to \vec{R}$). Given a developable surface $\vec{S}$ satisfying (1.1) and (2.1):

(i) specify the error tolerance ε_I and a partition $P_N = \{a = s_0 < s_1 < \cdots < s_N = b\}$ of $[a, b]$,

(ii) choose a quadrature method for evaluating I_i in (3.4) and a refinement $P_{\tilde{N}} = \{a = \tau_0 < \tau_1 < \cdots < \tau_{M_N} = b\}$ of P_N so that I_i is computed accurately to ε_I for each $i = 1, \cdots, N$, provided t^* and $t^{*\prime}$ are computed accurately to within ε_{t^*},

(iii) solve the system of differential equations (3.2) for t^* and $t^{*\prime}$ accurate to ε_{t^*} on $P_{\tilde{N}}$,

(iv) compute $\alpha_0 = \beta_0$ and θ_0 in (3.3) to obtain

$$
\begin{aligned}
&(x_0, y_0) = (-\alpha_0 \cos\theta_0, -\alpha_0 \sin\theta), \\
&(u_0, v_0) = (\beta_0 \cos\theta_0, \beta_0 \sin\theta_0);
\end{aligned}
$$

and

(v) for each $i=1, \cdots, N$ compute I_i, α_i, β_i and θ_i as in (3.5) and (3.6) to obtain the coordinates of the ruling-line endpoints of $\vec{S}$ mapped isometrically,

$$(x_i, y_i)=(I_i-\alpha_i \cos \theta_i, -\alpha_i \sin \theta_i),$$

$$(u_i, v_i)=(I_i+\beta_i \cos \theta_i, \beta_i \sin \theta_i),$$

onto the plane.

5. Discussion. The algorithm developed here has been applied successfully in the design and construction of developable hull steel ships and several improvements have already been proposed. Perhaps the most interesting suggestion has been to compute two geodesics $\vec{g}_1(s)$ and $\vec{g}_2(s)$ simultaneously. This would simplify the remaining calculations and provide an error check on the calculation of θ_i, $i=1, \cdots, N$, which could be sensitive to error propagation. The restriction of t^* to $[0, 1]$ is imposed here for simplicity only. What is required in practice is a check on the stepsize calculation in the differential equation solver for (3.2) or on the magnitude of $|\vec{r} \circ (\vec{g}' \times \vec{n})|$ at each step. A detailed error analysis for the coupled integral-differential equations (3.4) and (3.2) is currently in progress.

REFERENCES

[1] R. E. BARNHILL AND R. F. RIESENFELD, *Computer-Aided Geometric Design*, Academic Press, New York, San Francisco, London, 1974.

[2] H. CABELLO AND T. COSIN, *Shell plate expansion*, presented at Society of Naval Architects and Marine Engineers, Joint Meeting of the Chesapeake and Hampton Roads Section, September 26-28, 1968.

[3] J. C. CLEMENTS, *A computer system to derive developable hull surfaces and table of offsets*, Marine Tech., 18 (1981), pp. 227-233.

[4] G. DAHLQUIST AND A. BJORK, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[5] J. JOHNSSON, *A study of developable hull surfaces*, M.Sc. thesis, Chalmers Univ. of Technology, Gothenberg, 1984.

[6] U. KILGORE, *Developable hull surfaces*, in Fishing Boats of the World, Vol. 3, Fishing News Books, London, 1967, pp. 425-431.

[7] E. KREYSZIG, *Differential Geometry*, Univ. of Toronto Press, Toronto, Ontario, Canada, 1959.

[8] L. J. LEON, *A numerical procedure for the isometric mapping of a developable surface onto a subset of the plane*, M.Sc. thesis, Dalhousie University, Halifax, Nova Scotia, Canada, 1982.

[9] T. J. NOLAN, *Computer-aided design of developable surfaces*, Marine Tech., 8 (1971), pp. 233-242.

[10] SCAHD' 77, *Proceedings of the First International Symposium on Computer-Aided Hull Surface Definition*, SNAME, Annapolis, MD, Sept. 26-27, 1977.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

008

ON THE DYNAMIC SHEAR FLOW PROBLEM FOR VISCOELASTIC LIQUIDS*

HANS ENGLER†

Abstract. Initial-boundary value problems for a third order nonlinear integro-differential equation describing dynamic simple shear flow for viscoelastic liquids are studied on bounded one-dimensional spatial domains. Local and global existence results for arbitrary forces and initial data are given under suitable assumptions on the constitutive relations. Conditions on the forces and on the constitutive equations are formulated that imply that solutions of the equations tend to a rest state, and the convergence rates are estimated in terms of the force decay and of dissipation rates that can be derived from the constitutive equations.

Key words. parabolic integro-differential equations, non-Newtonian liquids, simple shear flow, energy estimates, asymptotic behavior

AMS(MOS) subject classifications. Primary 45K05, 45G10, 76A10; secondary 35B40, 35K55

Introduction. The purpose of this note is a study of unsteady simple shear flow for a class of non-Newtonian (viscoelastic) liquids. Using a general class of constitutive equations, the equations of motion reduce to a single third order partial integro-differential equation for the displacement u:

$$(0.1)\qquad u_{tt}(x,t)-\eta\cdot u_{xxt}(x,t)=\int_0^\infty (g(s,u_x(x,t)-u_x(x,t-s))_x\,ds+f(x,t)).$$

Here, $0\leqq x\leqq 1$, $t>0$, $\eta>0$, and g is function characterizing certain properties of the liquid. Details of the derivation of this equation are given in § 5.

To solve (0.1), one has to prescribe the initial history of the flow (i.e. u for $t\leqq 0$) and the initial velocity (i.e. $u_t(x,0)$). At the boundary, the displacement is prescribed ($u(j,t)=f_j(t)$), or traction forces act on the liquid ($\eta\cdot u_{xt}(j,t)+\int_0^\infty g(s,u_x(j,t)-u_x(j,t-s))\,ds=h_j(t)$) for $j=0,1$.

The plan and contents of this paper are as follows: in § 1, we give conditions under which (0.1), together with displacement or traction type boundary conditions, has local (in time) unique solutions in various regularity classes. We show that certain a priori estimates on the solution will guarantee that it can be continued for all times and comment on the relations between regularity properties of the solution and of the data.

In § 2, conditions on the constitutive function g are given under which the boundary displacement problem has a global classical solution for any choice of smooth data. If, e.g., $g(s,u)=a(s)\cdot g_0(u)$, then global existence will follow if $a(\cdot)$ is integrable over $\mathbb{R}^+$ and nonincreasing, and if g_0 is nondecreasing (up to an affine function in u). These conditions reflect a "fading memory" assumption and imply that a certain natural functional for the potential energy has suitable positivity properties and satisfies a dissipation inequality. We then use a transformation introduced in [1] to apply pointwise comparison arguments for scalar integro-differential equations and deduce a priori estimates that imply global existence of solutions. In § 3, the same argument is carried out for the boundary traction problem. In this case, global existence can be shown under slightly weaker conditions.

* Received by the editors August 19, 1985; accepted for publication (in revised form) May 13, 1986.

† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

In § 4, we use a "weighted" energy estimate, similar to an argument used in [8], to show that first time derivatives of the solution decay at essentially the same rate as the forces, provided this rate does not exceed the natural decay rate for Newtonian liquids with the same dynamic viscosity η in shear flow or the decay rate of the kernel $a(\cdot)$.

In the last section, we show in some more detail how an equation of type (0.1) can be derived from rheological models and interpret the results. For the rheological background of this paper, we refer to the monographs [3] and [15]. A general local existence result for unsteady flows of weakly non-Newtonian liquids is given in [18]. Global solutions for extensional flows and small forces are studied in [20] in the case of a special constitutive equation. There is also a relation between the equations studied here and those describing longitudinal motions of one-dimensional viscoelastic bars (cf. [1], [10], [21]). Abstract parabolic integro-differential equations related to those discussed below have been studied by various authors; see, e.g., [12] and its bibliography. Finally, model equations for other one-dimensional motions of visco-elastic materials have been studied in [6], [13], [16].

Throughout this paper, we use the usual notation for Sobolev and Hölder spaces. Single bars $|\cdot|$ denote absolute values or matrix norms; double bars $\|\cdot\|$ indicate function norms. In the proofs, we shall sometimes use the same letter C for various constants that may change from line to line.

1. Local existence of solutions. In this section, we want to give conditions under which the integro-differential equation

$$(1.1)\qquad u_{tt}(x,t)-u_{xxt}(x,t)=\int_0^\infty g(s,u_x(x,t)-u_x(x,t-s))_x\,ds+f(x,t)$$

with initial conditions

$$(1.2)\qquad u_t(x,0)=u_1(x),\ u(x,t)=u_0(x,t)\quad\text{for } t\leqq 0$$

and boundary conditions

$$(1.3)\qquad u(j,t)=f_j(t),\quad\text{or}$$

$$(1.4)\qquad u_{xt}(j,t)+\int_0^\infty g(s,u_x(j,t)-u_x(j,t-s))\,ds=h_j(t)$$

has a unique local (in time) smooth solution. Here, $x\in I=[0,1]$ and $j\in\{0,1\}$. We write (1.1) resp. (1.4) in the abstract forms

$$(1.5)\qquad u_{tt}(x,t)-u_{xxt}(x,t)=\mathbf{F}(u)(x,t),\quad\text{respectively,}$$

$$(1.6)\qquad u_{xt}(j,t)=\mathbf{G}_j(u)(x,t),$$

assuming that $\mathbf{F}$ and $\mathbf{G}_j$ are defined for smooth functions $u: I\times(-\infty,T]\to\mathbb{R}$ for any $T<\infty$ and are of "Volterra-type," that is, for any $0\leqq t\leqq T$

$$(1.7)\quad u=v \text{ on } I\times(-\infty,t]\Rightarrow \mathbf{F}(u)=\mathbf{F}(v)\text{ on } I\times[0,t],\ \mathbf{G}_j(u)=\mathbf{G}_j(v)\text{ on }[0,t].$$

Also, since the unknown function u is given on $I\times(-\infty,0]$ and since it should be continuous across $t=0$, one can incorporate u_0 into the definition of $\mathbf{F}$ and $\mathbf{G}_j$ and assume that these operators are actually defined on a suitable subspace of $C(I\times[0,T],\mathbb{R})$. The modified initial conditions for the problems (1.3), (1.5) resp. (1.5), (1.6) will then be

$$(1.8)\qquad u_t(x,0)=u_1(x),\qquad u(x,0)=u_0(x)\quad\text{for } x\in I.$$

One has corresponding linear boundary value problems which have the form (after integrating once with respect to t)

$$\begin{aligned} &w_t - w_{xx} = \varphi \quad \text{on } I \times [0, T], \\ &w(j, t) = \psi_j(t) \quad (j = 0, 1), \quad w(x, 0) = w_0(x), \end{aligned} \tag{1.9}$$

$$\begin{aligned} &w_t - w_{xx} = \varphi \quad \text{on } I \times [0, T], \\ &w_x(j, t) = \eta_j(t) \quad (j = 0, 1), \quad w(x, 0) = w_0(x). \end{aligned} \tag{1.10}$$

As is well known, these problems have unique solutions $w = \mathbf{L}_1(\varphi, \psi_j, w_0)$ (for (1.9)) and $w = \mathbf{L}_2(\varphi, \eta_j, w_0)$ (for (1.10)) in various classes of differentiable or integrable functions; $\mathbf{L}_1$ and $\mathbf{L}_2$ are in fact Volterra integral operators with weakly singular kernels (see [9], [14]). Therefore, (1.3), (1.5), (1.8), resp. (1.5), (1.6), (1.8), can be rewritten as integral equations. We shall first look for solutions of these integral equations and then show that in the situation studied here these "mild" solutions are in fact smooth functions, satisfying the original integro-differential equations (1.1)-(1.4). We abbreviate $C^{i,j}(T) = C^i([0, T], C^j(I, \mathbb{R}))$ and $C^i(T) = C^i([0, T], \mathbb{R})$ for various integer and noninteger values of i and j and write $\|u\|_{C^{i,j}(T)}$ etc. for the corresponding norms.

The main assumptions on $\mathbf{F}$ and $\mathbf{G}_j$ are

(1.11) $\mathbf{F} : C^{0,2}(T) \to C^{0,0}(T)$ is uniformly Lipschitz-continuous on any bounded set in $C^{0,2}(T)$,

(1.12) $\mathbf{G}_j : C^{0,2}(T) \to C^0(T)$ is uniformly Lipschitz-continuous on any bounded set in $C^{0,2}(T)$.

THEOREM 1.1. *Let $\mathbf{F}$, $\mathbf{G}_0$, $\mathbf{G}_1$ be Volterra-type operators with the Lipschitz properties* (1.11), (1.12), *and let u_0, $u_1 : I \to \mathbb{R}$, $f_0, f_1 : [0, T] \to \mathbb{R}$ be given functions.*

(a) *If $u_0 \in C^2(I)$, $u_1 \in C(I)$, $f_j' \in C^\alpha(T)$ for some $0 < \alpha < 1$, and*

$$u_0(j) = f_j(0), \qquad u_1(j) = f_j'(0) \quad \text{for } j = 0, 1, \tag{1.13}$$

then there exists a number $t_0 > 0$ and a unique solution u of

$$u_t(x, t) - u_{xx}(x, t) = \int_0^t \mathbf{F}(u)(x, s)\, ds + u_1(x) - u_0''(x) \tag{1.14}$$

on $I \times [0, t_0]$ that attains the initial and boundary values (1.3), (1.8) *and for which u_t and u_{xx} are still continuous on $I \times [0, t_0]$.*

(b) *If $u_0 \in C^2(I)$, $u_1 \in C(I)$, then there exists a unique solution u of* (1.14) *on some $I \times [0, t_0]$, satisfying* (1.8) *and the boundary conditions*

$$u_x(j, t) = u_0'(j) + \int_0^t \mathbf{G}_j(u)(s)\, ds, \qquad j = 0, 1 \tag{1.15}$$

and for which u_t and u_{xx} are continuous on $I \times [0, t_0]$.

The proof will be given by means of a standard fixed point argument. We first prepare a few well-known results concerning the solvability of the corresponding linear problems.

LEMMA 1.2. *Let $\mathbf{L}_1$ resp. $\mathbf{L}_2$ be the solution operators of the linear initial boundary value problems* (1.9) *resp.* (1.10).

(a) *If $\varphi \in C^{\alpha,0}(T)$, $0 < \alpha < 1$, $\varphi(j, 0) = 0$, then $\|\mathbf{L}_1(\varphi, 0, 0)\|_{C^{0,2}(T)} \leqq C_1 \cdot \|\varphi\|_{C^{\alpha,0}(T)}$ and $[\mathbf{L}_1(\varphi, 0, 0)]_{xx} \in C^{\beta,0}(T)$ for any $\beta < \alpha$.*

(b) *If $\psi_j' \in C^\alpha(T)$, $\psi_j(0) = \psi_j'(0) = 0$, then for $w = \mathbf{L}_1(0, \psi_j, 0)$, w_t and w_{xx} are Hölder-continuous in $I \times [0, T]$ (with exponents α with respect to t,* min $(2\alpha, 1)$ *with respect to x).*

(c) *If* $w_0\in C^2(I)$, $w_0(j)=w_0''(j)=0$, *then for* $w=\mathbf{L}_1(0,0,w_0)$, w_t *and* w_{xx} *are continuous on* $I\times[0,T]$ (*and analytic for* $t>0$).

(d) *If for some* $\alpha>0$, $\varphi\in C^{\alpha,0}(T)$, $\varphi(j,0)=0$, *then* $[\mathbf{L}_2(\varphi,0,0)]_{xx}\in C^{\gamma,0}(T)$ *for any* $\gamma<\alpha$, *and* $\|\mathbf{L}_2(\varphi,0,0)\|_{C^{0,2}(T)}\leqq C_2\cdot\|\varphi\|_{C^{\alpha,0}(T)}$.

(e) *If for some* $\beta>\frac{1}{2}, \eta_j\in C^{\beta}(T)$ *and* $\eta_j(0)=0$, *then* $[\mathbf{L}_2(0,\eta_j,0)]_{xx}\in C^{\gamma,0}(T)$ *for* $\gamma=2\beta-1$ *and* $\|\mathbf{L}_2(0,\eta_j,0)\|_{C^{0,2}(T)}\leqq C_3\cdot(\|\eta_0\|_{C^{\beta}(T)}+\|\eta_1\|_{C^{\beta}(T)})$.

(f) *If* $w_0\in C^2(I)$, $w_0'(j)=0$, *then for* $w=\mathbf{L}_2(0,0,w_0)$, w_t *and* w_{xx} *are continuous on* $I\times[0,T]$ (*and analytic for* $t>0$).

Here, the C_i depend only on T, α, and β.

These regularity results are neither optimal nor exhaustive, but they suffice for our purposes. Parts (a) and (d) follow from the representation formulae for inhomogeneous heat equations,

$$\mathbf{L}_j(\varphi,0,0)(x,t)=\int_0^t\int_0^1 H_j(x,y,t-s)\varphi(y,s)\,dy\,ds \quad \text{with}$$

$$H_1(x,y,t)=(4\pi t)^{-1/2}\sum_{-\infty}^{\infty}(\exp(-(x-y+2k)^2/4t)-\exp(-(x+y+2k)^2/4t))$$

and $H_2(x,y,t)=\int_x^1 H_{1y}(\zeta,y,t)\,d\zeta+1$ (see [9]), and from the facts that the solution semigroups $v\to S_i(t)(v)$, $S_i(t)(v)(x)=\int_0^1 H_i(x,y,t)v(y)\,dy$, map $C(I)$ into $C^2(I)$ with norms that can be estimated by $C\cdot t^{-1}$, as can be checked directly (see [1]).

The estimates in (a) and (d) then follow in a standard fashion, first for smooth φ and then by approximation in the general case (this is where the compatibility condition in (a) is needed). Parts (b) and (e) are special cases of the general regularity theory for parabolic equations ([9], [14]). Part (c) and (f) again follow from the facts that $\mathbf{L}_i(0,0,w_0)(x,t)=S_i(t)(w_0)(x)$ and that the S_i are suitable analytic semigroups.

Proof of Theorem 1.1. We abbreviate $\mathbf{JF}(u)(x,t)=\int_0^t \mathbf{F}(u)(x,s)\,ds$. To prove part (a), we look for a solution of (1.14) with the corresponding boundary conditions, i.e., for a solution u of

$$\begin{aligned}u&=\mathbf{L}_1(u_1-u_0''+\mathbf{JF}(u),f_j,u_0)\\&=\mathbf{L}_1(u_1-u_0'',f_j,u_0)+\mathbf{L}_1(\mathbf{JF}(u),0,0)=v_1+\mathbf{L}_1(\mathbf{JF}(u),0,0).\end{aligned}\tag{1.16}$$

Since $v_1(x,t)=u_0(x)+t\cdot u_2(x)+\mathbf{L}_1(u_1-u_2,0,0)+\mathbf{L}_1(0,g_j,0)$, with $u_2(x)=(1-x)u_1(0)+xu_1(1)$, $g_j(t)=f_j(t)-u_0(j)-t\cdot u_1(j)$, Lemma 1.2 shows that $v_1\in C^{0,2}(T)$, and we rewrite (1.16) with $v=u-v_1$ as

$$v=\mathbf{L}_1(\mathbf{JF}(v+v_1),0,0).\tag{1.17}$$

Fix $\gamma\in(0,1)$, let $M=\|v_1\|_{C^{0,2}(T)}$ and let K be the Lipschitz constant for $\mathbf{F}$ on the set $B_0=\{w\in C^{0,2}(T)\,|\,\|w\|_{C^{0,2}(T)}\leqq 2M\}$. Let $t_0>0$ be so small that

$$C_1\cdot t_0^{1-\gamma}(\|\mathbf{L}_1(\mathbf{F}(v_1),0,0)\|_{C^{0,0}(T)}+2M)\leqq M,\tag{1.18}$$

and

$$C_1\cdot t_0^{1-\gamma}K\leqq\tfrac{1}{2},\tag{1.19}$$

where $C_1=C_1(\gamma,T)$ is as in Lemma 1.2(a). Then for $w_1,w_2\in B_0$

$$\begin{aligned}\|\mathbf{L}_1(\mathbf{JF}(w_1),0,0)-\mathbf{L}_1(\mathbf{JF}(w_2),0,0)\|_{C^{0,2}(t_0)}&\leqq C_1\cdot\|J(\mathbf{F}(w_1)-\mathbf{F}(w_2))\|_{C^{\beta,0}(t_0)}\\&\leqq C_1\cdot t_0^{1-\gamma}\|\mathbf{F}(w_1)-\mathbf{F}(w_2)\|_{C^{0,0}(t_0)}\\&\leqq C_1\cdot t_0^{1-\gamma}\cdot K\cdot\|w_1-w_2\|_{C^{0,2}(t_0)}.\end{aligned}\tag{1.20}$$

Using (1.18) and (1.19), it follows that the operator $v \to \mathbf{L}_1(\mathbf{JF}(v_1+v), 0, 0)$ is a contraction from the complete metric space $B_1 = \{v \in C^{0,2}(t_0) \mid \|v\|_{C^{0,2}(t_0)} \leqq 2M\}$ into itself. Therefore it has a unique fixed point v. By Lemma 1.2(a), v_t and v_{xx} are also continuous; therefore $u = v + v_1$ is the desired solution of (1.14) with the corresponding boundary conditions. To prove part (b), we look for a solution of

$$u = \mathbf{L}_2(u_1 - u_0'' + \mathbf{JF}(u), u_0'(j) + \mathbf{JG}_j(u), u_0)$$
$$= \mathbf{L}_2(u_1 - u_0'', u_0'(j), u_0) + \mathbf{L}_2(\mathbf{JF}(u), 0, 0) + \mathbf{L}_2(0, \mathbf{JG}_j(u), 0)$$

and apply a similar argument, noting that $\mathbf{L}_2(u_1 - u_0'', u_0'(j), u_0)$ will again have continuous first time and second space derivatives in $I \times [0, T]$ by Lemma 1.2(d)-(f). □

It follows from the proofs that the solutions of both problems will exist on all $I \times [0, T]$ as soon as an a priori estimate in $C^{0,2}(T)$ is known, since the existence interval $[0, t_0]$ can then be fixed a priori and the solutions can successively be continued on $[t_0, 2t_0]$, $[2t_0, 3t_0]$, etc.

As a consequence, we obtain an existence-uniqueness theorem for the original integro-differential equations. Here, the general assumptions will be

(1.21) $g: (0, \infty) \times \mathbb{R} \to \mathbb{R}$, g_t and g_u are jointly continuous on $(0, \infty) \times \mathbb{R}$,

(1.22) $g_u(\cdot, 0) \in L^1(0, \infty; \mathbb{R})$; if $|v|, |w| \leqq R$, then,

$|g_u(t, v) - g_u(t, w)| \leqq a_R(t) \cdot |v - w|$ with some $a_R(\cdot) \in L^1(0, \infty; \mathbb{R})$,

(1.23) u_0 and $u_{0,xx}$ are bounded and continuous on $I \times (-\infty, 0]$.

THEOREM 1.3. (a) *If $u_1 \in C(I)$, $f_j' \in C^\alpha(T)$ for some $0 < \alpha < 1$, $f \in C^{0,0}(T)$, and if* (1.13) *holds, then there are a $t_0 > 0$ and a unique classical solution u on $I \times [0, t_0]$ of*

$$u_t(x, t) - u_{xx}(x, t) = u_1(x) - u_0''(x)$$
$$(1.24) \qquad + \int_0^t \left\{ \int_0^\infty g(s - \tau, u_x(x, s) - u_x(x, s - \tau))_x \, d\tau + f(x, s) \right\} ds$$

that satisfies the initial and boundary conditions (1.2), (1.3).

(b) *If $u_1 \in C(I)$, $h_j \in C^0(T)$, then there are $t_0 > 0$ and a unique classical solution u of* (1.24) *on $I \times [0, t_0]$ that satisfies the initial condition* (1.3) *and the boundary condition*

$$(1.25) \qquad u_x(j, t) + \int_0^t \int_0^\infty g(s - \tau, u_x(j, s) - u_x(j, s - \tau)) \, d\tau \, ds = u_0'(j) + \int_0^t h_j(s) \, ds.$$

Proof. For $v \in C^{0,2}(T)$, define $\mathbf{F}(v)(x, t) = \int_0^\infty g(s - \tau, v_x(x, t) - v_x(x, t - \tau))_x \, d\tau + f(x, t)$, and $v(x, t) = u_0(x, t)$ for $t < 0$; $v(x, t) = v(x, t) - v(x, 0) + u_0(x, 0)$ for $t \geqq 0$. One easily checks that all conditions in Theorem 1.1(a) are met, and part (a) follows. Part (b) is proved similarly by defining suitable operators $\mathbf{G}_j$. □

One can use the quasilinear structure of (1.1) to deduce a $C^{0,2}(T)$-bound for the solution from a $C^{0,1}(T)$-bound.

COROLLARY 1.4. *If in* (1.22) *$a_R(\cdot)$ can be chosen independently from R, then the solution found in Theorem* 1.3 *can be continued on all $I \times [0, T]$. In particular, if the derivative of a solution u_x can be a priori bounded uniformly on any existence set $I \times [0, t_0]$, then u exists as a solution on $I \times [0, T]$.*

Proof. Carrying out the differentiation, one sees that (1.24) can be written as

$$(1.26) \qquad u_t(x, t) - u_{xx}(x, t) = \int_0^t b(x, t, s) \cdot u_{xx}(x, s) \, ds + f_2(x, t)$$

with a suitable $f_2 \in C^{1,0}(T^*)$ and certain continuous coefficients b, both depending on the data u_0 and on the solution, where $I \times [0, T^*]$ is any existence set. Under the assumptions given above, both b and f_2 will be uniformly bounded, with bounds depending only on the data. Abbreviating the integro-differential operator on the right-hand side of (1.24) by **K**, we then have for all $1 < p < \infty$ and all $t > 0$ with $\|u\|_{L^p(t)} = \|u\|_{L^p(I \times [0,t])}$

$$\|\mathbf{K}(u)\|_{L^p(t)} \leqq C(p) \cdot \int_0^t \|u_{xx}\|_{L^p(s)}\, ds. \tag{1.27}$$

By standard regularity results for parabolic equations in L^p-spaces [14], it follows that for the solution of (1.2), (1.3), (1.24)

$$\|u_{xx}\|_{L^p(t)} + \|u_t\|_{L^p(t)} \leqq C \cdot \left(1 + \int_0^t \|u_{xx}\|_{L^p(s)}\, ds\right).$$

Gronwall's lemma implies bounds $\|u_{xx}\|_{L^p(T^*)} + \|u_t\|_{L^p(T^*)} \leqq C(p)$ on $I \times [0, T^*]$ and by standard imbedding theorems [14], u_x is a priori bounded in $C^{\beta,0}(T^*) \cap C^{0,2\beta}(T^*)$ for any $0 < \beta < \frac{1}{2}$. This implies that in (1.26) the derivatives b_t are also uniformly Hölder-continuous in x and t and that f_2 is Hölder-continuous in t, with exponents and bounds depending only on the data. Therefore, we have estimates

$$\|\mathbf{K}(u)\|_{C^{1,0}(t)} \leqq C \cdot \|u_{xx}\|_{C^{0,0}(t)}, \tag{1.28}$$

$$\|\mathbf{K}(u)\|_{C^{0,0}(t)} \leqq C \cdot \int_0^t \|u_{xx}\|_{C^{0,0}(s)}\, ds. \tag{1.29}$$

Hence by interpolation

$$\|\mathbf{K}(u)\|_{C^{\alpha,0}(t)} \leqq \varepsilon \cdot \|u_{xx}\|_{C^{0,0}(t)} + C(\varepsilon) \cdot \int_0^t \|u_{xx}\|_{C^{0,0}(s)}\, ds \tag{1.30}$$

for any $\varepsilon > 0$, $0 < \alpha < 1$, with some $C(\varepsilon) > 0$. Thus by Lemma 1.2(a)-(c)

$$\|u_{xx}\|_{C^{0,0}(t)} \leqq C^* \cdot \left(\varepsilon \cdot \|u_{xx}\|_{C^{0,0}(t)} + C(\varepsilon) \cdot \int_0^t \|u_{xx}\|_{C^{0,0}(s)}\, ds + 1\right) \tag{1.31}$$

with C^* a universal constant. Choosing $C^* \cdot \varepsilon < \frac{1}{2}$ and using Gronwall's lemma, an a priori estimate for u in $C^{0,2}(T^*)$ follows, which implies that the solution exists on $I \times [0, T]$.

A similar argument applies in case (b). In addition to (1.27), we use here that the boundary operators $J\mathbf{G}_j$ satisfy sublinear estimates as mappings from $L^p(0, T; W^{2,p}(I)) \cap W^{1,p}([0, T], L^p(I))$ into $C^0(T)$, if $p > 3$ (cf. [14]), using first L^p-theory and then the regularity estimates of Lemma 1.2(d)-(f).

Finally, if u_x is a priori uniformly bounded on any existence interval by, say, $R > 0$, then $g(t, v)$ can be made constant for $|v| > R + 1$ without changing the equation satisfied by u. Therefore, we can choose $a(\cdot) = a_{R+1}(\cdot)$ in (1.22) and the previous arguments imply a $C^{0,2}$-bound and the continuability for the solution. □

We conclude this section with some remarks concerning the question under which additional assumptions the solutions found in Theorem 1.3 will actually satisfy (1.1) and (1.4).

First, since the right-hand side of (1.1) is always bounded for the solutions found above, (1.1) will hold a.e. (in fact, in any $L^p_{\text{loc}}((0, 1) \times (\varepsilon, t_0))$, $p < \infty$, $\varepsilon > 0$). Also, if in case (a) the f_j are in $C^2(T)$, then u_{tt} and u_{xxt} will still be in any $L^p((0, 1) \times (\varepsilon, t_0))$, $\varepsilon > 0$. For such a result to be true in case (b), one will need that $h_j' \in C^1(T)$, and one can either assume more regularity for the data u_0 (e.g., $u_{0x}(j,\cdot)\in C^{1/2}((-\infty,0])$ will suffice), or more t-regularity for g, e.g.,

$$\int_0^h |g(s,v)|\,ds \leqq C_R\cdot h^\gamma, \tag{1.32}$$

$$|g(t+h,v)-g(t,v)|\leqq b_R(t)\cdot h^\gamma \quad \text{if } |v|\leqq R, \tag{1.33}$$

for any small h, with some $\gamma\geqq\frac{1}{2}$.

To obtain classical solutions of (1.1) on all $I\times[0,t_0]$ in case (a), it will be sufficient to assume that also $u_1''\in C(I)$, $f\in C^{\alpha,0}(T)$, $f_j''\in C^\alpha(T)$, and that either $u_{0,xx}\in C^\alpha((-\infty,0],C(I))$ or that (1.33) and (1.32) hold also for g_u instead of g with some $\gamma>0$. Also, the natural compatibility condition $f_j''(0)-u_1''(j)=\int_0^\infty g(s,u_{0x}(j,0)-u_{0x}(j,-s))_x\,ds+f(j,0)$ will be needed for $j=0,1$. Similarly, in case (b) we will have a classical solution of (1.1), (1.4), if $u_1''\in C(I)$, $f\in C^{\alpha,0}(T)$, $h_j\in C^\gamma(T)$ with $\gamma>\frac{1}{2}$, and if corresponding conditions hold that guarantee that the right-hand side of (1.1) is in $C^{\alpha,0}(T)$ and the right-hand side of (1.4) is in $C^\gamma(T)$ with some $\gamma>\frac{1}{2}$. We omit the details, which are similar to the conditions in case (a). All arguments leading to these conditions are straightforward applications of the regularity theory for parabolic equations.

2. Global existence for the boundary displacement problem. In this section conditions are given under which problem (1.1)–(1.3) has global solutions for arbitrary forces f, initial histories u_0 and boundary displacements f_j. We assume without loss of generality that $g(t,0)=0$ and define

$$G(t,u)=\int_0^u g(t,v)\,dv \quad \text{for } t>0, \qquad u\in\mathbb{R}. \tag{2.1}$$

We shall obtain pointwise a priori estimates on the displacement gradient u_x, based on

(i) a natural energy estimate, using the variational structure of the integral operator in (1.1);

(ii) a comparison argument for scalar integro-differential equations, using a technique from [1].

We always assume in this and the following sections that $t\cdot G_t(t,u)$ is still integrable for $0<t\leqq 1$; more precisely, it will be assumed that for all $t>0$, $|u|\leqq R$, $|G_t(t,u)\cdot\min(1,t)|\leqq d_R(t)$, with $d_R(\cdot)\in L^1(0,\infty)$. Also, the L^2-norm on the interval I will be abbreviated by $\|\cdot\|$, and we write $\langle\cdot,\cdot\rangle$ for the inner product on $L^2(I)$.

To carry out the first step, the key assumption on the integral operator is the following.

(H1) There exist $C_0\geqq 0$, $C_1\in L^1(0,\infty;\mathbb{R}^+)$ such that for all $t>0$, $u\in\mathbb{R}$,

$$G(t,u)\geqq -C_1(t)\cdot(u^2+1), \qquad G_t(t,u)\leqq C_0\cdot(G(t,u)+C_1(t)\cdot(u^2+1)), \tag{2.2}$$

$$|g(t,u)|\leqq C_0\cdot(G(t,u)+C_1(t)\cdot(u^2+1)). \tag{2.3}$$

LEMMA 2.1. *Let g satisfy* (H1) *and the assumptions from Theorem* 1.3, *let u_0, u_1 and f be as in Theorem* 1.3 *and let $f_1,f_2\in C^2(T)$. Then there exists a constant $C^*>0$, depending only on the data, T, C_0, and C_1, such that for any solution u of the integrated equations* (1.2), (1.3), (1.24) *on $I\times[0,T_0]$ and for any $0\leqq t\leqq T_0$*

$$\|u_t(\cdot,t)\|^2+\int_0^\infty\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds+\int_0^t \|u_{xt}(\cdot,s)\|^2\,ds\leqq C^*. \tag{2.4}$$

Proof. We can assume that $f_j(t)=w(j,t)$ and w_x is x-independent. Define

$$E(t)=\|u_t(\cdot,t)-w_t(\cdot,t)\|^2/2+\int_0^\infty\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds, \tag{2.5}$$

$$E_1(t)=E(t)+\int_0^\infty C_1(s)\{\|u_x(\cdot,t)-u_x(\cdot,t-s)\|^2+1\}\,ds \tag{2.6}$$

for $0\leqq t\leqq T_0$. Note that $E_1(t)\geqq 0$ by (2.2).

The remarks following Corollary 1.4 show that (1.1) will hold in any $L^p(I\times(\varepsilon,T_0))$, $\varepsilon>0$, p finite. Consequently, u_{xt} is still continuous in any $I\times(\varepsilon,T_0]$, and by the integrability assumption for $t\cdot G_t(t,u)$, $E(\cdot)$ is absolutely continuous on any $(\varepsilon,T_0]$. Differentiating (2.5) then gives, for a.e. t

$$\begin{aligned} d/dt\,E(t)&=\langle u_t(\cdot,t)-w_t(\cdot,t),u_{tt}(\cdot,t)-w_{tt}(\cdot,t)\rangle\\ &\quad+\int_0^\infty\int_I G_s(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds\\ &\quad+\int_0^\infty\langle g(s,u_x(\cdot,t)-u_x(\cdot,t-s)),u_{xt}(\cdot,t)\rangle\,ds\\ &\leqq\langle u_t(\cdot,t)-w_t(\cdot,t),u_{tt}(\cdot,t)\rangle+C\cdot(E_1(t)+1) \qquad (2.7)\\ &\quad+\int_0^\infty\langle g(s,u_x(\cdot,t)-u_x(\cdot,t-s)),u_{xt}(\cdot,t)-w_{xt}(\cdot,t)\rangle\,ds\\ &=\left\langle u_t(\cdot,t)-w_t(\cdot,t),u_{tt}(\cdot,t)-\int_0^\infty g(s,u_x(\cdot,t)-u_x(\cdot,t-s))_x\,ds\right\rangle\\ &\quad+C\cdot(E_1(t)+1), \end{aligned}$$

using an integration by parts. Inserting (1.1) into the last estimate, integrating the product $\langle u_t-w_t,u_{xxt}\rangle$ by parts and estimating further we obtain

$$\begin{aligned} d/dt\,E(t)&\leqq-\langle u_{xt}(\cdot,t)-w_{xt}(\cdot,t),u_{xt}(\cdot,t)\rangle+C\cdot(E_1(t)+1)\\ &\leqq-\tfrac{1}{2}\|u_{xt}(\cdot,t)\|^2+C\cdot(E_1(t)+1). \end{aligned} \tag{2.8}$$

Integrating (2.8) from ε to t and sending ε to 0, we obtain

$$\begin{aligned} &E_1(t)+\frac{1}{2}\int_0^t\|u_{xt}(\cdot,s)\|^2\,ds\\ &\leqq C\cdot\left(\int_0^t E_1(s)\,ds+1\right)+\int_0^\infty C_1(s)\{\|u_x(\cdot,t)-u_x(\cdot,t-s)\|^2+1\}\,ds \qquad (2.9)\\ &\leqq K\cdot\left(\int_0^t E_1(s)\,ds+\|u_x(x,t)\|^2+1\right)+2\int_0^t C_1(t-s)\cdot\|u_x(\cdot,s)\|^2\,ds \end{aligned}$$

with $K>0$. Now, for any δ and some constant C_δ (depending also on u_0)

$$\|u_x(\cdot,t)\|^2\leqq\delta\cdot\int_0^t\|u_{xs}(\cdot,s)\|^2\,ds+C_\delta\cdot\left(\int_0^t\int_0^s\|u_{xs}(\cdot,\tau)\|^2\,d\tau\,ds+1\right).$$

Pick δ such that $K\cdot\delta=\frac{1}{4}$; then (2.9) implies an estimate of the form

$$\begin{aligned}&E_1(t)+\frac{1}{4}\int_0^t \|u_{xt}(\cdot,s)\|^2\,ds\\ &\quad\leqq C\cdot\left(1+\int_0^t(1+C_1(t-s))\cdot\left(E_1(s)+\int_0^s\|u_{xt}(\cdot,\tau)\|^2\,d\tau\right)ds\right).\end{aligned}\tag{2.10}$$

This implies a bound for $E_1(t)$, and (2.4) follows. □

We now deduce a uniform a priori estimate for u_x, if additionally

(H2) $g(t,u)=g_0(t,u)+L(t)u$, with $L(\cdot)\in L^1(0,\infty;\mathbb{R})$ and g_0 nondecreasing.

THEOREM 2.2. *Assume that* u_0, u_1, *and* f *are as in Theorem* 1.3, *that* $f_0,f_1\in C^2(T)$, *and that* g *satisfies the assumptions of Theorem* 1.3 *as well as* (H1) *and* (H2). *Let* u *be a solution of* (1.2), (1.3), (1.24) *on some set* $I\times[0,T_0]$. *Then there exists a constant* C^*, *depending only on the data and on* T, *but not on* T_0, *such that*

$$|u_x(x,t)|\leqq C^*\quad\text{on } I\times[0,T_0],\tag{2.11}$$

and thus u *can be extended as a solution on* $I\times[0,T]$.

Proof. We use a version of a technique introduced in [1]. For fixed (x,t), integrate (1.1) (which holds a.e.) from $y\in[0,1]$ to x and then from 0 to 1 with respect to y. Using the abbreviation

$$p(x,t)=\int_0^1\int_y^x u_t(\zeta,t)\,d\zeta\,dy,\tag{2.12}$$

we get the identity (valid for all (x,t) due to the regularity of u)

$$\begin{aligned}&(u_x(x,t)-p(x,t))_t+\int_0^\infty g(s,u_x(x,t)-u_x(x,t-s))\,ds\\ &\quad=\int_0^\infty\int_I g(s,u_x(y,t)-u_x(y,t-s))\,dy\,ds+f_1(t)-f_0(t)\\ &\qquad-\int_0^1\int_y^x f(\zeta,t)\,d\zeta\,dy=:k(x,t).\end{aligned}\tag{2.13}$$

Now from Lemma 2.1 $|p(x,t)|$ is bounded independently of x and t, and (2.3) and (2.4) imply that also $|k(x,t)|$ is uniformly bounded. Dropping the x-dependence and using (H2), we obtain

$$\begin{aligned}&w'(t)+\int_0^t g_0(t-s,w(t)-w(s)+p(t)-p(s))\,ds+L_0\cdot w(t)\\ &\quad+\int_0^\infty g_0(t+s,w(t)+p(t)-k_0(s))\,ds=k_1(t)+\int_0^\infty L(t-s)w(s)\,ds\end{aligned}\tag{2.14}$$

with $w(t)=u_x(x,t)-p(x,t)$, $L_0=\int_0^\infty L(s)\,ds$, and suitable functions k_0, k_1 that are a priori bounded. Also, $|w(0)|\leqq\|u_{0,x}\|_{C(I)}+\|u_1\|_{C(I)}=M_0$. Assume now that for some minimal $t>0$ and all $0\leqq s<t$, $|w(t)|=(1+M_0)\cdot e^{Mt}$, $|w(s)|<(1+M_0)\cdot e^{Ms}$, where M is some arbitrary constant. If, e.g., $w(t)>0$, then consequently $w'(t)\geqq(1+M_0)\cdot M\cdot e^{Mt}=M\cdot w(t)$, and for $0\leqq s<t$, $w(s)\leqq|w(s)|\leqq e^{M(t-s)}w(t)$. Then (2.14)

implies

$$
\begin{aligned}
(1+M_0)\cdot M\cdot e^{Mt} &\leqq |L_0|\cdot(1+M_0)\cdot e^{Mt}+\int_0^t |L(s)|\,e^{-Ms}\,ds\cdot(1+M_0)\cdot e^{Mt}\\
&\quad+|k_1(t)|+\left|\int_0^t g_0(t-s,p(t)-p(s))\,ds\right|\\
&\quad+\left|\int_0^\infty g_0(t+s,p(t)-k_0(s))\,ds\right|\\
&\leqq C+(1+M_0)\cdot e^{Mt}\cdot\left(|L_0|+\int_0^\infty |L(s)|\cdot e^{-Ms}\,ds\right)
\end{aligned}
\tag{2.15}
$$

with some fixed constant C. It is clear that (2.15) can never hold if $M>(C+|L_0|+\int_0^\infty |L(s)|\cdot e^{-Ms}\,ds)$. A similar contradiction can be derived if $w(t)=-(1+M_0)\cdot e^{Mt}$ and M is sufficiently large. Therefore, $|w(t)|\leqq(1+M_0)\cdot e^{Mt}$ for all t, which gives the desired a priori bound for u_x. Corollary 1.4 then implies that the solution u exists on all $I\times[0,T]$. □

In [1] the second order equation $u_{tt}=u_{xxt}+\sigma(u_x)_x$ is treated, and a differential equation (instead of an integro-differential equation) is derived for w, which allows us to find a t-independent a priori bound. It is not clear how to do this in our situation. However, under suitable assumptions (that are satisfied in the setting of § 4), an a priori bound for $|u_x(x,t)|$ can be deduced that grows linearly in t.

COROLLARY 2.3. *Assume that in* (2.4) C^* *does not depend on* T, *that* f, f_0' *and* f_1' *are uniformly bounded, that in* (H2) $L=0$, *and that also* $\|u_x(\cdot,t)\|\leqq C^*$, *independent of* t. *Then there exists a constant* $K>0$, *dependent only on the data, such that for all* $(x,t)\in I\times[0,\infty)$

$$|u_x(x,t)|\leqq K\cdot(1+t).$$

Proof. Using the same arguments as in the proof above, one deduces (2.13) and observes that the assumptions imply that the right-hand side of (2.13) is bounded, independent of x and t. Thus (2.14) follows with $L_0=0=L(\cdot)$ and with uniformly bounded functions k_0, k_1 and p. Denote again by M_0 a common bound for $|w(0)|$; for arbitrary $M>0$, we then assume that for some $t>0$, $|w(t)|=(1+M_0)\cdot(1+M\cdot t)$ and $|w(s)|<(1+M_0)\cdot(1+M\cdot s)$ for $s<t$. An inequality similar to (2.15) follows which leads to a contradiction, if M is too big. This proves the corollary. □

3. Global solutions for the boundary traction problem. In this section, conditions are given under which the problem (1.1), (1.2), (1.4) has global solutions for arbitrary forces f, h_j and for arbitrary initial histories u_0. Again, we assume that $g(t,0)=0$ and define $G(t,u)$ as in (2.1). Instead of (H1), we shall use the weaker assumption

(H1′) There exist $C_0\geqq 0$, $C_1\in L^1(0,\infty;\mathbb{R}^+)$ such that for all $t>0$, $u\in\mathbb{R}$,

$$G(t,u)\geqq -C_1(t)\cdot(u^2+1),\tag{3.1}$$

$$G_t(t,u)\leqq C_0\cdot(G(t,u)+C_1(t)\cdot(u^2+1)).\tag{3.2}$$

We also assume the same integrability properties for $G_t(t,u)$ and the same abbreviations as in the previous section.

THEOREM 3.1. *Let* g *satisfy the assumptions of Theorem* 1.3, (H1′) *and* (H2). *Let* u_0, u_1, f *and* h_j *be as in Theorem* 1.3, $h_j\in C^1(T)$, *and assume that* $u_{0,x}(j,\cdot)\in C^{1/2}((-\infty,0],\mathbb{R})$ *for* $j=0,1$. *Then there exists a solution* u *of the integrated equations* (1.24), (1.25), (1.3) *on all* $I\times[0,T]$.

Proof. We use an energy estimate as in Lemma 2.1 and a comparison argument as in the proof of Theorem 2.2. Define

$$E(t)=\|u_t(\cdot,t)\|^2/2+\int_0^\infty\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds, \tag{3.3}$$

$$E_1(t)=E(t)+\int_0^\infty C_1(s)\cdot(\|u_x(\cdot,t)-u_x(\cdot,t-s)\|^2+1)\,ds. \tag{3.4}$$

By the same arguments as in the proof of Lemma 2.1, $E(\cdot)$ is absolutely continuous on any interval $[\varepsilon, T_0]$, and by differentiating we obtain (cf. (2.7))

$$\begin{aligned} d/dt\,E(t)&\leqq\langle u_t(\cdot,t),u_{tt}(\cdot,t)\rangle+C_0\cdot E_1(t)\\ &\quad+\left\langle u_{xt}(\cdot,t),\int_0^\infty g(s,u_x(\cdot,t)-u_x(\cdot,t-s))\,ds\right\rangle. \end{aligned}$$

Note that the assumptions imply that u_{tt} and u_{xxt} exist in $L^p(I\times(\varepsilon,T_0))$. Integrating the last integral on the right-hand side by parts with respect to x, inserting the equation and integrating by parts again gives

$$\begin{aligned} d/dt\,E(t)&\leqq-\|u_{xt}(\cdot,t)\|^2+\langle u_t(\cdot,t),f(\cdot,t)\rangle\\ &\quad+u_t(1,t)\cdot h_1(t)-u_t(0,t)\cdot h_0(t)+C_0\cdot E_1(t). \end{aligned}$$

Since the boundary power terms can be estimated by, e.g.,

$$|u_t(1,t)\cdot h_1(t)|\leqq\tfrac14\|u_{xt}(\cdot,t)\|^2+C\cdot(1+\|u_t(\cdot,t)\|^2),$$

we obtain an estimate

$$d/dt\,E(t)\leqq-\|u_{xt}(\cdot,t)\|^2/2+C\cdot(E_1(t)+1).$$

The rest of the argument is as in the proof of Lemma 2.1, and therefore

$$\|u_t(\cdot,t)\|^2\leqq C^*(t), \tag{3.5}$$

where the locally bounded function $C^*(\cdot)$ depends only on the data. Integrating now (1.1) (which holds locally in the L^p-sense) from $y=0$ to $y=x$ and abbreviating $q(x,t)=\int_0^x u_t(y,t)\,dy$, we obtain (for all (x,t))

$$\begin{aligned} (u_x(x,t)-q(x,t))_t&+\int_0^\infty g(s,u_x(x,t)-u_x(x,t-s))\,ds\\ &=h_0(t)-\int_0^x f(y,t)\,dy=k(x,t). \end{aligned} \tag{3.6}$$

Since (3.5) implies that q is uniformly bounded, we can use the same argument as in the proof of Theorem 2.2 to infer that $w=u_x-q$ is a priori bounded, which implies the theorem. □

We note that the argument given here still applies in the case where one of the boundary conditions is of traction type and the displacement is prescribed on the other boundary. Obviously a local existence result and the same continuability properties as in § 1 will hold for such a problem. If, e.g., $u(1,t)=f_1(t)$ is prescribed (with traction forces h_0 acting at $x=0$), then one defines the energy

$$E(t)=\|u_t(\cdot,t)-f_1'(t)\|^2/2+\int_0^\infty\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds$$

and uses the same arguments as above to arrive at the estimate (3.5). The comparison argument remains unchanged.

Also, $|u_x(x,t)|$ can be shown to grow linearly in t (uniformly in x) by using an argument that is similar to the one in Corollary 2.3 under the weaker assumptions that in (H2) $L(\cdot)=0$, that h_0 and f are uniformly bounded, and that in (3.5) $C^*(\cdot)$ does not depend on t (cf. the following section). Finally, by examining the proof of Theorem 3.1, one notices that both the estimates (3.5) and the bound on w depend only on the supremum norms of the h_j and of u_0 and $u_{0,x}$. Therefore, we can approximate any smooth h_j and any u_0 which is only continuous together with its second space derivatives by Hölder-continuous data and still obtain uniformly bounded solutions on all $I\times[0,T]$ which will approximate the solution of the original problem (by uniqueness). Consequently, we have the following corollary.

COROLLARY 3.2. *The statement of Theorem* 3.1 *remains true if only* $h_j\in C^0(T)$ *and* u_0 *and* $u_{0,xx}$ *are bounded and continuous on* $I\times(-\infty,0]$.

4. Asymptotic behavior of solutions. In this section we show decay estimates for the derivatives of the solutions of (1.1)-(1.4) which imply their convergence to a steady state u_∞. Instead of (1.1), the equation

$$u_{tt}(x,t)-\eta\cdot u_{xxt}(x,t)=\int_0^\infty g(s,u_x(x,t)-u_x(x,t-s))_x\,ds+f(x,t) \tag{4.1}$$

will be studied with $\eta>0$, and (1.4) will be replaced by

$$\eta\cdot u_{xt}(j,t)+\int_0^\infty g(s,u_x(j,t)-u_x(j,t-s))\,ds=h_j(t). \tag{4.2}$$

Of course, (4.1) and (4.2) are equivalent to (1.1) and (1.4) by rescaling; here we want to display the dependence of the estimates on the "viscosity" η.

The key hypothesis for g in addition to (H1), (H1′), (H2) for both the boundary displacement and the boundary traction problems will be

(H3) $0=G(t,0)\leqq G(t,u)$ for all $u\in\mathbb{R}$, $t>0$;

there is a constant $\delta>0$ such that for all $u\in\mathbb{R}$, $t>0$

$$G_t(t,u)+\delta\cdot G(t,u)\leqq 0. \tag{4.3}$$

THEOREM 4.1. *Let g satisfy the general assumptions of* § 1 *as well as* (H1), (H2), (H3). *Let* u_0, u_1 *and* f *be as in Theorem* 1.3(a) *and assume that* f_0 *and* f_1 *are t-independent. Let* $b:[0,\infty)\to[0,\infty)$ *be continuously differentiable,* $b(0)=1$, *and assume that* $0\leqq b'(t)\leqq\kappa\cdot b(t)$ *for some* $\kappa\leqq\delta$, $\kappa<2\eta\pi^2$. *Then for all* $t>0$ *and for any solution of* (1.2), (1.3), (4.1)

$$\|u_t(\cdot,t)\|\cdot b^{1/2}(t)\leqq C(u_0,u_1)^{1/2}+\int_0^t\|f(\cdot,s)\|\cdot b^{1/2}(s)\,ds, \tag{4.4}$$

$$\begin{aligned}\int_0^t b(s)\cdot\|u_{xt}(\cdot,s)\|^2\,ds&\leqq 2\pi^2(2\eta\pi^2-\kappa)^{-1}\\&\quad\cdot\left\{C(u_0,u_1)+\left(\int_0^t\|f(\cdot,s)\|\cdot b^{1/2}(s)\,ds\right)^2\right\},\end{aligned} \tag{4.5}$$

where $C(u_0,u_1)=\|u_1(\cdot)\|^2+2\cdot\int_0^\infty\int_I G(s,u_{0,x}(x,0)-u_{0,x}(x,-s))\,dx\,ds$.

COROLLARY 4.2. *Under the assumptions of Theorem* 4.1, *and if*

$$\int_0^\infty \|f(\cdot,s)\|\cdot b^{1/2}(s)\,ds<\infty, \tag{4.6}$$

$$\int_0^\infty b^{-1}(s)\,ds<\infty, \tag{4.7}$$

then there exists $u_\infty \in W^{1,2}(I,\mathbb{R})$ *with* $u_\infty(j)=f_j$ $(j=0,1)$ *such that*

$$\|u_x(\cdot,t)-u_{\infty x}(\cdot)\|^2=O\left(\int_t^\infty b^{-1}(s)\,ds\right)\quad \textit{as } t\to\infty. \tag{4.8}$$

Proof of the corollary. For $0<t<s$ we have

$$\begin{aligned}\|u_x(\cdot,t)-u_x(\cdot,s)\|^2&\leqq\left\|\int_t^s u_{xt}(\cdot,\tau)\,d\tau\right\|^2\\ &\leqq\int_t^s \|u_{xt}(\cdot,\tau)\|^2\cdot b(\tau)\,d\tau\cdot\int_t^s b^{-1}(\tau)\,d\tau\leqq C\cdot\int_t^s b^{-1}(\tau)\,d\tau.\end{aligned} \tag{4.9}$$

Then (4.5) and (4.7) imply that $t\to u(\cdot,t)$ has a limit in $W^{1,2}(I,\mathbb{R})$, as $t\to\infty$. Sending $s\to\infty$ in (4.9), the convergence estimate (4.8) follows. □

In the case $g=0$ (where (4.1) reduces to the heat equation), the existence of u_∞ already follows if (4.6) holds with $b(t)=1$; thus the convergence result certainly is not optimal. Also, we do not quite recover the "optimal" convergence rate $\|u(\cdot,t)-u_\infty\|\sim\exp(-\eta\pi^2 t)$ for $g=0$ and $f=0$.

For the case of the boundary traction problem (4.1), (4.2), (1.2) we introduce some additional notation: Let u be a solution. Define the mean displacement $U(t)=\int_I u(x,t)\,dx$ and $u^*(x,t)=u(x,t)-U(t)$. Then U can be computed from the equation; we have

$$U(t)=\int_I u_0(x)\,dx+t\cdot\int_I u_1(x)\,dx+\int_0^t (t-s)(F(s)+h_1(s)-h_0(s))\,ds, \tag{4.10}$$

where $F(t)=\int_I f(x,t)\,dx$ is the mean body force. Also, define $f^*(x,t)=f(x,t)-F(t)$. Then u^* will again satisfy (4.1) and (4.2), with f replaced by f^*, and additionally, $\int_I u^*(x,t)\,dx=0$ for all t.

THEOREM 4.3. *Let g satisfy the assumptions of* § 1, *as well as* (H1′), (H2) *and* (H3). *Let u_0, u_1, f and h_j be as in Theorem* 1.3(b), *and let b be as in Theorem* 4.1. *Then there is a constant $C_1>0$ such that for all $t>0$ and any solution u of* (1.2), (4.1), (4.2)

$$\begin{aligned}&b(t)\cdot\|u_t^*(\cdot,t)\|^2+\int_0^t b(s)\cdot\|u_{xt}(\cdot,s)\|^2\,ds\\ &\leqq C_1\left\{\left(\left(\int_0^t \|f(\cdot,s)\|+|h_0(s)|+|h_1(s)|\right)\cdot b^{1/2}(s)\,ds\right)^2\right.\\ &\qquad\left.+\int_0^t b(s)\cdot(h_0^2(s)+h_1^2(s))\,ds+C(u_0,u_1)\right\}\end{aligned} \tag{4.11}$$

where $C(u_0,u_1)$ is as in Theorem 4.1.

COROLLARY 4.4. *If*

$$\|f(x,\cdot)\|\cdot b^{1/2}(\cdot)\in L^1(0,\infty;\mathbb{R})\quad\textit{and}\quad h_j(\cdot)\cdot b^{1/2}(\cdot)\in L^1\cap L^2(0,\infty;\mathbb{R}), \tag{4.12}$$

and if (4.7) *holds, then there is a* $u_\infty \in W^{1,2}(I, \mathbb{R})$, $\int_I u_\infty(x)\,dx = 0$, *so that*

$$\|u(\cdot, t) - U(t) - u_\infty\|_{W^{1,2}} = O\left(\int_t^\infty b^{-1}(s)\,ds\right) \quad \text{as } t \to \infty. \tag{4.13}$$

The proof is similar to the one of Corollary 4.2.

In both convergence results, we are not able to deduce equations for the rest state u_∞. This reflects the fact that any t-independent function satisfies the equation (4.1), if $f = 0$. The only exception seems to be the case of a linear g, in which (4.1) is equivalent to a convolution equation (with a first time derivative), for which an equation for the rest state can be deduced: If $g(s, v) = a(s) \cdot v$ with $a(t) \geqq 0 \geqq a'(t) + \delta \cdot a(t)$ and $a(\cdot) \in L^1(0, \infty)$ (reflecting (H1) and (H3)), then (4.1) is equivalent to

$$u_t(x, t) - \eta \cdot u_{xx}(x, t) = \int_0^\infty a_0(s) \cdot u_{xx}(x, t - s)\,ds + f_0(x, t) \tag{4.14}$$

with $a_0(t) = \int_t^\infty a(s)\,ds$ and $f_0(\cdot, t) = u_1(\cdot) - \eta \cdot u_{0,xx}(\cdot, 0) - \int_0^\infty a_0(s) u_{0,xx}(\cdot, -s)\,ds + \int_0^t f(\cdot, s)\,ds$. Since (H1) and (H3) imply that $0 \leqq a_{00} = \int_0^\infty a_0(s)\,ds < \infty$, the limit u_∞ then satisfies

$$(\eta + a_{00}) \cdot u_{\infty,xx} + u_1 + \int_0^\infty f(\cdot, s)\,ds = \eta u_{0,xx}(\cdot, 0) + \int_0^\infty a_0(s) u_{0,xx}(\cdot, -s)\,ds. \tag{4.15}$$

This shows that $u_\infty \in W^{2,2}(I)$ in the case of a linear equation and suggests that the same should also be true for solutions of general nonlinear equations; however, it is not clear how to show this. A certain improvement is possible if a sublinear growth bound for $|u_x(x, t)|$ is known.

COROLLARY 4.5. *Under the assumptions of Corollary* 4.2, *and if in* (H2) $L(\cdot) = 0$, *then* $b^{-1}(t) = O(t^{-\alpha})$ *for some* $\alpha > 1$ *implies that*

$$\|u(\cdot, t) - u_\infty(\cdot)\|_{W^{1,p}} = O(t^{-\beta}) \tag{4.16}$$

for any $2 < p < \alpha + 1$, *with* $\beta = (\alpha + 1 - p)/p$. *Similarly,* $b^{-1}(t) = O(e^{-\gamma t})$ *implies*

$$\|u(\cdot, t) - u_\infty(\cdot)\|_{W^{1,p}} = O(e^{-\beta t}) \tag{4.17}$$

for any $2 < p < \infty$ *and any* $\beta < \gamma/p$.

Proof. The assumptions and conclusions of Corollary 4.2 together with the additional assumption $L(\cdot) = 0$ allow us to use Corollary 2.3 to conclude that $|u_x(x, t)| \leqq K \cdot (1 + t)$ for all (x, t), with $K > 0$ depending on the data. Now let $0 < t < s < \infty$, and let $k \geqq 0$ be such that $2^k t \leqq s < 2^{k+1} t$. Then for $2 < p$, writing $\|\cdot\|_p$ for the $L^p(I)$-norm and abbreviating $v = u_x$, we have

$$\begin{aligned}
\|v(\cdot, t) - v(\cdot, s)\|_p &\leqq \sum_{1 \leqq j \leqq k} \|v(\cdot, 2^{j-1}t) - v(\cdot, 2^j t)\|_p + \|v(\cdot, 2^k t) - v(\cdot, s)\|_p \\
&\leqq \sum_{1 \leqq j \leqq k} (\|v(\cdot, 2^{j-1}t) - v(\cdot, 2^j t)\|_\infty)^{1-2/p} (\|v(\cdot, 2^{j-1}t) - v(\cdot, 2^j t)\|_2)^{2/p} \\
&\quad + (\|v(\cdot, 2^k t) - v(\cdot, s)\|_\infty)^{1-2/p} (\|v(\cdot, 2^k t) - v(\cdot, s)\|_2)^{2/p} \\
&\leqq C \cdot \sum_{1 \leqq j \leqq k+1} (2^j t)^{1-2/p} \left(\int_{2^{j-1}t}^{2^j t} b^{-1}(s)\,ds\right)^{1/p},
\end{aligned} \tag{4.18}$$

using Hölder's inequality, the L^∞-estimate for u_x and (4.9). A direct calculation then shows that $u_x(\cdot, t)$ has a limit in $L^p(I)$ with the convergence rates given in (4.16) and (4.17). □

Proof of Theorem 4.1. Recall that under our assumptions, (4.1) still holds in any $L^p(I\times(T_0, T_1))$ for any finite p, T_0, T_1, and that u_{tt} and u_{xxt} are in these L^p-spaces. Also, by Theorem 2.2, u_x is uniformly bounded on any $I\times[0, T]$. Multiply (4.1) by $b(t)\cdot u_t(x, t)$, and integrate over $I\times[0, T]$. After an integration by parts with respect to x, this gives the identity

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}(b(t)\cdot\|u_t(\cdot,t)\|^2)-\frac{1}{2}b'(t)\cdot\|u_t(\cdot,t)\|^2\\ \text{(4.19)}\qquad &+\eta b(t)\cdot\|u_{xt}(\cdot,t)\|^2+\int_0^\infty b(t)\cdot\langle g(s,u_x(\cdot,t)-u_x(\cdot,t-s)),u_{xt}(\cdot,t)\rangle ds\\ &=b(t)\cdot\langle f(\cdot,t),u_t(\cdot,t)\rangle.\end{aligned}$$

We now note that for $t>0$, $t\to\int_0^\infty b(t)\cdot\int_I G(s, u_x(x,t)-u_x(x,t-s))\,dx\,ds$ is absolutely continuous in t (since u_{xt} is still bounded on any $I\times[T_0, T_1]$, $0<T_0<T_1$) and that for a.e. t

$$\begin{aligned}&\frac{d}{dt}\int_0^\infty b(t)\cdot\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds\\ \text{(4.20)}\qquad &\leqq\int_0^\infty b(t)\cdot\langle g(s,u_x(\cdot,t)-u_x(\cdot,t-s)),u_{xt}(\cdot,t)\rangle\,ds\end{aligned}$$

due to (H3), the assumptions on b' and the behavior of G_t near $t=0$. Also, since u_t vanishes at the endpoints of I, by Poincaré's inequality

$$\begin{aligned}&-\tfrac{1}{2}b'(t)\cdot\|u_t(\cdot,t)\|^2+\eta\cdot b(t)\cdot\|u_{xt}(\cdot,t)\|^2\\ \text{(4.21)}\qquad &\geqq b(t)\cdot(\eta\cdot\|u_{xt}(\cdot,t)\|^2-\tfrac{1}{2}\kappa\cdot\|u_t(\cdot,t)\|^2)\\ &\geqq b(t)\cdot(2\pi^2\eta-\kappa)(2\pi^2)^{-1}\cdot\|u_{xt}(\cdot,t)\|^2.\end{aligned}$$

Substituting (4.20) and (4.21) into (4.19), integrating over $[\varepsilon, t]$ and sending ε to 0, we obtain with $c=(2\pi^2\eta-\kappa)(2\pi^2)^{-1}$

$$\begin{aligned}&\frac{1}{2}b(t)\cdot\|u_t(\cdot,t)\|^2+c\cdot\int_0^t b(s)\cdot\|u_{xt}(\cdot,s)\|^2\,ds\\ \text{(4.22)}\qquad &+\int_0^\infty b(t)\cdot\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds\\ &\leqq\frac{1}{2}C(u_0,u_1)+\int_0^t\|f(x,s)\|\cdot b^{1/2}(s)\cdot\|u_t(\cdot,s)\|\cdot b^{1/2}(s)\,ds.\end{aligned}$$

We now drop the second integral on the left-hand side of (4.22) and apply Bihari's inequality [4]; then (4.4) follows. Inserting this estimate into the right-hand side of (4.22) and dropping the first and third integral on the left-hand side gives (4.5). □

Proof of Theorem 4.3. By an approximation argument similar to the one leading to Corollary 3.2, we can assume that the h_j and the data u_0 are so regular that (4.1) holds in any L^p_{loc}, that u_t and u_{xxt} are locally in L^p and that (4.2) holds pointwise for $t>0$. Recall that (4.1) and (4.2) also hold for u^*, with f^* instead of f. Multiply the equation for u^* with $b(t)\cdot u_t^*$, integrate with respect to x and use (4.20) to obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}(b(t)\cdot\|u_t^*(\cdot,t)\|^2)-\frac{1}{2}b'(t)\cdot\|u_t^*(\cdot,t)\|^2\\ &\quad+\eta b(t)\cdot\|u_{xt}(\cdot,t)\|^2+d/dt\int_0^\infty b(t)\cdot\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds\\ &\leqq b(t)\cdot(\langle f^*(\cdot,t),u_t^*(\cdot,t)\rangle+(h_0(t)\cdot u^*(0,t)-h_1(t)\cdot u^*(1,t))).\end{aligned}\tag{4.23}$$

Since $\int_I u^*(x,t)\,dx=0$, Poincaré's inequality implies again

$$-\tfrac{1}{2}b'(t)\cdot\|u_t^*(\cdot,t)\|^2+\eta\cdot b(t)\cdot\|u_{xt}(\cdot,t)\|^2\geqq b(t)\cdot c\cdot\|u_{xt}(\cdot,t)\|^2\tag{4.24}$$

with $c=(2\pi^2\eta-\kappa)(2\pi^2)^{-1}$. Also, for any $\gamma>0$ there exists $C_\gamma>0$ such that

$$\begin{aligned}|h_j(t)\cdot u_t^*(j,t)\cdot b(t)|&\leqq\gamma\cdot b(t)\cdot\|u_{xt}(\cdot,t)\|^2+b(t)\cdot h_j^2(t)\\ &\quad+C_\gamma\cdot b(t)\cdot|h_j(t)|\cdot\|u_t^*(\cdot,t)\|,\end{aligned}\tag{4.25}$$

due to the compactness of the imbedding $\{u^*\in W^{1,2}|\int_I u^*\,dx=0\}$ into $C(I)$. Picking $\gamma=C/4$, inserting (4.24) and (4.25) into (4.23), and integrating with respect to t then gives again (4.11) as before. □

Finally, it should be noted that the arguments above also give estimates

$$b(t)\cdot\int_0^\infty\int_I G(s,u_x(x,t)-u_x(x,t-s))\,dx\,ds\leqq C(t),\tag{4.26}$$

where $C(t)$ is the right-hand side in (4.4) resp. in (4.11). If, e.g., (2.3) is satisfied with $C_1(t)=0$, then (4.26) implies that the right-hand side of (1.1) still decays in a certain sense in $W^{-1,2}(I)$. This, together with suitable asymptotic results for weak solutions of inhomogeneous heat equations, can be used to show that $u_{xxt}(\cdot,t)$ and $u_{tt}(\cdot,t)$ decay to zero in $W^{-1,2}(I)$, as $t\to\infty$.

5. Simple shear flow for viscoelastic liquids. In this section, we apply the previous results to a class of equations that model a certain unsteady flow for certain non-Newtonian liquids. Typical examples for such liquids are polymer solutions, such as glues, paints, engine fuels with polymer additives or protein mixtures. The liquid is assumed to occupy a reference configuration at time $t=-\infty$ with a Lagrangian coordinate system $\boldsymbol{\xi}=(\xi^1,\xi^2,\xi^3)$. Let $\mathbf{y}(\boldsymbol{\xi},t)$ denote the position of the fluid particle $\boldsymbol{\xi}$ at time $t>-\infty$, let $\mathbf{T}(\boldsymbol{\xi},t)$ denote the Cauchy stress at $\boldsymbol{\xi}$ and t (i.e. the stress measured with respect to the deformed configuration), and let

$$\mathbf{F}(\boldsymbol{\xi},t)=\mathbf{F}=\left(\frac{\partial y^i}{\partial\xi^j}\right)_{ij}$$

be the deformation gradient. Assuming that the material is incompressible (i.e. det $\mathbf{F}=1$) and has density ρ, the equations of motion are

$$\rho\cdot\mathbf{y}_{tt}=\operatorname{div}_{\boldsymbol{\xi}}(\mathbf{T}\cdot\mathbf{F}^{-T})+\mathbf{f},\tag{5.1}$$

where $\mathbf{f}$ denotes body forces and the superscript $^{-T}$ indicates the operation of taking the inverse of the transpose of a matrix (see [11]). It is customary to write constitutive equations for elastic liquids in terms of the "upper convected stress" $\pi=\mathbf{F}^{-1}\cdot\mathbf{T}\cdot\mathbf{F}^{-T}$ and the Cauchy strains $\boldsymbol{\gamma}(\boldsymbol{\xi},t)=\mathbf{F}^T(\boldsymbol{\xi},t)\cdot\mathbf{F}(\boldsymbol{\xi},t)$. Using the relative strain invariants

$$I_1(\xi, t, s) = \operatorname{tr}(\gamma(\xi, t) \cdot \gamma^{-1}(\xi, s)), \qquad I_2(\xi, t, s) = \operatorname{tr}(\gamma^{-1}(\xi, t) \cdot \gamma(\xi, s)), \tag{5.2}$$

a wide class of constitutive equations can then be written as

$$\begin{aligned} \pi(t) &= -p(t)\gamma^{-1}(t) - \eta \cdot \partial_t \gamma^{-1}(t) \\ &\quad + \int_0^\infty W_1(s, I_1(t, t-s), I_2(t, t-s)) \cdot \gamma^{-1}(t-s)\, ds \\ &\quad - \int_0^\infty W_2(s, I_1(t, t-s), I_2(t, t-s)) \cdot \gamma^{-1}(t) \cdot \gamma(t-s) \cdot \gamma^{-1}(t)\, ds, \end{aligned} \tag{5.3}$$

where $W_i : \mathbb{R}^+ \times [3, \infty)^2 \to \mathbb{R}$ are suitable functions, $p(t)$ is an undetermined (reactive) pressure and $\eta \geqq 0$ is (Newtonian) dynamic viscosity. The dependence on the particle ξ has been suppressed. Special cases include viscous ($\eta > 0$) and inviscid ($\eta = 0$) Newtonian fluids ($W_i = 0$); various "rubberlike liquid" models, in which typically $W_i(t, I_1, I_2) = a_i(t)$ (cf. [3], [15] and the literature given there); the "K-BKZ-model" [2] in which $W_j = \partial_{I_j} W(t, I_1, I_2)$ for some scalar function W; in particular the "Doi-Edwards" model [7], which is a K-BKZ-model [5], derived from molecular considerations, with the separable structure $W(t, I_1, I_2) = a(t) \cdot W_0(I_1, I_2)$; and various empirical models, in which one often assumes a semi-separable structure

$$W_j(t, I_1, I_2) = \sum_{k \leqq N} a_{kj}(t) \cdot g_{kj}(I_1, I_2) \tag{5.4}$$

(see, e.g., [17]). Some of these models have originally been proposed for the case of vanishing dynamic viscosity η (corresponding to, e.g., concentrated polymer solutions or polymer melts); we shall, however, always assume that $\eta > 0$.

We want to study simple shear flow between two parallel infinite plates (parallel to the y^1-y^2-plane) of unit separation (in the y^3-direction). Using Lagrangian coordinates ξ that agree with the spatial coordinates $\mathbf{y}$, we are thus looking for a motion of the special form $y^1(\xi, t) = \xi^1 + u(\xi^3, t)$, $y^i(\xi, t) = \xi^i$ for $i = 2, 3$, with volume forces acting only in the y^1-direction and depending only on t and ξ^3. It turns out that such an *ansatz* is consistent with the equations of motion and with the constitutive assumption (5.3). One finds that in this situation $I_1(\xi, t, s) = I_2(\xi, t, s) = 3 + |u_x(x, t) - u_x(x, s)|^2$ (writing x for ξ^3), and (5.1) becomes, assuming constant density $\rho = 1$,

$$u_{tt}(x, t) = \eta \cdot u_{xxt}(x, t) + \int_0^\infty g(s, u_x(x, t) - u_x(x, t-s))_x\, ds + f(x, t), \tag{5.5}$$

where $g(s, v) = (W_1(s, 3 + v^2, 3 + v^2) + W_2(s, 3 + v^2, 3 + v^2)) \cdot v$ and f is the sum of the force component f^1 and a contribution from the pressure gradient in the flow direction.

If the liquid adheres to the plates which move parallel to the flow direction, one obtains the boundary conditions (1.2); if traction forces of magnitude $h_1(t)$ and $-h_0(t)$ in the y^1-direction act on the liquid at the plates, boundary conditions (1.4) arise. In applications, one frequently has $h_j = 0$, which means that there is a lubricant between the plates and the liquid. We do not discuss friction type boundary conditions, although similar results could be shown with the methods employed above. The flow history for $t \leqq 0$ is given by u_0; u_1 is the flow velocity at $t = 0$ (which is allowed to have a jump across $t = 0$). The results obtained in the previous section then have the following interpretations:

(i) For smooth flow histories and smooth forces and displacements, (5.1) with the corresponding boundary conditions has a unique local (in time) solution. This holds in fact for completely general flows and for a much wider class of constitutive equations, as was shown in [18].

(ii) One has global existence for arbitrary forces and initial histories, if, e.g., in (5.4) the a_{ki} are nonincreasing, C^1-smooth for $t>0$ and integrable on $(0,\infty)$ and if the g_{ki} satisfy $\partial/\partial v(g_{ki}(3+v^2, 3+v^2)\cdot v) \geqq -M$ for all $v \in \mathbb{R}$ and some constant M. This class of functions includes polynomially growing functions and the case in which the g_{ki} are rational functions of their arguments that vanish at ∞ (see [17]) or, as in the Doi-Edwards model, derivatives of certain elliptic integrals (see [5]). The first assumption would describe a "shear-thickening" liquid, the last two correspond to liquids that show a "shear-thinning" behavior (which is commonly observed). We note that in this latter case, the integral operator in (5.5) satisfies a global Lipschitz condition, such that the arguments of §§ 2 and 3 will in fact not be needed. Also, our assumptions allow weakly singular kernels a_{kj}, which are predicted by some molecular theories (see [7]).

(iii) If, e.g., in (5.4) for some $\delta>0$

$$a'_{ki}(t)+\delta\cdot a_{ki}(t)\leqq 0, \tag{5.6}$$

and

$$\int_0^v g_{ki}(3+r^2, 3+r^2)\cdot r\,dr \geqq 0 \quad \text{for all } v\in\mathbb{R}, \tag{5.7}$$

then forces that decay like $e^{-\gamma t}$ with $\gamma<\min(\delta, 2\eta\pi^2)$ give a displacement u in a boundary traction experiment that reaches a steady state at the same rate. This will also hold for flows with prescribed boundary displacement, if the plates are kept fixed after a certain finite time. If the forces decay at a weaker (e.g. algebraic) rate, the flow velocity will again decay at a comparable rate. It is not claimed that these rates are optimal, but they show the interplay between the force decay (measured by γ), the Newtonian dissipation mechanism (given by η) and the dissipation due to the history dependence in the constitutive law (expressed by δ). Also, no ellipticity conditions are needed for the differential operators under the integrals in (5.5); such conditions will rather play a role in the case of vanishing Newtonian viscosity η (see [19]) or for the study of steady flows. Assumptions (5.6) and (5.7) are always true for various "rubber-like liquid" models as well as for the Doi-Edwards model with additional Newtonian viscosity. Finally, using (4.22), one can also show that for decaying volume forces, space averages of shear stresses and of normal stress differences will decay at the same rate.

Acknowledgment. The author wishes to thank the referee of an earlier version of this paper for several helpful hints.

REFERENCES

[1] G. ANDREWS, *On the existence of solutions to the equation* $u_{tt}=u_{xxt}+\sigma(u_x)_x$, J. Differential Equations, 35 (1980), pp. 200-231.

[2] B. BERNSTEIN, E. A. KEARSLEY AND L. J. ZAPAS, *A study of stress relaxation with finite strain*, Trans. Soc. Rheol., 7 (1963), pp. 391-410.

[3] R. B. BIRD, O. HASSAGER, R. C. ARMSTRONG AND C. F. CURTISS, *Dynamics of Polymeric Liquids*, I, II, John Wiley, New York, 1977.

[4] H. BREZIS, *Opérateurs maximaux monotones*, North-Holland, Amsterdam, 1973.

[5] P. K. CURRIE, *Constitutive equations for polymer melts predicted by the Doi-Edwards and Curtiss-Bird kinetic theory models*, J. Non-Newtonian Fluid Mech., 11 (1982), pp. 53-68.

[6] C. M. Dafermos and J. A. Nohel, *A nonlinear hyperbolic Volterra equation in viscoelasticity*, Amer. J. Math. Suppl., (1981), pp. 87-116.

[7] M. Doi and S. F. Edwards, *Dynamics of concentrated polymer systems*, J. Chem. Soc. Faraday, 74 (1978), pp. 1789-1832 and 75 (1979), pp. 38-54.

[8] H. Engler, *Stabilization of solutions for a class of parabolic integro-differential equations*, Nonlinear Anal. Theory, Methods, Applications, 8 (1984), pp. 1337-1371.

[9] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[10] J. M. Greenberg, *On the existence, uniqueness and stability of the equation* $\rho_0 X_{tt} = E(X_x)X_{xx} + \lambda X_{xxt}$, J. Math. Anal. Appl., 25 (1969), pp. 575-591.

[11] M. Gurtin, *Topics in Finite Elasticity*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[12] M. L. Heard, *An abstract parabolic Volterra integro-differential equation*, this Journal, 13 (1982), pp. 81-105.

[13] W. J. Hrusa and J. A. Nohel, *The Cauchy problem in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 59 (1985), pp. 388-412.

[14] O. A. Ladyzenskaya, V. A. Uraltseva and N. N. Solonnikov, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[15] A. S. Lodge, *Body Tensor Fields in Continuum Mechanics*, Academic Press, New York, 1974.

[16] R. C. MacCamy, *A model for one-dimensional nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21-33.

[17] A. C. Papanastasiou, L. E. Scriven and C. W. Macosko, *An integral constitutive equation for mixed flows: Viscoelastic characterization*, J. Rheol., 27 (1983), pp. 387-410.

[18] M. Renardy, *Local existence theorems for the first and second initial boundary value problems for a weakly non-Newtonian fluid*, Arch. Rational Mech. Anal., 83 (1983), pp. 229-244.

[19] ———, *A local existence theorem for a* K-BKZ *fluid*, Arch. Rational Mech. Anal., 88 (1985), pp. 83-94.

[20] ———, *A quasilinear parabolic equation describing the elongation of thin filaments of polymeric liquids*, this Journal, 13 (1982), pp. 226-238.

[21] Y. Yamada, *Some remarks on the equation* $y_{tt} - \sigma(y_x)y_{xx} - y_{xtx} = f$, Osaka J. Math., 17 (1980), pp. 303-323.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

009

SOLUTIONS TO THE KORTEWEG–DE VRIES EQUATION WITH INITIAL PROFILE IN $L_1^1(\mathbb{R}) \cap L_N^1(\mathbb{R}^+)$*

AMY COHEN† AND THOMAS KAPPELER‡

Abstract. The Cauchy problem for the Korteweg-deVries equation is considered with initial profile integrable against $(1+|x|)\,dx$ on $\mathbb{R}$ and against $(1+|x|)^N\,dx$ on $\mathbb{R}^+$. Classical solutions are constructed for $N \geqq 11/4$. Under mild additional hypotheses the solution evolves in $L^2(\mathbb{R})$.

Key words. Korteweg-de Vries equation, inverse scattering method

AMS(MOS) subject classification. 35Q20

1. Introduction and summary of results. This paper considers the initial value problem for the Korteweg-deVries equation (KdV),

$$u_t - 6uu_x + u_{xxx} = 0, \tag{1.1}$$

$$u(x,0) = U(x), \tag{1.2}$$

under the hypothesis that

$$\int_{-\infty}^{\infty} |U(x)|(1+|x|)\,dx < \infty, \tag{1.3a}$$

$$\int_{0}^{\infty} |U(x)|(1+|x|)^N\,dx < \infty. \tag{1.3b}$$

No differentiability is assumed at all. The goal is to find the range of N such that the problem (1.1), (1.2) has a solution. Our existence theorem is based on a construction suggested by the inverse scattering method. We show that if $N \geqq 11/4$, then a classical solution exists in $t > 0$ which approaches its initial profile in an appropriate distribution sense as $t \to 0^+$.

These results improve considerably on earlier work of the first author [3], which required that U be at least piecewise C^1 as well as that U be integrable against $(1+|x|)^N\,dx$ on $\mathbb{R}$ for large enough N. By using Kappeler's new L^2 inverse scattering result [8], we are also able to get control over our solution as $x \to -\infty$, at least for U satisfying a rather mild additional hypothesis. These results also improve on work of Sachs [14], who requires that $U(x)$ be integrable against $(1+|x|)^N\,dx$ on all of $\mathbb{R}$ with $N > 11/4$ rather than only on $\mathbb{R}^+$ with $N \geqq 11/4$. Sachs claims convergence to initial profile in a weighted L^1 norm on each halfline $[a, +\infty)$; it appears that his proof of this point is flawed.

There is no direct comparison between our results and the very interesting paper of Kruzhkov and Faminskii [11], in which they prove the existence of a weak solution to KdV with arbitrary L^2 initial data, and show that the solution is classical if the datum is not only L^2 on $\mathbb{R}$ but also L^2 with respect to $(1+|x|)^3\,dx$ on $\mathbb{R}^+$. While Sachs' paper uses a different inverse scattering construction from ours (Deift and Trubowitz [5] rather than Faddeev [6]), Kruzhkov and Faminskii use a different approach altogether: they cut off and mollify their initial profile, apply results of Yakupov [19] and Shabat [16] solving KdV with data in $C_0^\infty(\mathbb{R})$, and then take limits.

* Received by the editors October 8, 1984; accepted for publication (in revised form) April 7, 1986.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

‡ Forschungsinstitut für Mathematik, ETH-Zentrum, CH-8092 Zürich, Switzerland. The author was supported in part by the Swiss National Foundation.

In their pioneering paper [7], Gardner, Greene, Kruskal and Miura showed that if $u(x, t)$ solved KdV and evolved in the Schwartz class $\mathcal{S}$, then the scattering data of the Schrödinger equation

$$-y''+u(x,t)y=k^2y \tag{1.4}$$

evolved according to simple first order linear o.d.e.'s in the variable t. By appealing to Faddeev's inverse scattering theory [6], they showed that $u(x, t)$ for $t>0$ could be recovered from $u(x, 0)$. This idea has been the basis for a succession of existence theorems [17], [3], [13], [14] employing progressively weaker hypotheses on the initial profile U.

Rather than give a detailed exposition of the forward scattering theory of (1.4) we refer the reader to Cohen's paper [4].

In § 2 we analyze the scattering data associated to (1.4) under the hypothesis $U\in L^1_N(\mathbb{R})$, i.e., U is integrable with respect to $(1+|x|^N)\,dx$, with $N\geqq 1$. The main result is Proposition 2.5 which says that generically the reflection coefficient R_+ is in $C^{N-1}(\mathbb{R})\cap C^N(\mathbb{R}\sim\{0\})$ and $\lim_{k\to 0} kR_+{}^{(N)}(k)$ exists—but that if U is exceptional, then R_+ is only in $C^{N-2}(\mathbb{R})\cap C^{N-1}(\mathbb{R}\sim\{0\})$ and $\lim_{k\to 0} kR_+{}^{(N-1)}(k)$ exists.

In § 3 we analyze the kernels $\Omega_+(x, t)$ and $\Omega_-(x, t)$ used in the Marchenko equations

$$B_+(x,y,t)+\Omega_+(x+y,t)+\int_0^\infty B_+(x,z,t)\Omega_+(x+y+z,t)\,dz=0, \tag{M+}$$

$$B_-(x,y,t)+\Omega_-(x+y,t)+\int_{-\infty}^0 B_-(x,z,t)\Omega_-(x+y+z,t)\,dz=0. \tag{M−}$$

What Gardner, Green, Kruskal and Miura showed was that if $u(x, t)$ solves KdV, and $\Omega_\pm$ are as defined below, then

$$u(x,t)=-\partial_x B_+(x,0,t)=+\partial_x B_-(x,0,t).$$

The kernels are defined as follows:

$$\Omega_+(x,t)=F_+(x,t)+2\sum c_{+j}\exp(-2\kappa_j x+8\kappa_j^3 t)$$

where

$$F_+(x,t)=\pi^{-1}\int_{-\infty}^\infty R_+(k)\exp(2ikx+8ik^3t)\,dk$$

and

$$\Omega_-(x,t)=F_-(x,t)+2\sum c_{-j}\exp(2\kappa_j x-8\kappa_j^3 t)$$

where

$$F_-(x,t)=\pi^{-1}\int_{-\infty}^\infty R_-(k)\exp(-2ikx-8ik^3t)\,dk.$$

Clearly the existence, regularity and decay of the $B_\pm(x, y, t)$ depend on the regularity and decay of the $\Omega_\pm$. In § 3, we show that for each fixed $t>0$, $\partial_x^\nu\Omega_+(x, t)$ is continuous for $0\leqq\nu\leqq 2N+3/2$ and establish algebraic decay rates as $x\to+\infty$ for these derivatives. We also analyze the decay and regularity of $\Omega_+(x, t)$ using the properties of R_+ proved in Proposition 2.5. To study F_-, we note just that R_- is quite similar to R_+ in its regularity and decay. Then we see that the decay of Ω_- is controlled by that of F_- and that the integral for F_- has stationary points when $x<0$. Nonetheless we find that if $U \in L^1_N$, $N \geqq 5$, and $R_-^{(n)}(k) = O(k^{-\lambda})$ for $\lambda \geqq 5/2$, then $\partial_x \Omega_-(x, t) = O(|x|^{-\lambda/2+1/4})$ as $x \to -\infty$.

In § 5 we prove sharper versions of the following results.

Result 1. Suppose U satisfies (1.3a) and (1.3b) with $N \geqq 11/4$. Then there is a classical solution $u(x, t)$ of KdV in $t > 0$ such that

$$u(x, t) \to U(x) \quad \text{in } H^{-1}(+\infty).$$

Result 2. Suppose that $U \in L^1_5(\mathbb{R}) \cap L^2(\mathbb{R})$ and that $R_+^{(n)}(k) = O(|k|^{-\lambda})$ as $k \to \pm\infty$ for some $\lambda > 5/2$, and $n = 0, 1, 2$. Then the solution given by Result 1 evolves in $L^2(\mathbb{R})$ for $t > 0$.

Result 3. Suppose that $N \geqq 3$ and that $U \in L^1_N(\mathbb{R})$ if U is generic but that $U \in L^1_{N+1}(\mathbb{R})$ if U is nongeneric. Suppose further that $(1+|x|)^{N-1}U(x)$ is in $L^2(\mathbb{R})$. If $u(x, t)$ is the solution to KdV given by Result 1, then $x^\alpha u(x, t) \to x^\alpha U(x)$ in $L^2(+\infty)$ as $t \to 0$ for $\alpha = 0$ and $\alpha = N - 1$.

We should also remark that the question of uniqueness is still largely open. Uniqueness is known for the initial value problem for KdV if the initial profile is in H^s with $s \geqq 3/2$ [2], [10], [15]. Uniqueness is also known within the class of solutions $u(x, t)$ such that $u(x, t)$ and $u_x(x, t)$ go to 0 as $x \to \pm\infty$ and $u_{xx}(x, t)$ is bounded as $x \to \pm\infty$ [12]. Kruzhkov and Faminskii [11] have shown that the problem (1.1), (1.2) is well posed in the class of functions U which are L^2 on $\mathbb{R}$ and L^2 with respect to a weight on $\mathbb{R}^+$. Unless we add to our minimal hypotheses we cannot show that our solution $u(x, t)$ evolves in a class where either of these uniqueness theorems applies.

Notational conventions. The operator ∂_x denotes the partial derivative with respect to the subscript variable.

$$f^*(x, k) = \text{the complex conjugate of } f(x, k).$$

In dealing with functions of x and k, prime (′) always denotes the x-derivative and dot (·) always denotes the k-derivative; thus

$$f'(x, k) = \partial_x f(x, k), \qquad \dot{f}(x, k) = \partial_k f(x, k).$$

The space $L^1_N(+\infty)$ consists of functions $g(x)$ such that

$$\int_X^\infty |g(x)|(1+|x|)^N \, dx < \infty \quad \text{for all finite } X.$$

The space $L^2(+\infty)$ consists of functions $g(x)$ such that

$$\int_X^\infty |g(x)|^2 \, dx < \infty \quad \text{for all finite } X.$$

We use aVb to denote $\max\{a, b\}$.

2. Analysis of the initial scattering data.

2.1. The Jost functions. Suppose that $U(x)$ belongs to $L^1_N(\mathbb{R})$ with $N \geqq 1$. Then the Jost functions for

$$-y'' + U(x)y = k^2 y \tag{2.1}$$

are the solutions $f_+(x, k)$ and $f_-(x, k)$ with the asymptotic behavior

$$f_+(x, k) \sim e^{+ikx} \quad \text{as } x \to +\infty, \qquad f_-(x, k) \sim e^{-ikx} \quad \text{as } x \to -\infty. \tag{2.2}$$

These exist for $\operatorname{Im} k \geqq 0$ and can be represented as

$$f_+(x, k) = e^{ikx} h_+(x, k), \qquad f_-(x, k) = e^{-ikx} h_-(x, k) \tag{2.3}$$

where

$$(2.4)\qquad h_+(x,k)=1+\int_0^\infty B_+(x,y)\,e^{2iky}\,dy,\qquad h_-(x,k)=1+\int_{-\infty}^0 B_-(x,y)\,e^{-2iky}\,dy$$

where, in turn,

$$B_\pm(x,\cdot)\in L^1(\mathbb{R}^\pm)\cap L^\infty(\mathbb{R}^\pm)\subset L^2(\mathbb{R}^\pm),$$

$$B_\pm(x,y)\quad\text{is continuous on }\mathbb{R}\times\mathbb{R}^\pm.$$

Here $\mathbb{R}^+=[0,\infty)$ and $\mathbb{R}^-=(-\infty,0]$. Moreover, the maps $x\to B_\pm(x,\cdot)$ are absolutely continuous and Fréchet differentiable from $\mathbb{R}$ to $L^1(\mathbb{R}^\pm)$. The following estimates are valid since $U\in L^1_1(\mathbb{R})$:

$$(2.5+)\qquad |B_\pm(x,y)|\leqq\left[\exp\left\{\int_x^\infty (t-x)|U(t)|\,dt\right\}\right]\int_{x+y}^\infty |U(t)|\,dt,$$

$$(2.6+)\qquad \begin{aligned}&|\partial_x B_+(x,y)+U(x+y)|\\&\leqq\left[\exp\left\{\int_x^\infty (t-x)|U(t)|\,dt\right\}\right]\int_x^\infty |U(t)|\,dt\int_{x+y}^\infty |U(s)|\,ds,\end{aligned}$$

$$(2.7+)\qquad \begin{aligned}&|\partial_y B_+(x,y)+U(x+y)|\\&\leqq 2\left[\exp\left\{\int_x^\infty (t-x)|U(t)|\,dt\right\}\right]\int_x^\infty |U(t)|\,dt\int_{x+y}^\infty |U(s)|\,ds.\end{aligned}$$

Analogous bounds (2.5−), (2.6−) and (2.7−) hold for $B_-(x,y)$, $\partial_x B_-(x,y)-U(x+y)$ and $\partial_y B_-(x,y)-U(x+y)$ in terms of integrals over left-half-lines. See [1], [3]-[6] for details. Applying these bounds to the forms (2.3), (2.4), one obtains the following.

PROPOSITION 2.1. *For any fixed* x, *the functions* $y^nB_+(x,y)$, $y^n\partial_xB_+(x,y)$, *and* $y^n\partial_yB_+(x,y)$ *are integrable over* $0<y<\infty$ *for* $0\leqq n\leqq N-1$. *Similar results hold for* B_- *with integrability over* $-\infty<y<0$.

It follows that $h_+(x,k)$ and $\partial_xh_+(x,k)$ are $(N-1)$ times continuously differentiable with respect to k. Indeed, if $1\leqq n\leqq N-1$ and $\operatorname{Im}k\geqq0$, then

$$(2.8)\qquad \partial_k^n[h_+(x,k)]=\int_0^\infty (2iy)^nB_+(x,y)\,e^{2iky}\,dy.$$

If in addition $k\neq0$, then an integration by parts yields

$$\partial_k^n[h_+(x,k)]\frac{-1}{2ik}\int_0^\infty [n(2iy)^{n-1}B_+(x,y)+(2iy)^n\partial_yB_+(x,y)]\,e^{2iky}\,dy.$$

Further

$$(2.9)\qquad \partial_k^n\partial_x^1[h_+(x,k)]=\int_0^\infty (2iy)^n\partial_xB_+(x,y)\,e^{2iky}\,dy.$$

Thus for $1\leqq n\leqq N-1$ and for each finite X, $k\partial_k^nh_+(x,k)$ and $\partial_k^n\partial_x^1h_+(x,k)$ are uniformly bounded on $\{(x,k)\colon x\geqq X \text{ and } \operatorname{Im}k\geqq0\}$.

It is possible to get better information about the regularity of $h_+(x,k)$ by using the approach of Deift and Trubowitz [5, p. 130]. Let

$$(2.10)\qquad D_k(y)=\int_0^y e^{2ikt}\,dt=(e^{2iky}-1)/2ik.$$

Deift and Trubowitz show that $h_+(x, k) = 1 + \int_x^\infty D_k(t-y)U(t)h_+(t, k)\,dt$. The next several propositions are similar to results in [4]-[6].

PROPOSITION 2.2. *Assume that $U \in L_N^1(\mathbb{R})$ with $N \geqq 1$. As functions of k with x fixed in $\mathbb{R}$, $h_+(x, k)$ and $\partial_x h_+(x, k)$ are C^{N-1} on $\{k\colon \operatorname{Im} k \geqq 0\}$ and C^N on $\{k\colon \operatorname{Im} k \geqq 0, k \neq 0\}$. Further $k\partial_k^N h_+(x, k)$ and $k\partial_k^N \partial_x h_+(x, k)$ extend continuously to $k = 0$. Moreover there are nonincreasing functions $K(x)$ such that for $0 \leqq n \leqq N$, $\operatorname{Im} k \geqq 0$, and $x \leqq t < \infty$*

(i) $|k\,\partial_k^n h_+(t, k)| \leqq K(x)$,

(ii) $\dfrac{|k|}{|k|+1}|\partial_k^n \partial_t^1 h_+(t, k)| \leqq K(x)$, *and*

(iii) $k\,\partial_k^n[h_+(x, k)] \to 0$ *and* $\partial_k^n \partial_x^1[h_+(x, k)] \to 0$ *as* $|k| \to \infty$, *uniformly in* $\operatorname{Im} k \geqq 0$.

Proof. We have already noted the claimed regularity on $\{k\colon \operatorname{Im} k \geqq 0\}$. To get the Nth derivative away from $k = 0$, we differentiate (2.10) N times formally and multiply by k. Thus if $\partial_k^N h_+(x, k)$ exists then $w \equiv k\,\partial_k^N h_+(x, k)$ satisfies the integral equation

$$(\mathbf{I} - \mathbf{T})w = r \tag{2.11}$$

where

$$\mathbf{T}[g](x) \equiv \int_x^\infty D_k(t-x)U(t)g(t)\,dt$$

and

$$r(x) \equiv \sum_{\nu=1}^N C_\nu \int_x^\infty k\,\partial_k^\nu[D_k(t-x)]U(t)\partial_k^{N-\nu} h_+(t, k)\,dt$$

for easily computable C_ν. For any finite X, $\mathbf{T}$ is a bounded operator on $L^\infty(X, \infty)$; indeed for $m \in \mathbb{N}$

$$\|\mathbf{T}^m g\| \leqq \|g\| \left(\int_X^\infty |U(t)|\,dt/|k| \right)^m.$$

Since $|k\partial_k^\nu[D_k(y)]| \leqq |2y|^\nu$ for all $\nu \geqq 0$, it follows that

$$\|r\| \leqq \sum_{\nu=1}^N C_\nu \int_x^\infty |2(t-x)|^\nu |U(t)| A(x)\,dt$$

where

$$A(x) = \sup\{|\partial_k^\mu h_+(t, k)|\colon \operatorname{Im} k \geqq 0, 0 \leqq \mu \leqq N-1, x \leqq t \leqq \infty\}.$$

Note that $A(x)$ is finite and nonincreasing. One can also verify that

$$\int_x^\infty |2(t-x)|^\nu |U(t)|\,dt \leqq K_\nu(x) \quad \text{for } 1 \leqq \nu \leqq N$$

where K_ν is the nonincreasing function

$$K_\nu(x) = \begin{cases} \displaystyle\int_{-\infty}^\infty |2t|^\nu |U(t)|\,dt + |2x|^\nu \int_x^\infty |U(t)|\,dt & \text{if } x < 0, \\ \displaystyle\int_{-\infty}^\infty |2t|^\nu |U(t)|\,dt & \text{if } x \geqq 0. \end{cases}$$

So $\|r\|$ in $L^\infty(X, \infty)$ is bounded by a nonincreasing function $B(X) = K(X)\sum_1^N C_\nu K_\nu$. It follows that the solution of (2.11) is given by

$$w = \sum_{m=0}^\infty \mathbf{T}^m r$$

and that w is continuous in x and k in $\mathbb{R}\times\{\operatorname{Im} k\geqq 0\}$. Further analysis reveals that $w(x,k)/k$ is indeed $\partial_k^N h_+(x,k)$ and that

$$|k\partial_k^N h_+(x,k)|\leqq \exp\left(\int_x^\infty |U(t)|\,dt/|k|\right)B(x).$$

Continuing in this vein one finds that $k\partial_k^N h_+(x,k)$ has a derivative with respect to x in distribution sense, and then that $\partial_x^1[k\partial_k^N h_+(x,k)]$ is a classical derivative as well, and satisfies (ii) and (iii). □

Remark. The factor $(1+|k|)^{-1}$ in (ii) is necessary because the term with $n=N$ in the sum for $r(x)$ involves $kh_+(x,k)$, which grows like $|k|$ as $|k|\to\infty$.

2.2. Regularity and decay of $W[f_-,f_+]$ and $W[f_+^*,f_-]$. Let $W(k)$ and $V(k)$ be defined on $\operatorname{Im} k\geqq 0$ by $W(k)=W[f_-,f_+]$ and $V(k)=W[f_+^*,f_-]$. Since f_-, f_+, and f_+^* solve (2.1), these Wronskians are independent of x. Evaluating at $x=0$, we get

$$W(k)=h_-(0,k)h_+'(0,k)-h_-'(0,k)h_+(0,k)+2ikh_-(0,k)h_+(0,k) \tag{2.12}$$

and

$$V(k)=h_+^*(0,k)h_-'(0,k)-h_-(0,k)h_+^{*\prime}(0,k). \tag{2.13}$$

Where ambiguity is possible we reserve prime (′) for $\partial/\partial x$ and dot ($\cdot$) for $\partial/\partial k$.

The following propositions follow immediately from the results of § 2.1.

PROPOSITION 2.3. *Assume* $U\in L_N^1(\mathbb{R})$ *with* $N\geqq 1$. *Then* $W\in C^{N-1}(\mathbb{R})\cap C^N(\mathbb{R}\sim\{0\})$. *Moreover* $k\partial_k^N[W(k)]$ *extends continuously to* $k=0$. *For all* n with $0\leqq n\leqq N$, $\lim_{|k|\to\infty}\partial_k^n[W(k)-2ik]=0$.

PROPOSITION 2.4. *Assume* $U\in L_N^1(\mathbb{R})$ *with* $N\geqq 1$. *Then* $V\in C^{N-1}(\mathbb{R})\cap C^N(\mathbb{R}\sim\{0\})$; $k\partial_k^N[V(k)]$ *extends continuously to* $k=0$; *and* $\lim_{|k|\to\infty}\partial_k^n[V(k)]=0$ *for* $0\leqq n\leqq N$.

2.3. Regularity and decay of $R_+(k)$, $R_-(k)$. Recall that the reflection coefficients R_+ and R_- are defined for $k\neq 0$ by

$$R_+(k)=\frac{V(k)}{W(k)},\qquad R_-(k)=\frac{V^*(k)}{W(k)}.$$

We concentrate on $R_+(k)$; $R_-(k)$ can be analyzed by the same methods. Note that

$$W(k)R_+(k)=V(k)$$

so that formally

$$W(k)R_+^{(n)}(k)=V^{(n)}(k)-\sum_{\nu=0}^{n-1}\binom{n}{\nu}R_+^{(\nu)}(k)W^{(n-\nu)}(k). \tag{2.14}$$

PROPOSITION 2.5. *Assume that* $U\in L_N^1(\mathbb{R})$ *with* $N\geqq 1$. *Then* $R_+\in C^N(\mathbb{R}\sim\{0\})$ *and* $\lim_{|k|\to\infty}kR_+^{(n)}(k)=0$ *for* $0\leqq n\leqq N$.

Furthermore

(A) *If U is of generic type, then* $R_+\in C^{N-1}(\mathbb{R})$ *and* $kR_+^{(N)}(k)$ *extends continuously to* $k=0$.

(B) *If U is of exceptional type and* $N\geqq 2$, *then* $R_+\in C^{N-2}(\mathbb{R})$ *and both* $kR_+^{(N-1)}(k)$ *and* $k^2R_+^{(N)}(k)$ *extend continuously to* $k=0$.

Proof. The regularity away from $k=0$ and the decay as $k\to\pm\infty$ follow from Propositions 2.3 and 2.4.

If U is generic then $W(k)$ is nonzero on $\mathbb{R}$ and (A) follows by an induction using (2.14). Suppose next that U is exceptional and that $N\geqq 2$. Then instead of treating R_+ as the ratio V/W we treat R_+ as the quotient of V/k and W/k. In this case it is known that W/k is continuous on $\mathbb{R}$ and never zero. Since V/k and W/k are C^{N-2} on $\mathbb{R}$, so is R_+. Using (2.14) it is easy to complete the proof of (B). □

We now turn to results involving L^2 hypotheses as well as L^1 assumptions on U.

LEMMA 2.6. *Suppose that $y^\nu U(y) \in L^2(\mathbb{R})$ and $U(y) \in L^1_{\nu+1}(\mathbb{R})$ for $0 \leqq \nu \leqq n$. Then $V^{(\nu)} \in L^2$ for $0 \leqq \nu \leqq n$.*

Proof. Deift and Trubowitz [5, p. 159] have proved that

$$V(k) = \int_{-\infty}^{\infty} \Pi_1(y)\, e^{-2iky}\, dy$$

where there is a constant K such that

$$|\Pi_1(y)| \leqq |U(y)| + KL(y)$$

for

$$L(y) = \int_y^\infty |U(t)|\, dt \quad \text{if } y \geqq 0, \qquad L(y) = \int_{-\infty}^y |U(t)|\, dt \quad \text{for } y < 0.$$

To show that $V^{(\nu)} \in L^2$, it suffices to show that $y^\nu \Pi_1(y) \in L^2$. Since

$$|\Pi_1(y)|^2 \leqq (1+K^2)(|U(y)|^2 + L(y)^2)$$

it follows that

$$\int_0^\infty |y^\nu \Pi_1(y)|^2\, dy \leqq (1+K^2) \int_0^\infty |y^\nu U(y)|^2\, dy + (1+K^2) \int_0^\infty y^{2\nu} L(y)^2\, dy.$$

The first term is finite since $y^\nu U(y) \in L^2$. Further

$$\begin{aligned}\int_0^\infty y^{2\nu} L(y)^2\, dy &= \int_{y=0}^\infty \left(y^\nu \int_{s=y}^\infty |U(s)|\, ds \right)\left(y^\nu \int_{t=y}^\infty |U(t)|\, dt \right) dy \\ &\leqq \int_{y=0}^\infty \left(\int_{s=0}^\infty s^\nu |U(s)|\, ds \right)\left(y^\nu \int_{t=y}^\infty |U(t)|\, dt \right) dy \\ &= \int_{s=0}^\infty s^\nu |U(s)|\, ds \int_{t=0}^\infty \frac{t^{\nu+1}}{\nu+1} |U(t)|\, dt < \infty.\end{aligned}$$

Thus $y^\nu \Pi_1(y)$ is in L^2 on $\mathbb{R}^+$; the proof that it is in L^2 on $\mathbb{R}^-$ is similar. □

PROPOSITION 2.7. *Suppose that $U \in L^1_N(\mathbb{R})$ and that $y^n U(y) \in L^2(\mathbb{R})$ for $0 \leqq n \leqq M$.*

(A) *If U is of generic type, then $R_+^{(n)} \in L^1(\mathbb{R})$ and $kR_+^{(n)}(k) \in L^2(\mathbb{R})$ for $0 \leqq n \leqq \min\{M, N-1\}$.*

(B) *If U is of exceptional type, then $R_+^{(n)} \in L^1(\mathbb{R})$ and $kR_+^{(n)}(k) \in L^2(\mathbb{R})$ for $0 \leqq n \leqq \min\{M, N-2\}$.*

Proof. The proof is an induction based on the formula

$$R_+^{(n)}(k) = \left[V^{(n)}(k) - \sum_{\nu=1}^n \binom{n}{\nu} R_+^{(n-\nu)}(k)\, W^{(\nu)}(k) \right] \Big/ W(k).$$

We discuss (A) first. Since $W(k)$ is continuous, never zero, and grows like $|k|$ at $\pm\infty$ it follows that $1/W$ is in L^2 and that $k/W \in L^\infty$. Since $V \in L^2$, it follows that $R_+ = V/W$ is both L^1 and L^2, and that $kR_+ \in L^2$.

Keep $0 \leqq n \leqq \min\{M, N-1\}$. We then know that $V^{(n)} \in L^2$ and that $W^{(\nu)} \in L^\infty$ for $1 \leqq \nu \leqq n$, then $R_+^{(n)} \in L^1 \cap L^2$ and $kR_+^{(n)} \in L^2$. Result (A) now follows by induction.

The induction for result (B) is similar, except that in the exceptional case we have only $R_+ \in C^{N-2}$. □

3. Regularity and decay of $\Omega_+(x, t)$ and $\Omega_-(x, t)$ for $t>0$.

3.1. Properties of $F_+(x, t)$. Recall from the introduction that the kernel of the Marchenko equation $(M+)$ is

$$\Omega_+(x, t) = F_+(x, t) + G_+(x, t)$$

where

$$F_+(x, t) = \pi^{-1} \int_{-\infty}^{\infty} R_+(k) \exp(2ikx + 8ik^3 t)\, dk$$

and

$$G_+(x, t) = 2 \sum_{j \in J} c_{+j} \exp(-2\kappa_j x + 8\kappa_j^3 t).$$

Since $G_+(x, t)$ is C^∞ and decays exponentially as $x \to +\infty$ for fixed $t>0$, we need to concentrate on the properties of F_+. In the first part of this subsection we use a representation of $F_+(x, t)$ in terms of $F_+(x)$ and the Airy function to find out as much as possible about $F_+(x, t)$ without using differentiability of $R_+(k)$. Later we report on what can be said of $F_+(x, t)$ using derivatives of $R_+(k)$ by a careful extension of the methods of Cohen in [3]. For convenience, we set

$$F_+(x) := F_+(x, 0) = \pi^{-1} \int_{-\infty}^{\infty} R_+(k)\, e^{2ikx}\, dk.$$

LEMMA 3.1. *If $U(x) \in L_1^1(\mathbb{R})$, then $R_+(k) \in L^2(\mathbb{R})$ and $F_+(x) \in L^2(\mathbb{R})$.*

Proof. These results are well known; see [5].

LEMMA 3.2. *Suppose* $U(x) \in L_1^1(\mathbb{R}) \cap L_N^1(\mathbb{R}^+)$ *for some N with $N \geqq 11/4$. Then*

(a) $$\int_0^\infty |F_+(x)|(1+x)^{N-1}\, dx < \infty,$$

(b) $$\int_0^\infty |\partial_+ F_+(x)|(1+x)^N\, dx < \infty.$$

Proof. Because of the exponential decay of $G_+(x, 0)$ as $x \to +\infty$, it is enough to prove the analogues of (a) and (b) for Ω_+. By Faddeev [6, p. 155], we know that

$$|\Omega_+(x)| \leqq C(x) \int_x^\infty |U(z)|\, dz \tag{3.1}$$

and

$$|\partial_x \Omega_+(x) - U(x)| \leqq C(x) \left[\int_x^\infty |U(z)|\, dz \right]^2 \tag{3.2}$$

where each $C(x)$ is a nonincreasing function of x. Now by (3.1)

$$\begin{aligned}
\int_0^\infty |\Omega_+(x)| x^{N-1}\, dx &\leqq \int_{x=0}^\infty C(0) \int_x^\infty |U(z)|\, dz\, x^{N-1}\, dx \\
&= C(0) \int_{z=0}^\infty |U(z)| \int_{x=0}^z x^{N-1}\, dx\, dz \\
&= C(0) N^{-1} \int_{z=0}^\infty |U(z)| z^N\, dz < \infty.
\end{aligned}$$

Thus (a) follows. For (b), use (3.2):

$$\int_0^\infty |\partial_+\Omega_+(x)|x^N\,dx \leqq \int_0^\infty |U(z)|x^N\,dx + \int_0^\infty C(x)\left[\int_x^\infty |U(z)|\,dz\right]^2 x^N\,dx.$$

The first term is finite by hypothesis. For the second,

$$\begin{aligned}\int_{x=0}^\infty \left[\int_x^\infty |U(z)|\,dz\right]^2 x^N\,dx &\leqq \int_{x=0}^\infty \left[x\int_x^\infty |U(z)|dz\right]\left[x^{N-1}\int_x^\infty |U(z)|\,dz\right]dx\\ &\leqq \int_{x=0}^\infty \left[\int_x^\infty z|U(z)|\,dz\right]\left[x^{N-1}\int_{z=x}^\infty |U(z)|\,dz\right]dx\\ &\leqq \int_0^\infty |U(z)|z\,dz\int_{x=0}^\infty x^{N-1}\left[\int_{z=x}^\infty |U(z)|\,dz\right]dx\\ &\leqq \int_0^\infty |U(z)|z\,dz\left[\int_{z=0}^\infty |U(z)|\int_{z=0}^\infty x^{N-1}\,dx\,dz\right]\\ &\leqq \left[\int_0^\infty |U(z)|z\,dz\right]N^{-1}\left[\int_{z=0}^\infty |U(z)|z^N\,dz\right]\\ &< \infty\end{aligned}$$

since each factor is finite. □

By Lemmas 3.1 and 3.2 we know that $F_+(x)$ is a real valued function such that

$$F_+(x)\in L^2(\mathbb{R})\cap L^1_{N-1}(\mathbb{R}^+),\qquad \partial_x F_+(x)\in L^1_{\text{loc}}(\mathbb{R})\cap L^1_N(\mathbb{R}^+).$$

To analyze $F_+(x, t)$ with $t>0$, we use the observation [13] that $F_+(x, t)$ is essentially a convolution of $F_+(x)$ with an Airy function:

$$F_+(x,t)=(3t)^{-1/3}\int_{-\infty}^\infty F_+(y)\,\mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right)dy. \tag{3.3}$$

We use the following properties of the Airy function [13]:

$$|\mathrm{Ai}\,(z)|<1\quad \forall z \text{ in } \mathbb{R}, \tag{3.4}$$

$$\mathrm{Ai}\,(z)\in C^\infty(\mathbb{R})\quad\text{and}\quad \mathrm{Ai}''\,(z)=z\,\mathrm{Ai}\,(z), \tag{3.5}$$

$$\begin{aligned}&|\mathrm{Ai}^{(n)}\,(z)|\leqq C_n^-(1+|z|)^{n/2-1/4}\quad \text{as } z\to-\infty,\\ &|\mathrm{Ai}^{(n)}\,(z)|\leqq C_n^+(1+|z|)^{n/2-1/4}\exp\,(-2|z|^{3/2}/3)\quad \text{as } z\to+\infty.\end{aligned} \tag{3.6}$$

Because of the different behavior at $+\infty$ and $-\infty$, it is convenient to divide the integral in (3.3) into pieces. To this end let $\zeta_1(x)$ denote a nonincreasing C^∞ function such that

$$\zeta_1(x)=1\quad \text{on } -\infty<x\leqq\tfrac{1}{3},\qquad \zeta_1(x)=0\quad \text{on } \tfrac{2}{3}\leqq x<\infty.$$

Let $\zeta_2(x):=1-\zeta_1(x)$. Next set $F_i(x)=F_+(x)\zeta_i(x)$ for $i=1,2$. Next set

$$F_i(x,t)=\begin{cases}(3t)^{-1/3}\displaystyle\int_{-\infty}^\infty F_i(y)\,\mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right)dy & \text{if } t>0,\\ F_i(x) & \text{if } t=0.\end{cases} \tag{3.7}$$

Note that $F_+(x,t)=F_1(x,t)+F_2(x,t)$.

LEMMA 3.3.

(a) $F_1(x,t)$ *is* C^∞ *in* $\mathbb{R}\times\mathbb{R}^+$;

(b) $\lim_{x\to+\infty} x^n\,\partial_x^j F_1(x,t)=0$ *for nonnegative integers* n, j;

(c) $\int_0^\infty |\partial_x^j F_1(x,t)|x^m\,dx<\infty$ *for all* j, m.

Proof. Part (c) follows immediately from (a) and (b). The regularity (a) follows from (3.7), the rapid decay of all $\mathrm{Ai}^{(j)}(z)$ as $z \to +\infty$, and the fact that supp $F_1 \subseteq (-\infty, 1]$. Now

$$\partial_x^j F_1(x,t) = (3t)^{-1/3} \int_{-\infty}^{\infty} F_1(y)\, \mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)(3t)^{-j/3}\, dy$$

and

$$|\partial_x^j F_1(x,t)| \leqq (3t)^{-(j+1)/3}\left[\int_{-\infty}^{1} |F_1(y)|^2\, dy\right]^{1/2}\left[\int_{-\infty}^{1} \left|\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)\right|^2 dy\right]^{1/2}.$$

Setting $\xi = (x-y)/(3t)^{1/3}$ we see that the second integral is

$$I(x,t) = \int_{(x-1)(3t)^{-1/3}}^{\infty} |\mathrm{Ai}^{(j)}(\xi)|^2 (3t)^{1/3}\, d\xi.$$

Since $\mathrm{Ai}^{(j)}(\xi)$ decays faster than exponentially as $\xi \to +\infty$, it follows that $I(x,t)$ decays at least exponentially fast as $x \to +\infty$, and (b) follows. □

We next analyze $F_2(x,t)$. Note that supp $F_2 \subseteq [0, +\infty]$ and that $\mathrm{Ai}\,((x-y)/(3t)^{1/3})$ is less well behaved as $y \to +\infty$. A technical remark precedes the analysis.

LEMMA 3.4. *There is a constant C such that*

$$|\mathrm{Ai}\,(-\xi)| \leqq C(1 \vee \xi)^{-1/4} \quad \textit{for all real } \xi.$$

Proof. We know that $|\mathrm{Ai}\,(z)| < 1$ for all z, and that there is a K such that

$$|\mathrm{Ai}\,(z)| \leqq K(1+|z|)^{-1/4} \quad \text{for all } z \leqq 0.$$

Choose $C = \max\{1, K\}$. If $\xi \leqq 1$, then $(1 \vee \xi) = 1$ and

$$|\mathrm{Ai}\,(-\xi)| \leqq 1 = (1 \vee \xi)^{-1/4} \leqq C(1 \vee \xi)^{-1/4}.$$

If $\xi > 1$, then $(1 \vee \xi) = \xi$ and

$$|\mathrm{Ai}\,(-\xi)| \leqq K(1+\xi)^{-1/4} \leqq C(1 \vee \xi)^{-1/4}$$

since $(1 \vee \xi)/(1+\xi) \leqq 1$ for $\xi > 1$. □

LEMMA 3.5. (a) *$F_2(x,t)$ is continuous in $\mathbb{R} \times (0, \infty)$.*

(b) *If $0 \leqq n \leqq N-1$, then $F_2(x,t) = o(x^{-n})$ as $x \to +\infty$.*

(c) *If $0 \leqq n \leqq N - 7/4$, then $\int_0^\infty x^n |F_2(x,t)|\, dx < \infty$.*

Proof. Because of the support of $F_2(x)$ we have

$$F_2(x,t) = (3t)^{-1/3} \int_0^\infty F_2(y)\, \mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right) dy. \tag{3.8}$$

The integrand is continuous in $(x,t) \in \mathbb{R} \times \mathbb{R}^+$ for each y. By its definition, $F_2(y) \in L^1(\mathbb{R})$. By (3.4) the integrand is bounded by $|F_2(y)|$. Thus (a) follows by Lebesgue's dominated convergence theorem.

For (b) we assume $0 \leqq n \leqq N-1$, keep $x \geqq 2$, and fix $t > 0$. Let $A(x,t)$ denote the part of the integral in (3.8) over $[0, x/2]$, and $B(x,t)$, the part over $[x/2, \infty)$. We need to show that $x^n A(x,t)$ and $x^n B(x,t)$ go to 0 as $x \to +\infty$. Now

$$\begin{aligned} x^n A(x,t) &= x^n \int_0^{x/2} F_2(y)\, \mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right) dy \\ &= x^n \int_{x/2(3t)^{1/3}}^{x/(3t)^{1/3}} \mathrm{Ai}\,(\xi) F_2(x-(3t)^{1/3}\xi)(3t)^{1/3}\, d\xi. \end{aligned}$$

Note $\xi > x/2(3t)^{1/3}$ implies $x \leqq 2(3t)^{1/3}\xi$. So

$$|x^n A(x,t)| \leqq 2^n (3t)^{(n+1)/3} \int_{x/2(3t)^{1/3}}^{x/(3t)^{1/3}} \xi^n |\mathrm{Ai}\,(\xi) F_2(x-(3t)^{1/3}\xi)|\,d\xi$$

$$\leqq 2^n (3t)^{(n+1)/3} \left[\int_{-\infty}^{\infty} |F(s)|^2\,ds\right]^{1/2} \left[\int_{x/2(3t)^{1/3}}^{x/(3t)^{1/3}} \xi^{2n} |\mathrm{Ai}\,(\xi)|^2\,d\xi\right]^{1/2}.$$

The decay rate (3.6) of Ai at $+\infty$ is such that $\xi^{2n}|\mathrm{Ai}\,(\xi)|^2$ is integrable on $\mathbb{R}^+$. Thus $x^n A(x,t) \to 0$ as $x \to +\infty$.

Next, since $|\mathrm{Ai}\,(s)| \leqq 1$ for all s,

$$|x^n B(x,t)| = \left| x^n \int_{x/2}^{\infty} F_2(y) \mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right) dy \right| \leqq 2^n \int_{x/2}^{\infty} y^n |F_2(y)|\,dy.$$

Since $F \in L^1_{N-1}(\mathbb{R}^+)$ and $n \leqq N-1$, $x^n B(x,t) \to 0$ as $x \to +\infty$. Thus (b) is proved. For (c) note that

$$\int_0^{\infty} x^n |F_2(x,t)|\,dx = \int_0^{\infty} x^n \left| (3t)^{-1/3} \int_{y=0}^{\infty} F_2(y)\,\mathrm{Ai}\left(\frac{y-x}{(3t)^{1/3}}\right) dy \right| dx$$

$$\leqq (3t)^{-1/3} \int_{x=0}^{\infty} x^n \int_{y=x}^{\infty} |F_2(y)| \left|\mathrm{Ai}\left(\frac{y-x}{(3t)^{1/3}}\right)\right| dy\,dx$$

$$+ (3t)^{-1/3} \int_{x=0}^{\infty} x^n \int_{y=0}^{x} |F_2(y)| \left|\mathrm{Ai}\left(\frac{y-x}{(3t)^{1/3}}\right)\right| dy\,dx.$$

Call these terms T_1 and T_2. It suffices to show T_1 and T_2 are finite. Now by Lemma 3.4

$$T_1 \leqq (3t)^{-1/3} \int_{x=0}^{\infty} x^n \int_{y=x}^{\infty} |F_2(y)| C\left(1 \vee \left\{\frac{y-x}{(3t)^{1/3}}\right\}\right)^{-1/4} dy\,dx$$

$$= C(3t)^{-1/3} \int_{y=0}^{\infty} |F_2(y)| \int_{x=0}^{y} x^n \left(1 \vee \frac{y-x}{(3t)^{1/3}}\right)^{-1/4} dx\,dy$$

$$\leqq C(3t)^{-1/3} \int_{y=0}^{\infty} |F_2(y)| \int_0^{y} x^n \left(\frac{y-x}{(3t)^{1/3}}\right)^{-1/4} dx\,dy.$$

It is easy to prove by induction on n that

$$\int_0^{y} x^n \left(\frac{y-x}{(3t)^{1/3}}\right)^{-1/4} dx \leqq (3t)^{-1/12} K_n y^{n+3/4}.$$

Thus there is a function $C = C(t)$ such that

$$T_1 \leqq C(t) \int_{y=0}^{\infty} |F_2(y)| y^{n+3/4}\,dy.$$

Since $n + \frac{3}{4} \leqq N-1$ and $F_2 \in L^1_{N-1}(\mathbb{R}^+)$, $T_1 < \infty$. Next

$$T_2 = (3t)^{-1/3} \int_{y=0}^{\infty} |F_2(y)| \int_{x=y}^{\infty} x^n \left|\mathrm{Ai}\left(\frac{y-x}{(3t)^{1/3}}\right)\right| dx\,dy.$$

Consider the inside integral $I(y)$ and let $\xi = (x-y)/(3t)^{1/3}$. Thus

$$I(y) = \int_{\xi=0}^{\infty} (y + (3t)^{1/3}\xi)^n |\mathrm{Ai}\,(\xi)| (3t)^{1/3}\,d\xi$$

$$\leqq 2^n (3t)^{1/3} \int_{\xi=0}^{\infty} (y^n + (3t)^{n/3}\xi^n) |\mathrm{Ai}\,(\xi)|\,d\xi \leqq C_1(t) y^n + C_2(t)$$

for positive functions $C_1(t)$, $C_2(t)$, since $\xi^j|\mathrm{Ai}\,(\xi)|\in L^1(\mathbb{R}^+)$ for all $j\geqq 0$. Now

$$T_2\leqq (3t)^{-1/3}\int_{y=0}^{\infty}|F_2(y)|(C_2(t)+y^nC_1(t))\,dy<\infty$$

because $n\leqq N-7/4<N-1$ and $F_2\in L^1_{N-1}(\mathbb{R}^+)$. □

Lemma 3.6. *Suppose* $0\leqq j\leqq 2N+1/2$. *Then*

(a) $\partial^j_x\partial_xF_2(x,t)$ *is continuous in* $\mathbb{R}\times(0,\infty)$,

(b) $\partial_xF_2(x,t)=o(x^{-n})$ *as* $x\to+\infty$ *for* $0\leqq n\leqq N$. *If* $j\geqq 1$, *then* $\partial^j_x\partial_xF_2(x,t)=o(x^{-n})$ *as* $x\to+\infty$ *for* $0\leqq n\leqq N+\frac{1}{4}-j/2$,

(c) $\int_0^\infty x^n|\partial^j_x\partial_xF_2(x,t)|\,dx<\infty$ *for* $0\leqq n\leqq N-\frac{3}{4}-j/2$.

Proof. From (3.3) we see

$$\partial_xF_2(x,t)=(3t)^{-2/3}\int_0^\infty F_2(y)\,\mathrm{Ai}'\left(\frac{y-x}{(3t)^{1/3}}\right)dy. \tag{3.9}$$

The continuity of $\partial_xF_2(x,t)$ follows from the Lebesgue dominated convergence theorem and the facts $F_2(y)\in L^1_{N-1}$ with $N\geqq 11/4$ and $\mathrm{Ai}'\,(\xi)=O(|\xi|^{+1/4})$ as $\xi\to-\infty$. To show continuity of $\partial^j_x\partial_xF_2$ with $j\geqq 1$, we integrate by parts in (3.9) getting

$$\partial_xF_2(x,t)=+(3t)^{-1/3}\int_0^\infty F_2'(y)\,\mathrm{Ai}\left(\frac{x-y}{(3t)^{1/3}}\right)dy. \tag{3.10}$$

Now fix $j\geqq 1$

$$\partial^j_xF_2(x,t)=(3t)^{-(j+1)/3}\int_0^\infty F_2'(y)\,\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)dy. \tag{3.11}$$

We need to show the continuity of $\partial^j_x\partial_xF_2$ at $x_0\in\mathbb{R}$, $t_0>0$. We keep $x\geqq x_0-1$, $t\geqq t_0/3$. The integrand is continuous in (x,t) for almost all y. Further it is uniformly bounded for $x\geqq x_0-1$:

$$\begin{aligned}\left|F_2'(y)\,\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)\right|&\leqq C|F_2'(y)|\left(1\vee\left\{\frac{y-x}{(3t)^{1/3}}\right\}\right)^{j/2-1/4}\\&\leqq C|F_2'(y)|\left(1\vee\left\{\frac{y+1-x_0}{t_0^{1/3}}\right\}\right)^{j/2-1/4}\end{aligned}$$

since $j/2-\frac{1}{4}>0$ when $j\geqq 1$. Further this bound is integrable on $\mathbb{R}^+$ since $F_2'\in L^1_N$ and the hypothesis on j implies $j/2-\frac{1}{4}\leqq N$. This completes (a).

For the remainder of the proof fix j so $0\leqq j\leqq 2N+\frac{1}{2}$. For part (b) we pick n so $0\leqq n\leqq N+\frac{1}{4}-j/2$ and keep $x\geqq 2$. From (3.11) we get

$$|x^n\partial^j_x\partial_xF_2(x,t)|\leqq(3t)^{-(j+1)/3}x^n\int_0^\infty\left|F_2'(y)\,\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)\right|dy.$$

Let J_1 and J_2 be the two terms obtained by splitting the integral at $y=x/2$. Note that $y>x/2$ implies $x^n\leqq 2^ny^n$.

$$\begin{aligned}J_1&=(3t)^{-(j+1)/3}x^n\int_0^{x/2}\left|F_2'(y)\,\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)\right|dy\\&=(3t)^{-j/3}x^n\int_{\xi=x/2(3t)^{1/3}}^{x/(3t)^{1/3}}|F_2'(x-(3t)^{1/3}\xi)\,\mathrm{Ai}^{(j)}\,(\xi)|\,d\xi\end{aligned}$$

after setting $\xi=(x-y)/(3t)^{1/3}$. Note $0\leqq y\leqq x/2$ implies $x^n\leqq 2^n(3t)^{n/3}\xi^n$. Thus

$$J_1\leqq(3t)^{-(n-j)/3}2^n\int_{x/2(3t)^{1/3}}^{x/(3t)^{1/3}}\xi^n|F_2'(x-(3t)^{1/3}\xi)\,\mathrm{Ai}^{(j)}\,(\xi)|\,d\xi.$$

By (3.6) we can find a constant K such that $|\xi^n \operatorname{Ai}^{(j)}(\xi)| \leqq K \mathrm{e}^{-\xi}$ for $\xi \geqq 0$ and in particular $|\xi^n \operatorname{Ai}^{(j)}(\xi)| \leqq K \exp(-x/2(3t)^{1/3})$ for $\xi \geqq x/2(3t)^{-1/3}$. Thus

$$J_1 \leqq (3t)^{(j-n)/2} 2^n K \, \mathrm{e}^{-x/2(3t)^{1/3}} \int_0^{x/2} |F_2'(y)| \, dy.$$

Since $F_2' \in L^1(\mathbb{R}^+)$, $J_1 \to 0$ as $x \to +\infty$.

It remains to deal with

$$J_2 = J_{2,j}(x) = (3t)^{-(j+1)/3} x^n \int_{x/2}^{\infty} \left| F_2'(y) \operatorname{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right) \right| dy.$$

If $j=0$ and $0 \leqq n \leqq N$, then

$$J_{2,0}(x) \leqq (3t)^{-1/3} 2^n \int_{y/2}^{\infty} y^n |F_2'(y)| \, dy,$$

which goes to 0 as $x \to +\infty$ because $F_2' \in L_N^1$.

Note that (3.6) implies that there is a constant A_j such that

$$|\operatorname{Ai}^{(j)}(-\xi)| \leqq A_j (1 \vee \xi)^{j/2-1/4}$$

for real ξ. Thus when $j \geqq 1$ we get

$$\begin{aligned} J_{2,j}(x) &\leqq (3t)^{-(j+1)/3} 2^n \int_{x/2}^{\infty} y^n |F_2'(y)| A_j \left(1 \vee \frac{y-x}{(3t)^{1/3}}\right)^{j/2-1/4} dy \\ &\leqq C_1(t) \int_{x/2}^{x+(3t)^{1/3}} y^n |F_2'(y)| \, dy + C_2(t) \int_{x+(3t)^{1/3}}^{\infty} y^n |F_2'(y)| (y-x)^{j/2-1/4} \, dy \end{aligned}$$

where $C_1(t)$ and $C_2(t)$ are positive functions of t. The first integral goes to 0 as $x \to +\infty$ because $n \leqq N - \frac{1}{4} < N$. In the second integral note $(y-x)^{j/2-1/4}$ is a decreasing function of x since $j \geqq 1$. Keeping $x \geqq 1$, the second integral is bounded by

$$\int_{x+(3t)^{1/3}}^{\infty} y^n |F_2'(y)| (y-1)^{j/2-1/4} \, dy,$$

which goes to zero as $x \to +\infty$ since we assume $n + j/2 - \frac{1}{4} \leqq N$ and know $F_2' \in L_N^1$. This finishes (b).

We finally turn to part (c). Keep $0 \leqq n \leqq N - \frac{3}{4} - j/2$ and note $n < N$ for all $j \geqq 0$. Now

$$\int_0^{\infty} x^n |\partial_x^j \partial_x F_2(x,t)| \, dx \leqq \int_{x=0}^{\infty} x^n \left(\int_{y=0}^{x} + \int_{y=x}^{\infty} \right) (3t)^{-(j+1)/3} \left| \operatorname{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right) F_2'(y) \right| dy \, dx.$$

Let K_1 denote the integral $\int_{x=0}^{\infty} \int_{y=0}^{x} \cdots dy \, dx$, and let K_2 denote the other. Set $\xi = (x-y)/(3t)^{1/3}$ in K_1. We get

$$\begin{aligned} K_1 &= (3t)^{-j/3} \int_{x=0}^{\infty} x^n \left(\int_{\xi=0}^{x/(3t)^{1/3}} |\operatorname{Ai}^{(j)}(\xi) F_2'(x-(3t)^{1/3}\xi)| \, d\xi \right) dx \\ &= (3t)^{-j/3} \int_{\xi=0}^{\infty} |\operatorname{Ai}^{(j)}(\xi)| \int_{x=(3t)^{1/3}\xi}^{\infty} x^n |F_2'(x-(3t)^{1/3}\xi)| \, dx \, d\xi \\ &= (3t)^{-j/3} \int_{\xi=0}^{\infty} |\operatorname{Ai}^{(j)}(\xi)| \int_{y=0}^{\infty} |F_2'(y)| (y+(3t)^{1/3}\xi)^n \, dy \, d\xi \\ &\leqq (3t)^{-j/3} 2^n \left[\int_{\xi=0}^{\infty} |\operatorname{Ai}^{(j)}(\xi)| \int_{y=0}^{\infty} y^n |F_2'(y)| \, dy \, d\xi \right. \\ &\qquad \left. + (3t)^{n/3} \int_{\xi=0}^{\infty} |\operatorname{Ai}^{(j)}(\xi)| \xi^n \int_{y=0}^{\infty} |F_2'(y)| \, dy \right] < \infty \end{aligned}$$

since $F_2' \in L^1_N$ with $N \geqq 11/4$ and $\mathrm{Ai}^{(j)}(\xi)$ has faster than exponential decay as $\xi \to +\infty$, and $|y+(3t)^{1/3}\xi|^n \leqq 2^n(y^n+(3t)^{n/3}\xi^n)$.

$$K_2=(3t)^{-(j+1)/3}\int_{x=0}^{\infty} x^n \int_{y=x}^{\infty} \left| F_2'(y)\,\mathrm{Ai}^{(j)}\left(\frac{x-y}{(3t)^{1/3}}\right)\right| \,dy\,dx.$$

By (3.6)

$$\begin{aligned} K_2 &\leqq (3t)^{-(j+1)/3}\int_{x=0}^{\infty} x^n \int_{y=x}^{\infty} |F_2'(y)| C_j^+ \left(1+\frac{y-x}{(3t)^{1/3}}\right)^{-j/2-1/4} dy\,dx \\ &\leqq C_j^+(3t)^{-(j+1)/3}\int_{y=0}^{\infty} |F_2'(y)| \int_{x=0}^{y} x^n \left(1+\frac{y-x}{(3t)^{1/3}}\right)^{j/2-1/4} dx\,dy. \end{aligned}$$

In case $j \geqq 1$ we have $j/2-\frac{1}{4}>0$. So for $x \geqq 1$

$$\begin{aligned} K_2 &\leqq C_j^+(3t)^{-(j+1)/3}\int_{y=0}^{\infty} |F_2'(y)| \int_{x=0}^{y} x^n \left(1+\frac{y-1}{(3t)^{1/3}}\right)^{j/2-1/4} dx\,dy \\ &\leqq C(3t)^{-(j+1)/3}\int_{y=0}^{\infty} |F_2'(y)|(1+y)^{n+1+j/2-1/4}\,dy<\infty \end{aligned}$$

since $n+\frac{3}{4}-j/2 \leqq N$. If $j=0$, then $j/2-\frac{1}{4}<0$ and this argument fails. However, if $j=0$ we get

$$K_2 \leqq C_j^+(3t)^{-1/3}\int_{y=0}^{\infty} |F_2'(y)| \int_{x=0}^{y} x^n \left(1+\frac{y-x}{(3t)^{1/3}}\right)^{-1/4} dx\,dy.$$

By induction one shows that the inner integral is $O(y^{n+(3/4)})$. Thus

$$K_2 \leqq C(3t)^{-1/3}\int_{y=0}^{\infty} |F_2'(y)|(1+y)^{n+3/4}\,dy<\infty$$

since we are assuming $n \leqq N-\frac{3}{4}$ when $j=0$. □

The results of Lemmas 3.5 and 3.6 rely on the fact that $\partial_x F_+(x,0)$ is in $L^1(+\infty)$. By contrast our next result does not use estimates on $\partial_x F_+(x,0)$. The next result is used by Kappeler in [9] where he considers KdV with certain measures as initial data. Up to this point we have used the Airy function strenuously; the rest of our results in this section rely on the type of analysis found in Cohen's paper [3]. Also by way of contrast, note that the distinction between generic and nongeneric data does not arise in the Airy function approach, whereas it does arise using the method of [3].

PROPOSITION 3.7. *Suppose that $U \in L^1_M(\mathbb{R})$ with $M \geqq 3$. If U is generic, set $N=M$; otherwise, set $N=M-1$. Let R_+ be the reflection coefficient of $U(x)$. Then a function*

$$F_+(x,t)=\pi^{-1}\int_{-\infty}^{\infty} R_+(k)\exp(2ikx+8ik^3t)\,dk$$

may be well defined on $\mathbb{R}\times(0,\infty)$ as an improper Riemann integral. Further for each fixed $t>0$, $F_+(x,t)$ is $(2N-1)$-times continuously differentiable with respect to x. Moreover for arbitrarily small $\varepsilon(0<\varepsilon \ll 1/2)$ there are functions $K_{0,N}(t)$ and $K_{1,N}(t)$ such that

$$\begin{aligned} &|F_+(x,t)| \leqq K_{0,N}(t)x^{-N+1+\varepsilon}, \\ &|\partial_x^j F_+(x,t)| \leqq K_{j,N}(t)x^{-N+j/2+\varepsilon} \quad \textit{for } 1 \leqq j \leqq 2N-1 \end{aligned}$$

whenever $x>12t>0$. $K_{j,N}(t)$ can be chosen nonincreasing, bounded as $t\uparrow+\infty$, and $O(t^{-j/2-\varepsilon})$ as $t\downarrow 0$.

Proof. Use Proposition 2.7 and a careful adaptation of the methods of [3]. □

3.2. Properties of $F_-(x, t)$. Recall that the crucial term in the kernel of the left-side Marchenko equation is

$$F_-(x,t)=\pi^{-1}\int_{-\infty}^{\infty} R_-(k)\exp(-2ikx-8ik^3t)\,dk.$$

Since $R_-(-k)=R_-^*(k)$, this may be rewritten as

$$F_-(x,t)=\pi^{-1}\int_{-\infty}^{\infty} R_-^*(k)\exp(2ikx+8ik^3t)\,dk.$$

Because $R_-^*(k)=V^*(k)/W(k)$, the analysis of R_+ is easily adapted to R_- and the regularity of F_- is the same as that of F_+. The decay of F_- and its derivatives as $x\to-\infty$ requires different treatment because there will be stationary points when $x<0$, namely $k=\pm(|x|/12t)^{1/2}$.

The purpose of this subsection is to identify conditions on $R_-(k)$ sufficient to verify the hypotheses on F_- in Kappeler's L^2 inverse scattering theorem [8]. The crucial point is to see when $\partial_x^j F_-(x,t)$ and $|x|^{1/2}\partial_x^j F_-(x,t)$ are in $L^2(-\infty, X)$ for finite X and $j=0,1,\cdots,4$. We formulate the results in two ways to allow some flexibility as to whether we ask R_- to have many derivatives of slower decay or fewer derivatives of faster decay.

This subsection will not be used until late in § 5.

LEMMA 3.8. *Suppose the function g has property $A(\lambda, N)$, namely*

$$g\in C^N(\mathbb{R}) \quad \textit{for } N\geqq 2,$$
$$g^{(n)}(k)=O(|k|^{-\lambda}) \quad \textit{as } |k|\to\infty \textit{ for } n=0,1,2,$$
$$g^{(n)}(k)=O(|k|^{-\lambda(n)}) \quad \textit{as } |k|\to\infty \textit{ for } 3\leqq n\leqq N$$

where

$$\lambda\geqq 1,\quad \lambda(n)\geqq\max\{1,\lambda-n\}\quad \textit{and}\quad \lambda\geqq\lambda(3)\geqq\cdots\geqq\lambda(N).$$

Let $G(x,t)$ be defined by

$$G(x,t)=\int_{-\infty}^{\infty} g(k)\,e^{8ik^3t+2ikx}\,dk.$$

If $N\geqq\lambda+3/2$, then for $t>0$

(i) $\partial_x^j G(x,t)=O(|x|^{-(\lambda-j)/2-1/4})$ *as $x\to-\infty$ for $j=0,1,2$.*

If $N>(\lambda+1)/2$, then for $t>0$ there is a δ such that $0<\delta\ll 1$ and

(ii) $\partial_x^j G(x,t)=O(|x|^{-1/2(\lambda-j)-\delta})$ *as $x\to-\infty$, $j=0,1,2$.*

Proof. This proof requires the careful extension and correction of the Appendix B of [13], i.e., a careful analysis by the method of stationary phase. □

Remark. Result (ii) gets a weaker result but requires less regularity in g for fixed λ. The following applications will be used in discussion of the L^2 inverse scattering problem in § 5.

Application 1(i). Suppose g satisfies $A(\lambda, N)$ with $5/2<\lambda\leqq 7/2$, $N=5$, and $\lambda(n)=1$ for $3\leqq n\leqq 5$. Then $N\geqq\lambda+3/2$. Part (i) of Lemma 3.8 tells us that for $t>0$

$$\partial_x G(x,t)=O(|x|^{-(\lambda-1)/2-1/4})=O(|x|^{-\lambda/2+1/4})\quad\text{as } x\to-\infty.$$

Since $-\lambda/2+\frac{1}{4}<-1$, it follows that both $\partial_x G(x,t)$ and $|x|^{1/2}\,\partial_x G(x,t)$ are in $L^2(-\infty)$.

Application 1(ii). Suppose g satisfies $A(\lambda, N)$ with $\lambda=3, N=3, \lambda(3)=1$. This requires more decay but fewer derivatives than the previous application. Note that $N>(\lambda+1)/2$. By part (ii) of Lemma 3.8, if $t>0$, then

$$\partial_x G(x, t)=O(|x|^{-(\lambda-1)/2-\delta})=O(|x|^{-1-\delta})$$

for some very small positive δ. Thus, in this case also, both $\partial_x G(x, t)$ and $|x|^{1/2}\,\partial_x G(x, t)$ are in $L^2(-\infty)$.

Application 2(i). Suppose that $g \in C^8(\mathbb{R})$ and that

$$g^{(n)}(k)=O(|k|^{-\lambda_0}) \quad \text{as } |k| \to \infty \text{ for } n=0,1,2,$$
$$g^{(n)}(k)=O(|k|^{-\lambda_0(n)}) \quad \text{as } |k| \to \infty \text{ for } 3 \leqq n \leqq 8$$

where

$$\lambda_0=\tfrac{11}{2}+\varepsilon, \qquad 0<\varepsilon<\tfrac{1}{2},$$
$$\lambda_0(n) \geqq \max\{1, \lambda_0-n\} \quad \text{for } 3 \leqq n \leqq 8,$$
$$4=\lambda_0(3) \geqq \lambda_0(4) \geqq \cdots \geqq \lambda_0(8).$$

Let $g_0=g$ and $g_1=g'$. Then it is easy to see that g_0 satisfies $A(\lambda, N)$ with $\lambda=\lambda_0$ and $N=8$. One can also verify that g_1 satisfies $A(\lambda, N)$ with $\lambda=\lambda_1=4$ and $N=7$. Since $8>\lambda_0+3/2$ we can apply Lemma 3.8(i) to get

$$\partial_x^j G_0(x, t)=O(|x|^{-(\lambda_0-j)/2-1/4}) \quad \text{as } x \to -\infty, \qquad j=0,1,2.$$

Since $7>\lambda_1+3/2$, we can also obtain

$$\partial_x^j G_1(x, t)=O(|x|^{-(\lambda_1-j)/2-1/4}) \quad \text{as } x \to -\infty, \qquad j=0,1,2.$$

Recall that

$$\partial_x^2 G_0(x, t)=\frac{1}{6it}\,G_1(x, t)+\frac{x}{3t}\,G_0(x, t),$$

$$\partial_x^3 G_0(x, t)=\frac{1}{6it}\,\partial_x^1 G_1(x, t)+\frac{1}{3t}\,G_0(x, t)+\frac{x}{3t}\,\partial_x G_0(x, t),$$

$$\partial_x^4 G_0(x, t)=\frac{1}{6it}\,\partial_x^2 G_1(x, t)+\frac{2}{3t}\,G_0(x, t)+\frac{x}{3t}\,\partial_x^2 G_0(x, t).$$

It follows that

$$\partial_x^3 G_0(x, t)=O(|x|^{-1-(3/4)}) \quad \text{as } x \to -\infty$$

and

$$\partial_x^4 G_0(x, t)=O(|x|^{-1-\varepsilon/2}) \quad \text{as } x \to -\infty.$$

We can conclude that for $j=0,1,\cdots,4$ both $\partial_x^j G(x, t)$ and $|x|^{1/2}\partial_x^j G(x, t)$ belong to $L^2(-\infty)$.

Application 2(ii). Suppose that g satisfies $A(\lambda, N)$ with $\lambda=6$, $N=4$, $\lambda(3)=4$, and $\lambda(4)=2$. Then since $4>(6+1)/2$, Lemma 3.8(ii) gives us

$$\partial_x^j G_1(x, t)=O(|x|^{-(6-j)/2-\delta}) \quad \text{as } x \to -\infty, \quad j=0,1,2$$

for some small positive δ. Also g' satisfies $A(\lambda, N)$ with $\lambda=4$ and $N=3$. Since $3>(4+1)/2$, Lemma 3.8(i) gives us

$$\partial_x^j G_1(x, t)=O(|x|^{-(4-j)/2-1/4}) \quad \text{as } x \to -\infty, \quad j=0,1,2.$$

Thus

$$\partial_x^j G(x,t) = O(|x|^{-3+j/2-\delta}) \quad \text{as } x \to -\infty, \quad j = 0, 1, 2,$$
$$\partial_x^3 G(x,t) = O(|x|^{-3/2-\delta}) \quad \text{as } x \to -\infty,$$
$$\partial_x^4 G(x,t) = O(|x|^{-1-\delta}) \quad \text{as } x \to -\infty.$$

It follows that for $j = 0, 1, \cdots, 4$ both $\partial_x^j G(x,t)$ and $|x|^{1/2}\,\partial_x^j G(x,t)$ are in $L^2(-\infty)$.

4. The regularity of the solutions of the Marchenko equation. The right-hand side Marchenko equation is

$$B_+(x,y) + \Omega_+(x+y) + \int_0^\infty B_+(x,z)\Omega_+(x+y+z)\,dz = 0. \tag{4.1}$$

It is well known [1], [6] that if $\Omega_+ \in L^1(+\infty) \cap L^\infty(+\infty)$ and if $\Omega_+' \in L_1^1(+\infty)$, then Ω_+ generates a compact operator from $L^1(\mathbb{R}^+)$ to $L^1(\mathbb{R}^+)$ and from $L^2(\mathbb{R}^+)$ to $L^2(\mathbb{R}^+)$ for each x by

$$\mathbf{\Omega}_{+x}[f](y) = \int_0^\infty f(z)\Omega_+(x+y+z)\,dz.$$

Theorems 4.1 and 4.2 are similar to results in [4]-[6]. They are stated here in the form used later.

THEOREM 4.1. *Suppose $n \geqq 1$. Suppose that $\Omega_+ \in C^{n+1}(\mathbb{R})$ and that for all finite X*

(i) $\int_X^\infty |\Omega_+'(s)|(1+|s|)\,ds < \infty$;

(ii) *If $0 \leqq k \leqq n$, then* $\int_X^\infty |\Omega_+^{(k)}(s)|\,ds < \infty$;

(iii) *If $0 \leqq k \leqq n+1$, then* $\sup\{|\Omega_+^{(k)}(X+s)|: s \geqq 0\} < \infty$;

(iv) $\int_X^\infty |\Omega_+^{(n+1)}(s)|^2\,ds < \infty$.

Then (4.1) *has a unique solution $B_+(x,\cdot)$ in $L^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)$. Further $x \mapsto B_+(x,\cdot)$ is $(n+1)$-times Frechet-differentiable as a function from $\mathbb{R}$ to $L^2(\mathbb{R}^+)$. Moreover $\partial_x^k B_+(x,y)$ is continuous in $\mathbb{R} \times [0,\infty)$ for $k \leqq n+1$. Finally if $k \leqq n$, then $\partial_x^k B_+(x,\cdot) \in L^1(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$.*

Proof. Consider first the case $n = 1$:

The existence of $B_+(x,\cdot)$ in $L^1(\mathbb{R}^+)$ is well known [1], [6]. We need to show (a) that $x \mapsto B_x(x,\cdot)$ is twice differentiable as a function from $\mathbb{R}$ to $L^2(\mathbb{R}^+)$, (b) that $\partial_x^k B_+(x,y)$ is continuous on $\mathbb{R} \times \mathbb{R}^+$ for $k = 0, 1, 2$, and (c) that $\partial_x^k B_+(x,\cdot) \in L^1(\mathbb{R}^+)$ for $k = 0, 1$.

Since $\Omega_+ \in L^1(+\infty)$ it is easy to check that $x \mapsto \mathbf{\Omega}_{+x}$ is continuous in the uniform operator norm on both $L^1(\mathbb{R}^+)$ and $L^\infty(\mathbb{R}^+)$. Thus $(\mathbf{I} + \mathbf{\Omega}_{+x})^{-1}$ also depends continuously on x in the uniform norm in $\mathscr{L}(L^1(\mathbb{R}^+))$. It is also easy to see that $\mathbf{\Omega}_{+x}$ maps $L^1(\mathbb{R}^+)$ into $L^\infty(\mathbb{R}^+)$:

$$\sup_{y \geqq 0} |\mathbf{\Omega}_{+x}[g](y)| \leqq \sup_{y \geqq 0} \int_0^\infty |g(z)|\,|\Omega_+(x+y+z)|\,dz \leqq \sup_{s > x} |\Omega_+(s)| \cdot \int_{z=0}^\infty |g(z)|\,dz.$$

It now follows from (4.1) that $B_+(x,\cdot)$ is in $L^\infty(\mathbb{R}^+)$ as well as in $L^1(\mathbb{R}^+)$. To see the continuity of $B_+(x,y)$ we note

$$\begin{aligned} B_+(x_1,y_1) - B_+(x_2,y_2) = {} & -\Omega_+(x_1,y_1) + \Omega_+(x_2,y_2) \\ & - \int_0^\infty B_+(x_1,z)\{\Omega_+(x_1+y_1+z) - \Omega_+(x_2+y_2+z)\}\,dz \\ & - \int_0^\infty \{B_+(x_1,z) - B_+(x_2,z)\}\Omega_+(x_2+y_2+z)\,dz. \end{aligned}$$

Thus

$$
\begin{aligned}
|B_+(x_1, y_1) - B_+(x_2, y_2)| \leqq |\Omega_+(x_1+y_1) - \Omega_+(x_2+y_2)| \\
+ \sup_{s \geqq s_0} |\Omega'_+(s)|\,|x_1+y_1-x_2-y_2|\, \|B_+(x_1, \cdot)\|_{L^1} \\
+ \sup_{s \geqq s_0} |\Omega_+(s)| \cdot \|B_+(x_1, \cdot) - B_+(x_2, \cdot)\|_{L^1}
\end{aligned}
$$

where $s_0 \leqq \min\{x_1+y_1, x_2+y_2\}$. The continuity of $B_+(x, y)$ as a map $\mathbb{R} \times \mathbb{R}^+ \to R$ now follows easily.

For the rest of this section we will omit the subscripts "+" from B_+, Ω_+, and $\mathbf{\Omega}_+$. Where we intend B_- and Ω_-, the subscript "−" will appear.

Next we ask whether $x \mapsto B(x, \cdot)$ is differentiable as a map $\mathbb{R} \mapsto L^1(\mathbb{R}^+)$.

For $h \neq 0$, set

$$
\Phi_h(x, y) = \frac{\Omega(x+y+h) - \Omega(x+y)}{h} + \int_{z=0}^{\infty} B(x, z) \frac{\Omega(x+y+z+h) - \Omega(x+y+z)}{h}\, dz
$$

and for $h = 0$, set

$$
\Phi_0(x, y) = \Omega'(x+y) + \int_0^{\infty} B(x, z)\Omega'(x+y+z)\, dz.
$$

Note that

$$
\frac{B(x+h, y) - B(x, y)}{h} = -(\mathbf{I} + \mathbf{\Omega}_{(x+h)})^{-1}[\Phi_h(x, \cdot)](y).
$$

Clearly $\Phi_h(x, \cdot) \in L^1(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ for all h. Further $\Phi_h(x, \cdot) \mapsto \Phi_0(x, \cdot)$ in both $L^1(\mathbb{R}^+)$ and $L^\infty(\mathbb{R}^+)$ as $h \to 0$. Thus

$$
\lim_{h \to \infty} \frac{B(x+h, \cdot) - B(x, \cdot)}{h} = -(\mathbf{I} + \mathbf{\Omega}_x)^{-1}[\Phi_0(x, \cdot)] \quad \text{in } L^1(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+).
$$

Thus

$$
\begin{aligned}
\partial_x^1 B(x, y) + \int_{z=0}^{\infty} \partial_x^1 B(x, z)\Omega(x+y+z)\, dz &= -\Phi_0(x, y) \\
&= -\Omega'(x+y) - \int_0^{\infty} B(x, z)\Omega'(x+y+z)\, dz.
\end{aligned}
$$

It now follows that $\partial_x^1 B(x, y)$ depends continuously on x and y.

Finally consider the map $x \mapsto \partial_x^1 B(x, \cdot)$ as going from $\mathbb{R}$ to $L^2(\mathbb{R}^+)$. We must show that it is differentiable. Write $B^{(1,0)}(x, y)$ for $\partial_x^1 B(x, y)$. For $h \neq 0$, set

$$
\psi_h(x, \cdot) = -(\mathbf{I} + \mathbf{\Omega}_{(x+h)}) \left[\frac{B^{(1,0)}(x+h, \cdot) - B^{(1,0)}(x, \cdot)}{h} \right].
$$

Since all $B^{(1,0)}(x, \cdot)$ are in $L^1(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ it follows that $\psi_h(x, \cdot)$ is in $L^1(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$. Computation shows that

$$
\begin{aligned}
\psi_h(x, y) = {} & \frac{\Omega'(x+y+h) - \Omega'(x+y)}{h} + \int_0^{\infty} \frac{B(x+h, z) - B(x, z)}{h} \Omega'(x+y+h+z)\, dz \\
& + \int_0^{\infty} B^{(1,0)}(x, z) \frac{\Omega(x+y+h+z) - \Omega(x+y+z)}{h}\, dz \\
& + \int_0^{\infty} B(x, z) \frac{\Omega'(x+y+h+z) - \Omega'(x+y+z)}{h}\, dz.
\end{aligned}
$$

For $h=0$ set

$$\psi_0(x, y)=\Omega''(x+y)+2\int_0^\infty B^{(1,0)}(x, z)\Omega'(x+y+z)\,dz+\int_0^\infty B(x, z)\Omega''(x+y+z)\,dz.$$

Our hypotheses and earlier results tell us that

$$\psi_0(x,\cdot)\in L^2(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+).$$

Now we verify that $\psi_h(x,\cdot)\to\psi_0(x,\cdot)$ in $L^2(\mathbb{R}^+)$. As a function of y in $\mathbb{R}^+$, $\{\Omega'(x+y+h)-\Omega'(x+y)\}/h$ converges to $\Omega''(x+y)$ in $L^2(\mathbb{R}^+)$ as $h\to 0$. The remaining three terms in $\psi_h(x, y)$ are essentially convolutions. It is straightforward to verify the convergence of the factors in these convolutions in $L^2(\mathbb{R}^+)$. Thus, still in $L^2(\mathbb{R}^+)$, $\psi_h(x,\cdot)\to\psi_0(x,\cdot)$ as $h\to 0$. Thus

$$\begin{aligned}\lim_{h\to 0}\frac{B^{(1,0)}(x+h,\cdot)-B^{(1,0)}(x,\cdot)}{h}&=-\lim_{h\to 0}(\mathbf{I}+\mathbf{\Omega}_{(x+h)})^{-1}\Psi_h(x,\cdot)\\&=-(\mathbf{I}+\mathbf{\Omega}_x)^{-1}\psi_0(x,\cdot).\end{aligned}$$

So $\partial_x^2B(x,\cdot)$ exists in $L^2(\mathbb{R}^+)$ and satisfies

$$\begin{aligned}\partial_x^2B(x, y)+\int_0^\infty\partial_x^2B(x, z)\Omega(x+y+z)\,dz=-\Omega''(x+y)-2\int_0^\infty B^{(1,0)}(x, z)\Omega'(x+y+z)\,dz\\-\int_0^\infty B(x, z)\Omega''(x+y+z)\,dz.\end{aligned}$$

The continuity of $\partial_x^2B(x, y)$ follows from analysis of this equation.

This proves the theorem for $n=1$. The method extends in the obvious way to cases where $n>1$. □

THEOREM 4.2. *Suppose that $\Omega(x, t)$ has the following properties*:

(i) *For fixed $t>0$, $\Omega(x, t)$ is a C^1 function of x, and*

$$\int_X^\infty|\partial_x\Omega(x, t)|(1+|x|)\,dx<\infty\quad\textit{for all finite } X.$$

(ii) *The mapping $t\mapsto\Omega(\cdot, t)$ is differentiable both as a map from $(0,\infty)$ to $L^1(+\infty)$ and from $(0,\infty)$ to $L^\infty(+\infty)$; $\int_X^\infty|\partial_t\Omega(x, t)|\,dx<\infty$, for finite X.*

(iii) *The mapping $t\mapsto\Omega(\cdot, t)$ is differentiable as a map from $(0,\infty)$ to $L^2(+\infty)$, $\int_X^\infty|\partial_t\partial_x\Omega(x, t)|^2\,dx<\infty$ for finite X.*

(iv) *For fixed $t>0$, the functions $\Omega(x, t)$, $\partial_x\Omega(x, t)$, $\partial_t\Omega(x, t)$, and $\partial_t\partial_x\Omega(x, t)$ are in $L^\infty(+\infty)$.*

(v) *The functions mentioned in* (iv) *are continuous on $\mathbb{R}\times(0,\infty)$. For each $t>0$, let $B(x,\cdot, t)$ denote the solution of*

$$B(x, y, t)+\Omega(x+y, t)+\int_0^\infty B(x, z, t)\Omega(x+y+z, t)\,dz=0,$$

which is the Marchenko equation with $\Omega=\Omega(x, t)$. Then

(a) *The map $t\mapsto B(x,\cdot, t)$ is differentiable both as a map $(0,\infty)\to L^1(\mathbb{R}^+)$ and as a map $(0,\infty)\to L^2(\mathbb{R}^+)$. Further both $B(x, y, t)$ and $\partial_tB(x, y, t)$ are continuous in $\mathbb{R}\times[0,\infty)\times(0,\infty)$.*

(b) *The map $t\mapsto\partial_xB(x,\cdot, t)$ is differentiable as a map $(0,\infty)\to L^2(\mathbb{R}^+)$, and $\partial_t\partial_xB(x, y, t)$ is continuous in $\mathbb{R}\times[0,\infty)\times(0,\infty)$.*

Proof. By hypotheses we have $\Omega(x, t) \in L^1(+\infty)$ and $\partial_x \Omega(x, t) \in L_1^1(+\infty)$ for each fixed $t > 0$. Therefore the solutions $B(x, \cdot, t)$ exist in $L^1(\mathbb{R}^+)$. The continuity of $B(x, y, t)$ follows immediately, as does the existence and continuity of $\partial_x B(x, y, t)$.

Let $\mathbf{\Omega}_x^t$ denote the operator $\mathbf{\Omega}_x^t[g](y) = \int_0^\infty \Omega_+(x+y+z, t) g(z)\, dz$.

We use again the methods of the previous theorem. For $h \neq 0$, one gets

$$\frac{B(x, y, t+h) - B(x, y, t)}{h} + \int_{z=0}^\infty \frac{B(x, z, t+h) - B(x, z, t)}{h} \Omega(x+y+z, t+h)\, dz$$
$$= -\Phi_h(x, y, t)$$

where

$$\Phi_h(x, y, t) \equiv \frac{\Omega(x+y, t+h) - \Omega(x+y, t)}{h}$$
$$+ \int_0^\infty B(x, z, t) \left\{ \frac{\Omega(x+y+z, t+h) - \Omega(x+y+z, t)}{h} \right\} dz.$$

We have assumed that the map $t \mapsto \Omega(\cdot, t)$ is differentiable in $L^1(+\infty)$. Therefore as $h \to 0$, $\Phi_h(x, \cdot, t)$ converges in $L^1(\mathbb{R}^+)$ to

$$\Phi_0(x, y, t) \equiv \partial_t \Omega(x+y, t) + \int_0^\infty B(x, z, t) \partial_t \Omega(x+y+z, t)\, dz.$$

It is easy to see that $\Phi_h(x, \cdot, t) \to \Phi_0(x, \cdot, t)$ also in $L^\infty(\mathbb{R}^+)$, whence in $L^2(\mathbb{R}^+)$ as well. Now we have

$$\frac{B(x, \cdot, t+h) - B(x, \cdot, t)}{h} = (\mathbf{I} + \mathbf{\Omega}_x^{t+h})^{-1}[-\Phi_h(x, \cdot, t)].$$

The operator $(\mathbf{I} + \mathbf{\Omega}_x^{t+h})^{-1}$ depends continuously on h in the operator norms on both $L^1(\mathbb{R}^+)$ and $L^2(\mathbb{R}^+)$. So

$$\lim_{h \to 0} \frac{B(x, \cdot, t+h) - B(x, \cdot, t)}{h} = (\mathbf{I} + \mathbf{\Omega}_x^t)^{-1}[-\Phi_0(x, \cdot, t)]$$

in both spaces; equivalently $\partial_t B(x, \cdot, t)$ exists in $L^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)$ and satisfies

$$\partial_t B(x, y, t) = -\Omega_t(x+y, t) - \int_0^\infty B_t(x, z, t) \Omega(x+y+z, t)\, dz$$
$$- \int_0^\infty B(x, z, t) \Omega_t(x+y+z, t)\, dz.$$

From this it follows that $\partial_t B(x, \cdot, t)$ is in $L^\infty(\mathbb{R}^+)$ and that $\partial_t B(x, y, t)$ is continuous in $\mathbb{R} \times [0, \infty) \times (0, \infty)$.

Next we study the map $t \mapsto \partial_x B(x, \cdot, t)$. Set

$$-\Psi_h(x, \cdot, t) = (\mathbf{I} + \mathbf{\Omega}_x^{t+h})^{-1} \left[\frac{B(x, \cdot, t+h) - B(x, \cdot, t)}{h} \right].$$

Computation shows that

$$\Psi_h(x, y, t) = \{\Omega^{(1,0)}(x+y, t+h) - \Omega^{(1,0)}(x+y, t)\}/h$$
$$+ \int_0^\infty B_x(x, z, t)\{\Omega(x+y+z, t+h) - \Omega(x+y+z, t)\} h^{-1}\, dz$$
$$+ \int_0^\infty B(x, z, t)\{\Omega^{(1,0)}(x+y+z, t+h) - \Omega^{(1,0)}(x+y+z, t)\} h^{-1}\, dz$$

$$+\int_0^\infty \{B(x,z,t+h)-B(x,z,t)\}h^{-1}\Omega^{(1,0)}(x+y+z,t+h)\,dz.$$

Clearly as $h\to 0$, we get the convergence $\Psi_h(x,\cdot,t)\to\Psi_0(x,\cdot,t)$ in $L^2(\mathbb{R}^+)$, where

$$\begin{aligned}\Psi_0(x,y,t)=\partial_t\partial_x\Omega(x+y,t)&+\int_0^\infty B_x(x,z,t)\Omega(x+y+z,t)\,dz\\&+\int_0^\infty B(x,z,t)\Omega_{x,t}(x+y+z,t)\,dz\\&+\int_0^\infty B_t(x,z,t)\Omega_x(x+y+z,t)\,dz.\end{aligned}$$

Thus $t\mapsto B_x(x,\cdot,t)$ is differentiable in $L^2(\mathbb{R}^+)$ and

$$\partial_t\partial_x B(x,y,t)+\int_0^\infty \partial_t\partial_x B(x,z,t)\Omega(x+y+z,t)\,dz=-\psi_0(x,y,t)$$

whence $\partial_t\partial_x B(x,y,t)$ is continuous and belongs to $L^\infty(\mathbb{R}^+)$ as a function of y. □

We next investigate the consequences of a stronger decay assumption on $\Omega'(x)$, namely that there is an $\alpha\geqq 1$ such that

$$\int_X^\infty |\Omega'(x)|(1+|x|^\alpha)\,dx<\infty \quad \text{for all } X \text{ in } \mathbb{R}.$$

We make use of an inequality of Faddeev's [6, p. 160]

$$|\Omega(x)+\partial_x B(x,0)|\leqq C(x)\left[\int_x^\infty |\Omega'(t)|\,dt\right]^2 \tag{4.2}$$

where $C(x)$ is a nonincreasing function of x.

LEMMA 4.3. *Suppose that $\Omega\in L^1(+\infty)$ and $\Omega'\in L^1_\alpha(+\infty)$ with $\alpha\geqq 1$. Let $B(x,y)$ be the solution of* (4.1) *and set $u(x)=-\partial_x B(x,0)$. Then $u\in L^1_\alpha(+\infty)$.*

Proof. Because of (4.2) it suffices to prove that

$$Q\equiv\int_{y=x}^\infty y^\alpha\left(\int_{t=y}^\infty |\Omega'(t)|\,dt\right)^2 dy<\infty$$

for $x\geqq 0$.

Now

$$\begin{aligned}Q&=\int_{y=x}^\infty y^\alpha\left(\int_{t=y}^\infty |\Omega'(t)|\,dt\right)\left(\int_{s=y}^\infty |\Omega'(s)|\,ds\right)dy\\&=\int_{t=x}^\infty |\Omega'(t)|\int_{s=x}^t |\Omega'(s)|\int_{y=x}^s y^\alpha\,dy\,ds\,dt+\int_{t=x}^\infty |\Omega'(t)|\int_{s=t}^\infty |\Omega'(s)|\int_{y=x}^t y^\alpha\,dy\,ds\,dt\\&\leqq\frac{1}{\alpha+1}\int_{t=x}^\infty |\Omega'(t)|t^{(\alpha+1)/2}\int_{s=x}^\infty \Omega'(s)|s^{(\alpha+1)/2}\,ds\,dt\\&<\infty\end{aligned}$$

since $(\alpha+1)/2\leqq\alpha$. □

5. The existence theorem for KdV; properties of the solution. In this section we describe the properties of the function $u(x,t)$ constructed by the inverse scattering method and establish the sense in which it solves the problem (1.1), (1.2).

THEOREM 5.1. *Suppose that $U\in L_1^1(\mathbb{R})$ and that $U\in L_N^1(\mathbb{R}^+)$ for some $N\geqq 11/4$. Then there is a classical solution $u(x,t)$ of* KdV *in $t>0$ such that*

(i) *$\partial_x^j\partial_t^k u(x,t)$ is continuous in x for each positive t when $0\leqq j+3k\leqq 2N-3$;*

(ii) *$u(\cdot,t)\to U$ in $H^{-1}(+\infty)$ as $t\to 0$.*

(iii) *$x^n\partial_x^j u(x,t)\to 0$ as $x\to+\infty$ for $0\leqq n\leqq N+\frac{1}{4}-(j+1)/2$.*

Proof. For each fixed positive t we consider the Marchenko equation

$$B_+(x,y,t)+\Omega_+(x+y,t)+\int_{z=0}^{\infty}\Omega_+(x+y+z,t)B_+(x,z,t)\,dz=0 \tag{5.1}$$

where

$$\Omega_+(x,t)=F_+(x,t)+2\sum_{j\in J}c_{+j}\,e^{-2\kappa_j x+8\kappa_j^3 t}$$

and

$$F_+(x,t)=\pi^{-1}\int_{-\infty}^{\infty}R_+(k)\,e^{8ik^3t+2ikx}\,dk.$$

This $F_+(x,t)=F_1(x,t)+F_2(x,t)$ as in § 3. By appealing to Lemmas 3.3, 3.4, 3.5 and to the hypothesis $N\geqq 11/4$ we conclude that

(a) $\partial_x\Omega_+(x,t)\in L_1^1(+\infty)$,

(b) $\partial_x^\nu\Omega_+(x,t)\in L^1(+\infty)$ for $0\leqq\nu\leqq 2N-\frac{1}{2}$,

(c) $\partial_x^\nu\Omega_+(x,t)\in L^\infty(+\infty)$ for $0\leqq\nu\leqq 2N+\frac{1}{2}$, and

(d) $\partial_x^{\nu+1}\Omega_+(x,t)\in L^2(+\infty)$ for $0\leqq\nu\leqq 2N-3/2$.

Note that $N\geqq 11/4$ implies that $2N-3/2\geqq 4$. Thus our kernel $\Omega_+(x,t)$ satisfies the hypotheses of Theorem 4.1 with $1\leqq n\leqq 2N-\frac{3}{4}$. So we obtain a solution $B_+(x,y,t)$ to (5.1) such that

$$B_+(x,\cdot,t)\in L^1(\mathbb{R}^+)\cap L^2(\mathbb{R}^+).$$

$$\partial_x^\nu B_+(x,y,t)\text{ is continuous for }(x,y)\text{ in }\mathbb{R}\times(0,\infty)\text{ if }\nu\leqq n+1.$$

$$\partial_x^\nu B_+(x,\cdot,t)\in L^1(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+)\quad\text{if } 0\leqq\nu\leqq n.$$

Let $u(x,t)=-\partial_x B_+(x,0,t)$. We must now show that $u(x,t)$ is the desired solution of KdV.

In addition to properties (a)-(d) of Ω_+ we know that in the distribution sense

$$\partial_t\Omega_+(x,t)+\partial_x^3\Omega_+(x,t)=0\quad\text{for } t>0,\quad x\in\mathbb{R}.$$

Since $N\geqq 11/4$, and thus $2N-1\geqq 4$ we conclude that $\Omega_+(x,t)$ and $\partial_x\Omega_+(x,t)$ are continuously differentiable with respect to t, and that

$$\partial_t\Omega_+(\cdot,t)=-\partial_x^3\Omega_+(\cdot,t)\in L^1(+\infty)\cap L^\infty(+\infty),$$

$$\partial_t\partial_x\Omega_+(\cdot,t)=-\partial_x^4\Omega_+(\cdot,t)\in L^2(+\infty)\cap L^\infty(+\infty).$$

In order to apply Theorem 4.2 we need finally to check that $t\mapsto\Omega_+(\cdot,t)$ is differentiable in $L^\infty(+\infty)$ for $t>0$; but this follows from the continuity and decay rate of $\Omega_t=-\Omega_{xxx}$.

By applying Theorem 4.2 we now learn that

$$\partial_t B_+(x,\cdot,t)\in L^1(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+),$$

$$\partial_t\partial_x B_+(x,\cdot,t)\in L^2(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+),$$

and further that all $\partial_x^j\partial_t^k B_+(x,y,t)$ are continuous in $\mathbb{R}\times[0,\infty)\times(0,\infty)$ for $j+3k\leqq 2N-2$.

For fixed positive t, it is clear that u has the regularity (i) and the decay rate (iii). The proof that this function $u(x, t)$ satisfies the KdV equation (1.1) in $t>0$ follows Tanaka's argument in [17]. The condition $N \geqq 11/4$ gives enough regularity to justify the formal argument.

To prove (ii) we show that $u(x, t) \to U(x)$ in $H^{-1}([X, \infty))$ as $t \to 0$ for each finite X. Since $B_+(x, 0, t) = \int_x^\infty u(s, t)\, ds$ for $t \geqq 0$ we must show that $B_+(x, 0, t) \to B_+(x, 0, 0)$ in $L^2([X, \infty))$ as $t \to 0$. From the Marchenko equation (4.1) we obtain

$$B_+(x, 0, t) - B_+(x, 0, 0) = -Q_1(x, t) - Q_2(x, t) - Q_3(x, t)$$

where

$$Q_1(x, t) = \Omega_+(x, t) - \Omega_+(x, 0),$$

$$Q_2(x, t) = \int_0^\infty \{B_+(x, z, t) - B_+(x, z, 0)\}\Omega_+(x+z, t)\, dz,$$

$$Q_3(x, t) = \int_0^\infty B_+(x, z, 0)\{\Omega_+(x+z, t) - \Omega_+(x+z, 0)\}\, dz.$$

One easily sees that $Q_1(x, t) \to 0$ in $L^2([X, \infty))$ as $t \to 0$.

Next we show $Q_3(x, t) \to 0$ in $L^2([X, \infty))$ as $t \to 0$ by showing that

$$\|Q_3(\cdot, t)\|_{L^2([X,\infty))} \leqq C\|\Omega_+(\cdot, t) - \Omega_+(\cdot, 0)\|_{L^2([X,\infty))}. \tag{5.2}$$

For any $h \in L^2([X, \infty))$

$$\begin{aligned}\left|\int_X^\infty h(x) Q_3(x, t)\, dx\right| &= \left|\int_{x=X}^\infty h(x) \int_{s=x}^\infty B_+(x, s-x, 0)\{\Omega_+(s, t) - \Omega_+(s, 0)\}\, ds\, dx\right| \\ &\leqq \int_{s=X}^\infty |\Omega_+(s, t) - \Omega_+(s, 0)| \left\{\int_{x=X}^s |h(x) B_+(x, s-x, 0)| dx\right\} ds \\ &\leqq \|\Omega_+(\cdot, t) - \Omega_+(\cdot, 0)\|_{L^2([X,\infty))} \\ &\quad \cdot \left\{\int_{s=X}^\infty \left(\int_{x=X}^s |h(x) B_+(x, s-x, 0)|\, dx\right)^2 ds\right\}^{1/2}.\end{aligned}$$

Letting Φ denote the second factor, we have

$$\begin{aligned}\Phi &\leqq \left\{\int_{s=X}^\infty \left(\int_{x=X}^s |h(x)|^2\, dx\right)\left(\int_{x=X}^s |B_+(x, s-x, 0)|^2\, dx\right) ds\right\}^{1/2} \\ &\leqq \|h\|_{L^2([X,\infty))} \left\{\int_{s=X}^\infty \int_{x=X}^s |B_+(x, s-x, 0)|^2\, dx\, ds\right\}^{1/2}.\end{aligned}$$

From Tanaka's paper [17, § 1] we have

$$|B_+(x, y, 0)| \leqq K \int_{x+y}^\infty |U(z)|\, dz \quad \text{for all } x \geqq X, \quad y \geqq 0,$$

whence

$$|B_+(x, s-x, 0)| \leqq K \int_x^\infty |U(z)|\, dz \quad \text{for all } x \geqq X, \quad s \geqq x.$$

Thus

$$\Phi \leqq \|h\|_{L^2([X,\infty))} \left\{\int_{s=X}^\infty \int_{x=X}^s \left(K \int_{z=s}^\infty |U(z)|\, dz\right)^2 dx\, ds\right\}^{1/2}$$

whence $\Phi \leqq C\|h\|_{L^2([X,\infty))}$ where

$$C=\frac{1}{2}K\left\{\int_X^\infty |U(w)|\,dw \int_{z=X}^\infty (z-X)^2|U(z)|\,dz\right\}^{1/2}<\infty.$$

Now (5.2) follows since h was arbitrary.

Finally we look at $Q_2(x,t)$: One finds

$$\|Q_2(x,t)\|^2_{L^2([X,\infty))} \leqq \left\{\int_{x=X}^\infty \int_{z=0}^\infty |B_+(x,z,t)-B_+(x,z,0)|^2\,dz\,dx\|Q_1(\cdot,t)\|^2_{L^2([X,\infty))}.\right.$$

The second factor is bounded as $t\to 0$, so we look at the first factor, $\Psi(t)$.

$$B_+(x,z,t)-B_+(x,z,0)=-Q_4(x,z,t)-Q_5(x,z,t)$$

where

$$Q_4(x,\cdot,t)=(\mathbf{I}+\mathbf{\Omega}_x^t)^{-1}[\Omega_+(x+[\cdot],t)-\Omega_+(x+[\cdot],0)]$$

and

$$Q_5(x,\cdot,t)=[(\mathbf{I}+\mathbf{\Omega}_x^t)^{-1}-(\mathbf{I}+\mathbf{\Omega}_x^0)^{-1}]\Omega_+(x+[\cdot],0).$$

Thus

$$\psi(t)=\int_X^\infty \|Q_4(x,\cdot,t)+Q_5(x,\cdot,0)\|^2_{L^2(\mathbb{R}^+)}\,dx \leqq 2\int_X^\infty (\|Q_4(x,\cdot,t)\|^2+\|Q_5(x,\cdot,t\|^2)\,dx.$$

There is a bound M such that

$$\|(\mathbf{I}+\mathbf{\Omega}_x^t)^{-1}\|_{op,L^2(\mathbb{R}^+)} \leqq M \quad \text{for } X\leqq x\leqq\infty, \quad 0\leqq t\leqq 1$$

because the operator depends continuously on (x,t) and the kernel decays fast enough as $x\to+\infty$. Thus

$$\|Q_4(x,\cdot,t)\|^2_{L^2(\mathbb{R}^+)} \leqq M\int_x^\infty |\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds$$

and

$$\int_X^\infty \|Q_4(x,\cdot,t)\|^2\,dx \leqq M\int_X^\infty (s-X)|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds.$$

By the form of $\Omega_+, \Omega_+=F_++G_+$ in §3.1, we need only show

$$\int_X^\infty (s-X)|F_+(s,t)-F_+(s,0)|^2\,ds\to 0 \quad \text{as } t\to 0.$$

We already know $F_+(s,t)\to F_+(s,0)$ in $L^2(\mathbb{R})$ so it suffices to show $\int_{X+1}^\infty s^2|F_+(s,t)-F_+(s,0)|^2\,ds\to 0$. By Proposition 2.5 we know $R_+\in C^1(\mathbb{R})$ and $R_+(k)=O(k^{-1})$ as $k\to\pm\infty$. We may therefore follow Kappeler's proof of Theorem 3.1(iv) in [9] to conclude that $sF_+(s,t)\to sF_+(s,0)$ in $L^2(+\infty)$ as $t\to 0$. It remains to make $\int_X^\infty \|Q_5(x,\cdot,t)\|^2\,dx\to 0$ as $t\to 0$. But, using L^2 norms on $\mathbb{R}^+$,

$$\begin{aligned}\|Q_5(x,\cdot,t)\| &\leqq \|(\mathbf{I}+\mathbf{\Omega}_x^t)^{-1}(\mathbf{\Omega}_x^0-\mathbf{\Omega}_x^t)(\mathbf{I}+\mathbf{\Omega}_x^0)^{-1}\|_{op}\|\Omega_+(x+[\cdot],t)\| \\ &\leqq M^2\|\mathbf{\Omega}_x^0-\mathbf{\Omega}_x^t\|_{op}\|\Omega_+(x+[\cdot],t)\|.\end{aligned}$$

The last factor is bounded as $t \to 0$ so

$$\begin{aligned}\int_X^\infty \|Q_5(x,\cdot,t)\|^2\,dx &\leqq C\int_X^\infty \|\Omega_x^0-\Omega_x^t\|_{op}^2\,dx\\ &\leqq C\int_X^\infty\int_x^\infty |\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\,dx\\ &\leqq C\int_X^\infty (s-X)|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds,\end{aligned}$$

which we have already seen goes to 0 with t.

This completes the proof of Theorem 5.1. □

Under certain additional hypotheses we can also study $u(x,t)$ as $x \to -\infty$.

THEOREM 5.2. (i) *Assume that $U \in L_N^1(\mathbb{R})$, that $N \geqq 5$, and that $R_-^{(n)}(k) = O(|k|^{-\lambda})$ as $k \to \pm\infty$ for $n = 0, 1, 2$ and some $\lambda > 5/2$. Let $u(x,t)$ be the solution to* KdV *with initial profile U in the sense of Theorem* 5.1. *Then $u(\cdot,t)$ evolves in $L^2(\mathbb{R})$ for $t > 0$. If $N \geqq \lambda + 2$, then $\int_{-\infty}^\infty |u(x,t)|^2|x|^{2\alpha}\,dx < \infty$ for any α with $\lambda - 3/2 > 2\alpha \geqq 1$.*

(ii) *Suppose that U satisfies the hypothesis of Theorem* 5.1, *that $R_-^{(n)}(k) = O(k^{-3})$ for $n = 0, 1, 2$, and that $R_-^{(3)}(k) = O(k^{-1})$. Let $u(x,t)$ be the solution of* KdV *given by Theorem* 5.1. *Then $u(x,t)$ evolves in $L^2(\mathbb{R})$ for $t > 0$.*

Remark 1. The purpose of the extra hypothesis is to allow use of the left-side Marchenko equation as well as the right-side one, and thereby to study $u(x,t)$ as $x \to -\infty$. Sachs did not treat this point. Kruzhkov and Faminskii also show evolution in $L^2(\mathbb{R})$, but they consider weighted L^2 norms only on $\mathbb{R}^+$, and their construction is not conducive to analysis of the long time asymptotics of their solution.

Remark 2. For U in $L_N^1(\mathbb{R})$ it is known that additional regularity of U is a sufficient condition for additional decay of $R_-(k)$. For example, from [4] one learns that if $U(x)$ is absolutely continuous, $U'(x)$ is piecewise absolutely continuous, $U \in L_5^1(\mathbb{R})$, $U' \in L_5^1(\mathbb{R})$, and $U'' \in L_4^1(\mathbb{R})$, then the hypothesis of (ii) is satisfied. Another example [20] shows that if $x^m \partial_x^j U(x)$ is in $L^2(\mathbb{R})$ for $0 \leqq m \leqq 4$ and $0 \leqq j \leqq 4$, then the hypothesis of (ii) is also satisfied.

Proof. (i) We consider the solution $u(x,t)$ provided by Theorem 5.1. We know from § 3 that for $t > 0$

$$\begin{aligned}\partial_x F_+(x,t) &= O(|x|^{-N+1/2+\varepsilon}) \quad \text{as } x \to +\infty,\\ \partial_x F_-(x,t) &= O(|x|^{-\lambda/2+1/4}) \quad \text{as } x \to -\infty.\end{aligned}$$

The same decay rates hold for $\partial_x\Omega_+, \partial_x\Omega_-$. By the forward scattering theory [5], [6] all hypotheses of Theorem 3.3 in [8] are satisfied at each $t > 0$. Thus we conclude by Theorem 3.9 of [8] that

$$u(\cdot,t) = -\partial_x B_+(\cdot,0,t) = \partial_x B_-(\cdot,0,t)$$

in $L_{\text{loc}}^1(\mathbb{R})$ for each fixed positive t, where B_+ and B_- are the solutions of the Marchenko equations

$$\begin{aligned}B_+(x,y,t) + \Omega_+(x+y,t) + \int_{z=0}^\infty B_+(x,z,t)\Omega_+(x+y+z,t) &= 0,\\ B_-(x,y,t) + \Omega_-(x+y,t) + \int_{z=-\infty}^0 B_-(x,z,t)\Omega_-(x+y+z,t) &= 0\end{aligned}$$

in $L^1(\mathbb{R}^+)$ and $L^1(\mathbb{R}^-)$, respectively.

In the generic case $u(x, t) = O(|x|^{-N+1+\varepsilon})$ as $x \to \infty$ for $0 < \varepsilon \ll 1/2$. In the exceptional case $u(x, t) = O(x^{-N+2+\varepsilon})$ as $x \to \infty$ for $0 < \varepsilon \ll 1/2$. Since $N \geqq 5$, $u(\cdot, t) \in L^2(\mathbb{R}^+)$. If we know that both $\lambda > 5/2$ and $N \geqq \lambda + 2$, then we can pick α so $\lambda - 3/2 > 2\alpha \geqq 1$ and conclude that

$$\int_0^\infty |s|^{2\alpha} |u(s, t)|^2 \, ds < \infty$$

since $2\alpha + 2(-N + 2\varepsilon) < -1$.

If we take α so $\lambda - 3/2 > 2\alpha \geqq 1$, then by Kappeler's Theorem 3.9 [8] we get

$$\int_{-\infty}^0 |s|^{2\alpha} |u(s, t)|^2 \, ds < \infty \quad \text{for } t > 0.$$

Since $2\alpha \geqq 1$ we get $\int_{-\infty}^{-1} |u(s, t)|^2 \, ds < \infty$ also. But since $u(x, t)$ is in $L^1(+\infty) \cap L^\infty(+\infty)$ we can conclude that $\int_{-1}^0 |u(s, t)|^2 \, ds < \infty$, and further that $u(\cdot, t) \in L^2(\mathbb{R}^-)$.

The proof of (i) is completed by combining results on $\mathbb{R}^+, \mathbb{R}^-$; the proof of (ii) is similar. □

THEOREM 5.3. *For $N \geqq 3$, assume that $U \in L^1_N(\mathbb{R})$ if U is generic, but that $U \in L^1_{N+1}(\mathbb{R})$ if U is nongeneric. Let $u(x, t)$ be the solution of* KdV *provided by Theorem* 5.1. *Recall $B_+(x, 0, t) = \int_x^\infty u(z, t) \, dz$. Then*

(i) *For $0 \leqq n \leqq N - 1$, $x^n B_+(x, 0, t) \to x^n B_+(x, 0, 0)$ in $L^2(+\infty)$ as $t \to 0$;*

(ii) *For each n with $1 \leqq n \leqq N - 1$, if $(1 + x^n) U(x)$ is in $L^2(\mathbb{R})$, then $x^\alpha u(x, t) \to x^\alpha U(x)$ in $L^2(+\infty)$ as $t \to 0$ for all α with $0 \leqq \alpha \leqq n$;*

(iii) *For each n with $1 \leqq n \leqq N - 1$, if both $(1 + x^n) U(x)$ and $(1 + x^n) U'(x)$ are in $L^2(+\infty)$, then also $x^\alpha \partial_x u(x, t) \to x^\alpha U'(x)$ in $L^2(+\infty)$ as $t \to 0$ for all α with $0 \leqq \alpha \leqq n$.*

Proof. Because $R_+(k)$ is at least C^1 and $R_+'(k)$ is $O(k^{-1})$ as $k \to \pm\infty$, the proof of (i) may be taken over from the proof of Kappeler's [9, Thm. 3.1]. We prove (ii) below; the proof of (iii) is similar.

Start with the representation

$$\begin{aligned} u(x, t) - U(x) = {} & \partial_x \Omega_+(x, t) - \partial_x \Omega_+(x, 0) \\ & + \int_0^\infty \{B_+(x, z, t) - B_+(x, z, 0)\} \partial_x \Omega_+(x + z, t) \, dz \\ & + \int_0^\infty B_+(x, z, 0) \{\partial_x \Omega_+(x + z, t) - \partial_x \Omega + (x + z, 0)\} \, dz \\ & + \int_0^\infty \{\partial_x B_+(x, z, t) - \partial_x B_+(x, z, 0)\} \Omega_+(x + z, t) \, dz \\ & + \int_0^\infty \partial_x B_+(x, z, 0) \{\Omega_+(x + z, t) - \Omega_+(x + z, 0)\} \, dz, \end{aligned}$$

which is based on the Marchenko equation. Call the five terms on the right $T_\nu(x, t)$ for $\nu = 1, \cdots, 5$. We must show that $x^\alpha T_\nu(x, t) \to 0$ in $L^2([X, \infty))$ as $t \to 0$ for arbitrary X and $\nu = 1, \cdots, 5$. We do this by assuming three technical points which will be stated when first used but not proved until the end.

Since $x^\alpha U(x) \in L^2(\mathbb{R})$ for $0 \leqq \alpha \leqq n$, Proposition 2.7 tells us that $k R_+^{(\alpha)}(k)$ is in $L^2(\mathbb{R})$ for $0 \leqq \alpha \leqq n$, and that $R_+^{(\alpha)}(k)$ is in $L^2(\mathbb{R})$ for $0 \leqq \alpha \leqq N$. Thus

$$x^\alpha F_+(x) \in L^2(\mathbb{R}) \quad \text{for } 0 \leqq \alpha \leqq N,$$

and

$$x^\alpha \partial_x F_+(x) \in L^2(\mathbb{R}) \quad \text{for } 0 \leqq \alpha \leqq n.$$

By Kappeler's method of proof [9, Thm. 3.1] one may show that if $0 \leqq \beta \leqq n$, then

$$x^\beta \Omega_+(x,t) \to x^\beta \Omega_+(x,0) \quad \text{in } L^2(+\infty) \text{ as } t \to 0 \tag{5.3a}$$

and that

$$\partial_x \Omega_+(x,t) \to \partial_x \Omega_+(x,0) \quad \text{in } L^2(+\infty) \text{ as } t \to 0. \tag{5.3b}$$

The convergence of $x^\alpha T_1(x,t)$ to 0 in $L^2([X,\infty))$ as $t \to 0$ is one part of our first technical result:

Point 1. If $0 \leqq \alpha \leqq n$ and X is fixed, then

$$x^\alpha \partial_x \Omega_+(x,t) \to x^\alpha \partial_x \Omega_+(x,0) \quad \text{in } L^2([X,\infty)) \quad \text{as } t \to 0, \tag{5.4}$$

$$\int_X^\infty |x|^{2\alpha} \left(\int_x^\infty |\partial_s \Omega_+(s,t) - \partial_s \Omega_+(s,0)| \right)^2 dx \to 0 \quad \text{as } t \to 0. \tag{5.5}$$

Looking at $T_2(x,t)$ we see

$$\begin{aligned}\int_X^\infty |x^\alpha T_2(x,t)|^2 \, dx &\leqq \int_X^\infty \left\{ |x|^\alpha \int_0^\infty |B_+(x,z,t) - B_+(x,z,0)|^2 \, dz \right\} \\ &\quad \cdot \left\{ |x|^\alpha \int_0^\infty |\partial_x \Omega_+(x+z,t)|^2 \, dz \right\} dx \\ &\leqq \left\{ \int_X^\infty |x|^{2\alpha} \left(\int_0^\infty |B_+(x,z,t) - B_+(x,z,0)|^2 \, dz \right)^2 dx \right\}^{1/2} \\ &\quad \cdot \left\{ \int_X^\infty |x|^{2\alpha} \left(\int_0^\infty |\partial_x \Omega_+(x+z,t)|^2 \, dz \right)^2 dx \right\}^{1/2}.\end{aligned}$$

The second factor is bounded as $t \to 0$ by (5.4). Part of our next technical point tells us that the first factor vanishes as $t \to 0$:

Point 2. Fix X. Then, as $t \to 0$,

$$B_+(x,\cdot,t) \to B_+(x,\cdot,0) \quad \text{in } L^2(\mathbb{R}^+) \text{ uniformly for } x \geqq X, \tag{5.6}$$

$$\int_X^\infty |x|^{2\alpha} \left(\int_0^\infty |B_+(x,z,t) - B_+(x,z,0)|^2 \, dz \right)^2 dx \to 0 \quad \text{for } 0 \leqq \alpha \leqq n. \tag{5.7}$$

Looking at the third term we see

$$\begin{aligned}\int_X^\infty |x^\alpha T_3(x,t)|^2 \, dx &\leqq \int_X^\infty \left\{ |x|^\alpha \int_0^\infty |B_+(x,z,0)|^2 dz \right\} \\ &\quad \cdot \left\{ |x|^\alpha \int_0^\infty |\partial_x \Omega_+(x+z,t) - \partial_x \Omega_+(x+z,0)|^2 \, dz \right\} dx \\ &\leqq \left\{ \int_X^\infty |x|^{2\alpha} \left(\int_0^\infty |B_+(x,z,0)|^2 \, dz \right)^2 dx \right\}^{1/2} \\ &\quad \cdot \left\{ \int_X^\infty |x|^{2\alpha} \left(\int_0^\infty |\partial_x \Omega_+(x+z,t) - \partial_x \Omega_+(x+z,0)|^2 \, dz \right)^2 dx \right\}^{1/2}.\end{aligned}$$

The first factor is finite by (5.6). The second factor vanishes because of the form of Ω_+ and result (5.5).

For the fourth term, we see

$$\int_X^\infty |x^\alpha T_4(x,t)|^2\,dx \leqq \int_X^\infty \left\{|x|^\alpha \int_0^\infty |\partial_x B_+(x,z,t)-\partial_x B_+(x,z,0)|^2\,dz\right\} \cdot \left\{|x|^\alpha \int_0^\infty |\Omega_+(x+z,t)|^2\,dz\right\}dx$$

$$\leqq \left\{\int_X^\infty |x|^{2\alpha}\left(\int_0^\infty |\partial_x B_+(x,z,t)-\partial_x B_+(x,z,0)|^2\,dz\right)^2 dx\right\}^{1/2} \cdot \left\{\int_X^\infty |x|^{2\alpha}\left(\int_0^\infty |\Omega_+(x+,z,t)|^2\,dz\right)^2 dx\right\}^{1/2}.$$

The second factor is bounded as $t\to 0$ by Point 1. The first factor vanishes by the final technical result:

Point 3. As $t\to 0$

(5.8) $$\partial_x B_+(x,z,t)\to\partial_x B_+(x,z,0)\quad \text{in } L^2(\mathbb{R}^+) \text{ uniformly for } x\geqq X,$$

(5.9) $$\int_X^\infty |x|^{2\alpha}\left(\int_0^\infty |\partial_x B_+(x,z,t)-\partial_x B_+(x,z,0)|^2\,dz\right)^2 dx\to 0\quad \text{for } 0\leqq\alpha\leqq n.$$

Finally,

$$\int_X^\infty |x^\alpha T_5(x,t)|^2\,dx \leqq \int_X^\infty \left\{|x|^\alpha \int_0^\infty |\partial_x B_+(x,z,0)|^2\,dz\right\} \cdot \left\{|x|^\alpha \int_0^\infty |\Omega_+(x+z,t)-\Omega_+(x+z,0)|^2\,dz\right\}dx$$

$$\leqq \left\{\int_X^\infty |x|^{2\alpha}\left(\int_0^\infty |\partial_x B_+(x,z,0)|^2\,dz\right)^2 dx\right\}^{1/2} \cdot \left\{\int_X^\infty |x|^{2\alpha}\left(\int_0^\infty |\Omega_+(x+z,t)-\Omega_+(x+z,0)|^2\,dz\right)^2 dx\right\}^{1/2}.$$

Point 3 says the first factor is finite; the second factor vanishes by (5.3a).

To complete the proof of Theorem 5.3 we must now prove three technical points. Recall that

$$\Omega_+(x,t)=F_+(x,t)+2\sum c_{+j}\exp(8\kappa_j^3 t-2\kappa_j x).$$

Thus for Point 1 it suffices to prove

Point 1′. If $0\leqq\alpha\leqq n$, $1\leqq n\leqq N-1$ and $X\in\mathbb{R}$, then as $t\to 0$

(5.4′) $$x^\alpha\partial_x F_+(x,t)\to x^\alpha\partial_x F_+(x,0)\quad \text{in } L^2([X,\infty)),$$

(5.5′) $$\int_X^\infty |x|^{2\alpha}\left(\int_x^\infty |\partial_s F_+(s,t)-\partial_s F_+(s,0)|^2\,ds\right)^2 dx\to 0.$$

Proof of (5.4′). It suffices to treat $\alpha=0$ and $\alpha=n$. Recall that

(5.10) $$\partial_x F_+(x,t)=\pi^{-1}\int_{-\infty}^\infty 2ikR_+(k)\,e^{8ik^3t+2ikx}\,dk.$$

By Proposition 2.7 we know $2ikR_+(k)$ is in $L^2(\mathbb{R})$. It follows that $\partial_x F_+(x, t) \to \partial_x F_+(x, 0)$ in $L^2(\mathbb{R})$ as $t \to 0$. It is the exponential terms in $\partial_x \Omega_+(x, t)$ that restrict the convergence (5.4) to halflines. This takes care of the case $\alpha = 0$. Next we take $1 \leqq \alpha = n \leqq N-1$. By the case $\alpha = 0$ we know

$$x^\alpha \partial_x F_+(x, t) \to x^\alpha \partial_x F_+(x, 0) \quad \text{in } L^2_{\text{loc}}.$$

Thus it suffices to prove

$$x^\alpha \partial_x F_+(x, t) \to x^\alpha \partial_x F_+(x, 0) \quad \text{in } L^2([2, \infty)).$$

Set $\mathcal{R}(k) = 2ikR_+(k)$ in (5.10). Note that

$$\partial_x F_+(x, t) = \pi^{-1} \int_{-\infty}^{\infty} \mathcal{R}(k) \frac{1}{2i(12k^2t + x)} \frac{\partial[e^{8ik^3t + 2ikx}]}{\partial k}\, dk$$

$$= (-1)\pi^{-1} \int_{-\infty}^{\infty} \frac{\partial}{\partial k}\left[\frac{\mathcal{R}(k)}{2i(12k^2t + x)}\right][e^{8ik^3t + 2ikx}]\, dk.$$

Repeating this procedure one finds

$$\partial_x F_+(x, t) = (-1)^\alpha \pi^{-1} \int_{-\infty}^{\infty} \mathcal{T}^\alpha[\mathcal{R}]\, e^{8ik^3t + 2ikx}\, dk$$

where $\mathcal{T}[g] \equiv \partial_k[g/12i(12k^2t + x)]$. Now $\mathcal{T}^\alpha[\mathcal{R}]$ is a linear combination of terms of the form

$$\frac{\mathcal{R}^{(\lambda)}(k) k^\mu t^\nu}{(12k^2t + x)^{\alpha+\nu}} \quad \text{where } 0 \leqq \lambda \leqq \alpha,\ 0 \leqq \mu \leqq \nu,\ 0 \leqq \lambda + \nu \leqq \alpha.$$

Since $x^\alpha \partial_x F_+(x, t)$ is a linear combination of the terms

$$I_{\lambda,\mu,\nu}(x, t) = t^\nu \pi^{-1} \int_{-\infty}^{\infty} \mathcal{R}^{(\lambda)}(k) \frac{k^\mu x^\alpha}{(12k^2t + x)^{\alpha+\nu}} e^{8ik^3t + 2ikx}\, dk,$$

it will suffice to show that as $t \to 0$

$$I_{\lambda,\mu,\nu}(x, t) \to I_{\lambda,\mu,\nu}(x, 0) \quad \text{in } L^2([2, \infty)).$$

Case $\nu > 0$. Here we need $I_{\lambda,\mu,\nu}(x, t) \to 0$ in $L^2([2, \infty))$. Since $\lambda \leqq \alpha = n \leqq N-1$ we know $\mathcal{R}^{(\lambda)} \in L^2(\mathbb{R})$. For each t let $W_\lambda(x, t)$ denote the inverse Fourier transform of $\mathcal{R}^{(\lambda)}(k) \exp(8ik^3t)$. Now we can see $I_{\lambda,\mu,\nu}$ as the result of a pseudo-differential operator acting on W_λ. The symbol

$$p_{\mu,\nu}(x, k) = \frac{k^\mu x^\alpha}{(12k^2t + x)^{\alpha+\nu}}$$

has the property that there is a C such that

$$|\partial_k^i \partial_x^j p_{\mu,\nu}(x, k)| \leqq C \quad \text{for } x \geqq 1, \qquad 0 \leqq t \leqq 1.$$

Choose a nonnegative C^∞ cutoff function $\zeta(x)$ such that $\zeta(x) = 0$ for $x \leqq 0$, $\zeta(x) = 1$ for $x \geqq 2$, and $|\partial_x^m \zeta(x)| \leqq M_0$ for all $m \in \mathbb{N}$. Now for $x \geqq 2$

$$I_{\lambda,\mu,\nu}(x, t) = t^\nu \mathcal{F}^{-1}[\zeta(x) p_{\mu,\nu}(x, t) \mathcal{F}[W_\lambda(x, t)]].$$

By a result of Calderon and Vaillancourt (as present in [18, Thm. 3.1, p. 347]) we see there is a constant M_1 independent of t for $0 \leqq t \leqq 1$ such that

$$\|t^\nu \mathcal{F}^{-1}[\zeta(x)p_{\mu,\nu}(x, t)\mathcal{F}[W_\lambda(x, t)]]\|_{L^2(-\infty<x<\infty)}$$
$$\leqq t^\nu M_1 \|\mathcal{F}[W_\lambda(x, t)]\|_{L^2(-\infty<x<\infty)} = t^\nu M_1 \|\mathcal{R}^{(\lambda)}\|_{L^2(-\infty<k<\infty)}.$$

Thus, as required, $I_{\lambda,\mu,\nu}(x, t) \to 0$ in $L^2([2, \infty))$ as $t \to 0$.

Case $\nu = 0$. Since $0 \leqq \mu \leqq \nu$, we must show

$$I_{\lambda,0,0,}(x, t) \to I_{\lambda,0,0,}(x, 0) \quad \text{in } L^2([2, \infty)) \text{ as } t \to 0.$$

Now

$$I_{\lambda,0,0,}(x, t) - I_{\lambda,0,0,}(x, 0) = \pi^{-1} \int_{-\infty}^{\infty} \mathcal{R}^{(\lambda)}(k) \frac{x^\alpha}{(12k^2t + x)^\alpha} [e^{8ik^3t} - 1]\, e^{2ikx}\, dk$$
$$+ \pi^{-1} \int_{-\infty}^{\infty} \mathcal{R}^{(\lambda)}(k) \left\{ \frac{x^\alpha}{(12k^2t + x)^\alpha} - 1 \right\} e^{2ikx}\, dk$$

and $\mathcal{R}^{(\lambda)}(k)[e^{8ik^3t} - 1]$ is in $L^2(\mathbb{R})$. Apply the same result of Calderon and Vaillancourt to obtain

$$\left\| \pi^{-1} \int_{-\infty}^{\infty} \mathcal{R}^{(\lambda)}(k)[e^{8ik^3t} - 1]\zeta(x) \left(\frac{x}{12k^2t + x} \right)^\alpha e^{2ikx}\, dk \right\|_{L^2(-\infty<x<\infty)}$$
$$\leqq M_1 \|\mathcal{R}^{(\lambda)}(k)[e^{8ik^3t} - 1]\|_{L^2(-\infty<k<\infty)},$$

the right side of which clearly goes to 0 as $t \to 0$. The second term in the difference $I_{\lambda,0,0}(x, t) - I_{\lambda,0,0}(x, 0)$ is treated similarly to complete the proof of (5.4′).

In proving (5.5′) we may assume $X \leqq 1$. Let $E(x, t)$ denote $\int_x^\infty |\partial_s F_+(s, t) - \partial_s F_+(s, 0)|^2\, ds$. We must show

$$\int_X^\infty |x|^{2\alpha} E(x, t)^2\, dx \to 0 \quad \text{as } t \to 0.$$

Divide the integral at $x = 1$. Clearly

$$\int_X^1 \cdots dx \leqq \max\{1, |x|^{2\alpha}\} \int_X^\infty E(x, t)^2\, dx$$
$$\leqq \max\{1, |x|^{2\alpha}\} E(X) \int_X^\infty E(x, t)\, dx$$
$$\leqq KE(X) \int_X^\infty \int_x^\infty |\partial_s F_+(s, t) - \partial_s F_+(s, 0)|^2\, ds\, dx$$
$$\leqq KE(X) \int_{s=X}^\infty (s - X)|\partial_s F_+(s, t) - \partial_s F_+(s, 0)|^2\, ds,$$

which goes to zero by (5.4′). Next

$$\int_1^\infty \cdots dx = \int_1^\infty \{x^{2\alpha} E(x, t)\} E(x, t)\, dx$$
$$\leqq \int_1^\infty \left\{ \int_x^\infty s^{2\alpha} |\partial_s F_+(s, t) - \partial_s F_+(s, 0)|^2\, ds \right\} E(x, t)\, dx$$
$$\leqq \left\{ \int_1^\infty s^{2\alpha} |\partial_s F_+(s, t) - \partial_s F_+(s, 0)|^2\, ds \right\} \int_1^\infty E(x, t)\, dx.$$

The first factor goes to zero by (5.4′). The second also does since

$$\int_1^\infty E(x,t)\,dx=\int_1^\infty (s-1)|\partial_s F_+(s,t)-\partial_s F_+(s,0)|^2\,ds.$$

This completes the proof of Point 1. □

Point 2.

(i) $B_+(x,\cdot,t)\to B_+(x,\cdot,0)$ in $L^2(\mathbb{R}^+)$ as $t\to 0$ uniformly in $x\geqq X$;

(ii) $\int_1^\infty |x|^{2\alpha}\|B_+(x,\cdot,t)\to B_+(x,\cdot,0)\|^4_{L^2(\mathbb{R}^+)}\,dx\to 0$ as $t\to 0$ for $0\leqq \alpha\leqq n$.

Proof. We have

$$\begin{aligned}&\|B_+(x,\cdot,t)-B_+(x,\cdot,0)\|_{L^2(\mathbb{R}+)}\\ &\quad\leqq C_0(x)\|\Omega_+(x+[\cdot],t)-\Omega_+(x+[\cdot],0)\|_{L^2(\mathbb{R}+)}\\ &\qquad+\|(\mathbf{I}+\mathbf{\Omega}^t_x)^{-1}-(\mathbf{I}+\mathbf{\Omega}^0_x)^{-1}\|_{op}\|\Omega_+(x+[\cdot],0)\|_{L^2(\mathbb{R}^+)}\end{aligned}\tag{5.11}$$

where $C_0(x)=\sup\{\|(\mathbf{I}+\mathbf{\Omega}^t_w)^{-1}\|_{op}: x\leqq w, 0\leqq t\leqq 1\}$. $C_0(x)$ is finite and nonincreasing. We show that each term on the right of (5.11) vanishes as $t\to 0$. First

$$\sup\{\|\Omega_+(x+[\cdot],t)-\Omega_+(x+[\cdot],0)\|^2_{L^2(\mathbb{R}^+)}: x\geqq X\}\leqq\int_X^\infty |\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds,$$

which, as we have already seen, vanishes as $t\to 0$. Second,

$$\begin{aligned}\|(\mathbf{I}+\mathbf{\Omega}^t_x)^{-1}-(\mathbf{I}+\mathbf{\Omega}^0_x)^{-1}\|_{op}&\leqq C_0(x)^2\|(\mathbf{I}+\mathbf{\Omega}^t_x)^{-1}\|_{op}\\ &\leqq C_0(x)^2\left(\int_{y=x}^\infty (y-x)|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\right)^{1/2}.\end{aligned}$$

Thus

$$\begin{aligned}&\sup\{\|(\mathbf{I}+\mathbf{\Omega}^t_x)^{-1}-(\mathbf{I}+\mathbf{\Omega}^0_x)^{-1}\|_{op}: x\geqq X\}\\ &\qquad\leqq C_0(X)\left(\int_{y=x}^\infty (y-X)|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\right)^{1/2}\end{aligned}$$

which vanishes as $t\to 0$ by Point 1. This completes (i). Because of (i) it suffices to prove (ii) for $X=0$. Now by (5.11)

$$\begin{aligned}&\|B_+(x,\cdot,t)-B_+(x,\cdot,0)\|^4\\ &\quad\leqq C_1(x)\left(\int_x^\infty|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\right)^2\\ &\qquad+C_0(x)\left(\int_x^\infty (y-x)|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\right)^2\left(\int_x^\infty|\Omega_+(s,0)|^2\,ds\right)^2\end{aligned}$$

where $C_1(x)$ and $C_2(x)$ are nonincreasing. Now first

$$\begin{aligned}\int_0^\infty |x|^{2\alpha}\left(\int_x^\infty|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\right)^2dx&\leqq\int_0^\infty\left(\int_x^\infty s^\alpha|\Omega_+(s,t)-\Omega_+(s,0)|\,ds\right)^2dx\\ &=\int_0^\infty\int_x^\infty s^\alpha|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\,dx\cdot\int_x^\infty s^\alpha|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\end{aligned}$$

$$= \int_0^\infty s^{\alpha+1}|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds \cdot \int_0^\infty s^\alpha|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds,$$

which goes to 0 as $t \to 0$ by (5.3a) since $\alpha \leqq N-1$. Next, we see

$$\int_0^\infty |x|^{2\alpha}\left(\int_x^\infty (y-x)|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\right)^2\left(\int_x^\infty |\Omega_+(s,0)|^2\,ds\right)^2 dx$$

$$\leqq \int_0^\infty \left(\int_x^\infty (y-x)|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\right)^2\left(\int_x^\infty s^\alpha|\Omega_+(s,0)|^2\,ds\right)^2 dx$$

$$\leqq K^2\int_0^\infty y|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy \int_0^\infty\int_x^\infty y|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy\,dx$$

$$= K^2\int_{y=0}^\infty y|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy \int_{y=0}^\infty y^2|\Omega_+(y,t)-\Omega_+(y,0)|^2\,dy.$$

Both of the last two factors go to 0 with t by (5.3a). This completes part (ii).

Point 3.

(i) $\partial_x B_+(x,\cdot,t) \to \partial_x B_+(x,\cdot,0)$ in $L^2(\mathbb{R}^+)$ as $t \to 0$ uniformly in $x \geqq X$;

(ii) $\int_X^\infty |x|^{2\alpha}\|\partial_x B_+(x,\cdot,t)-\partial_x B_+(x,\cdot,0)\|^4_{L^2(\mathbb{R}^+)}\,dx \to 0$ as $t \to 0$ for $0 \leqq \alpha \leqq n$.

Proof. One may verify that

$$(\mathbf{I}+\mathbf{\Omega}_x^t)[\partial_x B_+(x,\cdot,t)-\partial_x B_+(x,\cdot,0)] = -\sum_{\nu=1}^4 Q_\nu(x,\cdot,t)$$

where

$$Q_1(x,\cdot,t) = [(\mathbf{I}+\mathbf{\Omega}_x^t)-(\mathbf{I}+\mathbf{\Omega}_x^0)]\partial_x B_+(x,\cdot,0),$$

$$Q_2(x,y,t) = \partial_x\Omega_+(x+y,t)-\partial_x\Omega_+(x+y,0),$$

$$Q_3(x,y,t) = \int_0^\infty \{B_+(x,z,t)-B_+(x,z,0)\}\partial_x\Omega_+(x+y+z,t)\,dz,$$

$$Q_4(x,y,t) = \int_0^\infty B_+(x,z,0)\{\partial_x\Omega_+(x+y+z,t)-\partial_x\Omega_+(x+y+z,0)\}\,dz.$$

Since $\|(\mathbf{I}+\mathbf{\Omega}_x^t)^{-1}\|_{op} \leqq C_0(X)$ for all $x \geqq X$, $0 \leqq t \leqq 1$, it suffices to prove for each ν that as $t \to 0$

(i) $Q_\nu(x,\cdot,t) \to 0$ in $L^2(\mathbb{R}^+)$ uniformly for $x \geqq X$, and

(ii) $\int_X^\infty |x|^{2\alpha}\|Q_\nu(x,\cdot,t)\|^4_{L^2(\mathbb{R}^+)}\,dx \to 0$.

Case $\nu = 1$.

$$\|Q_1(x,\cdot,t)\|^2_{L^2(\mathbb{R}^+)} = \|\mathbf{\Omega}_x^t-\mathbf{\Omega}_x^0\|^2_{op}\|\partial_x B_+(x,\cdot,0)\|^2_{L^2(\mathbb{R}^+)}.$$

So

$$\sup\{\|Q_1(x,\cdot,t)\|^2_{L^2(\mathbb{R}^+)}: x \geqq X\} \leqq K\int_{s=x}^\infty (s-X)|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds,$$

which goes to 0 with t by (5.3). Thus (i) holds.

$$\begin{aligned}
\int_0^\infty |x|^{2\alpha}\|Q_1(x,\cdot,t)\|^4\,dx &\leqq K\int_0^\infty x^{2\alpha}\left(\int_{s=x}^\infty s|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\right)^2 dx\\
&\leqq K\int_0^\infty\left\{x^2\int_{s=x}^\infty s|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\right\}\\
&\qquad\cdot\left\{x^{2\alpha-2}\int_x^\infty s|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\right\}dx\\
&\leqq K\int_{s=0}^\infty s^3|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\\
&\qquad\cdot\int_0^\infty\int_x^\infty s^{2\alpha-1}|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\,dx\\
&\leqq K\int_0^\infty s^3|\Omega_+(s,t)-\Omega_+(s,0)|^2\,ds\int_0^\infty s^{2\alpha}|\Omega_+(s,t)-(s,0)|^2\,ds,
\end{aligned}$$

which goes to zero with t by (5.3) since $N\geqq 3$ says $3/2\leqq n\leqq N-1$. This finishes (ii).

Case $\nu=2$. This follows directly from Point 1.

Case $\nu=3$.

$$\begin{aligned}
\|Q_3(x,\cdot,t)\|^2_{L^2(\mathbb{R}^+)} &= \int_{y=0}^\infty\left|\int_{z=0}^\infty\{B_+(x,z,t)-B_+(x,z,0)\}\partial_x\Omega_+(x+y+z,t)\,dz\right|^2 dy\\
&\leqq\int_{z=0}^\infty\|B_+(x,\cdot,t)-B_+(x,\cdot,0)\|^2_{L^2(\mathbb{R}^+)}\int_{s-x+y}^\infty|\partial_x\Omega_+(s,t)|^2\,ds\,dy\\
&\leqq\|B_+(x,\cdot,t)-B_+(x,\cdot,0)\|^2\int_{s=x}^\infty(s-x)|\partial_x\Omega_+(s,t)|^2\,ds.
\end{aligned}$$

Convergence (i) follows by Point 1 and Point 2. For (ii) note

$$\begin{aligned}
&\int_0^\infty |x|^{2\alpha}\|Q_3(x,\cdot,t)\|^4\,dx\\
&\qquad\leqq\int_0^\infty |x|^{2\alpha}\|B_+(x,\cdot,t)-B_+(x,\cdot,0)\|^4\left(\int_{s=x}^\infty(s-x)|\partial_s\Omega_+(s,t)|^2\,ds\right)^2 dx.
\end{aligned}$$

Since $B_+(x,\cdot,t)\to B_+(x,\cdot,0)$ in $L^2(\mathbb{R}^+)$ uniformly in $x\geqq 0$ we check the boundedness of the rest:

$$\begin{aligned}
&\int_0^\infty |x|^{2\alpha}\left(\int_{s=x}^\infty(s-x)|\partial_s\Omega_+(s,t)|^2\,ds\right)^2 dx\\
&\qquad\leqq\int_0^\infty\left\{x^2\int_x^\infty s|\partial_s\Omega_+(s,t)|^2\,ds\right\}^2\left\{x^{2\alpha-2}\int_{s=x}^\infty s|\partial_x\Omega_+(s,t)|^2\,ds\right\}^2 dx\\
&\qquad\leqq\int_0^\infty s^3|\partial_s\Omega_+(s,t)|^2\,ds\int_0^\infty s^{2\alpha}|\partial_x\Omega_+(s,t)|^2\,ds.
\end{aligned}$$

Both factors are bounded for $0\leqq t\leqq 1$ by Point 1.

Case $\nu = 4$.

$$\|Q_4(x,\cdot,t)\|^2 = \int_{y=0}^{\infty}\left|\int_0^\infty B_+(x,z,0)\{\partial_x\Omega_+(x+y+z,t)-\partial_x\Omega_+(x+y+z,0)\}\,dz\right|^2 dy$$

$$\leqq \int_0^\infty\left(\int_0^\infty |B_+(x,z,0)|^2\,dz\right)$$

$$\cdot\left(\int_0^\infty |\partial_x\Omega_+(x+y+z,t)-\partial_x\Omega_+(x+y+z,0)|^2\,dz\right)dy$$

$$\leqq \int_0^\infty |B_+(x,z,0)|^2\,dz \int_{y=0}^\infty \int_{s=x+y}^\infty |\partial_s\Omega_+(s,t)-\partial_s\Omega_+(s,0)|^2\,ds\,dy$$

$$\leqq \int_0^\infty |B_+(x,z,0)|^2\,dz \int_{s=x}^\infty (s-x)|\partial_s\Omega_+(s,t)-\partial_s\Omega_+(s,0)|^2\,ds.$$

Now

$$\sup\{\|Q_4(x,\cdot,t)\|^2 : x \geqq X\} \leqq \sup\{\|B_+(x,\cdot,0)\|^2 : x \geqq X\}$$

$$\cdot \int_{x=X}^\infty (s-X)|\partial_s\Omega_+(s,t)-\partial_s\Omega_+(s,0)|^2\,ds$$

which goes to 0 by Point 1. Further

$$\int_0^\infty |x|^{2\alpha}\|Q_4(x,\cdot,t)\|^4\,dx$$

$$\leqq \int_0^\infty |x|^{2\alpha}\left(\int_0^\infty |B_+(x,z,0)|^2\,dz\right)\left(\int_{s=x}^\infty (s-x)|\partial_s\Omega_+(s,t)-\partial_s\Omega_+(s,0)|^2\,ds\right)dx$$

$$\leqq \int_{s=x}^\infty s|\partial_s\Omega_+(s,t)-\partial_s\Omega_+(s,0)|^2\,ds \int_0^\infty x^{2\alpha}\int_0^\infty |B_+(x,z,0)|^2\,dz\,dx.$$

The second factor is finite and the first goes to 0 by Point 1.

This finishes the proof of the last of the three technical points needed to complete the proof of Theorem 5.3. □

Acknowledgments. The authors are grateful to the University of California at Berkeley for its hospitality during the time this work was begun. We also gratefully acknowledge several very helpful discussions with S. Agmon.

REFERENCES

[1] Z. S. AGRANOVICH AND V. A. MARCHENKO, *The Inverse Problem of Scattering Theory*, Gordon and Breach, New York, 1963 (English trans.).

[2] J. L. BONA AND R. SCOTT, *Solutions of the Korteweg-deVries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87-99.

[3] A. COHEN, *Existence and regularity for solutions of the Korteweg-de Vries equation*, Arch. Rat. Mech. Anal., 71 (1979), pp. 143-175.

[4] ———, *Decay and regularity in the inverse scattering problem*, J. Math. Anal. Appl., 87 (1982), pp. 395-426.

[5] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121-252.

[6] L. FADDEEV, *Properties of the S-matrix of the one-dimensional Schrodinger equation*, Amer. Math. Soc. Transl., 65 (1967), pp. 129-166.

[7] C. GARDNER, J. GREENE, M. KRUSKAL AND R. MIURA, *Method for solving the Korteweg-deVries equation*, Phys. Rev. Lett., 19 (1967), pp. 1095-1097.

[8] TH. KAPPELER, *The inverse scattering problem for scattering data with poor regularity or slow decay*, J. Integral Equations, to appear.

[9] ———, *Solutions to the Korteweg-deVries equation with irregular initial profile*, Comm. Partial Differential Equations, 11 (1986), pp. 927-945.

[10] T. KATO, *On the Cauchy problem for the (generalized) Korteweg-deVries equation*, Stud. Appl. Math., Adv. in Math. Suppl. Stud., 8 (1983), pp. 93-128.

[11] S. N. KRUZHKOV AND A. V. FAMINSKII, *Generalized solutions of the Cauchy problem for the Korteweg-deVries equation*, Math. USSR-Sb., 48 (1984), pp. 391-421.

[12] P. D. LAX, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467-490.

[13] A. C. MURRAY, *Solutions of the Korteweg-de Vries equation from irregular data*, Duke Math. J., 45 (1978), pp. 141-181.

[14] R. SACHS, *Classical solutions of the Korteweg-deVries equation for nonsmooth initial data via inverse scattering*, Comm. Partial Differential Equations, 10 (1985), pp. 29-98.

[15] J. C. SAUT AND R. TEMAM, *Remarks on the Korteweg-deVries equation*, Israel J. Math., 24 (1976), pp. 78-87.

[16] A. B. SHABAT, *On the Korteweg-deVries equation*, Soviet Math. Dokl., 14 (1973), pp. 1266-1269.

[17] S. TANAKA, *Korteweg-deVries equation: construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49-59.

[18] M. E. TAYLOR, *Pseudodifferential Operators*, Princeton Univ. Press, Princeton, NJ, 1981.

[19] V. M. YAKUPOV, *The Cauchy problem for the Korteweg-deVries equation*, Differential Equations, 11 (1975), pp. 419-423.

[20] TH. KAPPELER AND E. TRUBOWITZ, *Properties of the scattering map*, Comm. Math. Helv., 61 (1986), pp. 442-480.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

010

ON A COUPLED REACTION DIFFUSION SYSTEM WITH TIME DELAYS*

C. V. PAO†

Abstract. In a two-compartment model for cellular control by repression there arises a coupled system of reaction diffusion equations which governs the densities of the chemical species in the compartments. The reaction mechanism in this model is affected by time delays and the coupling of the equations is through the interface of the two compartments. The purpose of this paper is to give a qualitative analysis of this unusual coupled reaction diffusion system with time delays. This analysis includes the existence and uniqueness of a global time-dependent solution and the corresponding steady state solution, iterative methods for the construction of the solution, and local and global stability of the steady-state solution. Also included is the uniform boundedness of the solution.

Key words. reaction-diffusion equations, upper-lower solutions, time-delayed equations, existence-uniqueness, stability

AMS(MOS) subject classifications. 35K60, 35R10, 35K20

1. Introduction. In the compartment analysis of certain biochemical reactions there arise some mathematical models which are governed by a coupled system of differential equations with time delays. A well-known model for biochemical control of genes was first proposed by Goodwin [3], [4] and later analyzed by many authors (e.g. [1], [5], [7]). Goodwin's model was recently extended to a two-compartment and a three-compartment model where the spatial effect of diffusion is taken into consideration (cf. [6]). This extension leads to a coupled system of parabolic partial differential equations with time-delays. A mathematical analysis for the three-compartment model has been given in [8] where the time-delays only appear in the ordinary differential equations. However, in the two-compartment model the time-delays also occur in the parabolic equation which is coupled with an ordinary differential equation through the interface of the two compartments. This yields an unusual coupling system of reaction diffusion equations. To include possible spherical or cylindrical interface between the two compartments, we consider a multi-dimensional diffusion medium Ω in $\mathbb{R}^m (m=1,2,\cdots,)$ for the second compartment. In this situation the system of reaction diffusion equations is given by

$$\text{(1.1a)}\quad \begin{aligned} L_1[u_1] &\equiv (u_1)_t + (a_1+b_1)u_1 = a_1u_2 + f(v_1(x, t-r_1)) \\ \mathcal{L}_1[v_1] &\equiv (v_1)_t + (a_2+b_2)v_1 = a_2v_2 \end{aligned} \qquad (x \in \Gamma_1,\ t>0),$$

$$\text{(1.1b)}\quad \begin{aligned} L_2[u_2] &\equiv (u_2)_t - D_1\nabla^2 u_2 + b_1u_2 = 0 \\ \mathcal{L}_2[v_2] &\equiv (v_2)_t - D_2\nabla^2 v_2 + b_2v_2 = g(u_2(x, t-r_2)) \end{aligned} \qquad (x \in \Omega,\ t>0)$$

where ∇^2 is the Laplace operator and Γ_1 is the interface between the two compartments (see [6] for a derivation). In the above equations, (u_1, v_1) and (u_2, v_2) represent the densities of the chemical species in the first and second compartment, respectively, a_i and b_i $(i=1,2)$ are the various reaction rates, D_i is the diffusion coefficient, r_i is the time-delay, and f and g are, in general, nonlinear reaction functions with time-delays. On the interface Γ_1, the concentrations (u_1, v_1) and (u_2, v_2) are related through the

* Received by the editors October 29, 1984; accepted for publication May 22, 1986.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695.

boundary condition

$$B(u_2) \equiv \frac{\partial u_2}{\partial \nu} + \beta_1 u_2 = \beta_1 u_1 \qquad\qquad (x \in \Gamma_1, t > 0), \tag{1.2a}$$
$$\mathscr{B}(v_2) \equiv \frac{\partial v_2}{\partial \nu} + \beta_2 v_2 = \beta_2 v_1$$

where β_1 and β_2 are positive constants and $\partial/\partial\nu$ is the outward normal derivative on $\partial\Omega$. On the nonintersecting boundary surface Γ_2 the concentration (u_2, v_2) is required to satisfy the no-flex boundary condition

$$\frac{\partial u_2}{\partial \nu} = 0, \quad \frac{\partial v_2}{\partial \nu} = 0 \quad (x \in \Gamma_2, t > 0). \tag{1.2b}$$

In view of the time-delay the initial condition is given in the form

$$\begin{aligned} &u_1(x, 0) = \xi_1(x), \qquad v_1(x, t) = \eta_1(x, t) \qquad (x \in \Gamma_1, -\gamma_1 \leqq t \leqq 0),\\ &u_2(x, t) = \xi_2(x, t), \qquad v_2(x, 0) = \eta_2(x) \qquad (x \in \Omega, -\gamma_2 \leqq t \leqq 0). \end{aligned} \tag{1.3}$$

Throughout the paper we assume that Ω is a smooth bounded domain in $\mathbb{R}^m$, the initial functions ξ_i, η_i are continuous nonnegative in their respective domain, and f and g are C^1-functions on $\mathbb{R}^+ \equiv [0, \infty)$. Of special interest are the reaction functions

$$f_0(v_1) = \sigma_0(1 + k_0 v_1^\rho)^{-1}, \qquad g_0(u_2) = c_0 u_2, \tag{1.4}$$

where σ_0, k_0, ρ and c_0 are positive constants (cf. [6]). Notice that when Ω is the one-dimensional interval $(0, l)$ and f and g are given by (1.4), problem (1.1)-(1.3) is reduced to the model given in [6].

The purpose of this paper is to study the existence and stability problem of the system (1.1)-(1.3) and its corresponding steady-state system. Our discussion includes the existence and uniqueness of global time-dependent solution and steady-state solution, method of construction of these solutions and the stability property of steady-state solutions. Also included is the uniform boundedness of the time-dependent solution. In § 2 we use the method of upper-lower solutions to establish an existence-comparison theorem for the problem (1.1)-(1.3) and to construct some explicit upper and lower bounds of the solution. Through suitable construction of upper-lower solutions in § 3 we give some sufficient conditions which ensure the local stability and global stability of a steady-state solution. The existence and uniqueness of a steady-state solution is discussed in § 4 where similar upper and lower bounds of the solution are also given.

2. The existence–comparison theorem. In this section we investigate the existence and uniqueness of a global solution to the time-dependent system (1.1)-(1.3). The basic tool for the existence problem is the monotone method and the associated upper-lower solutions for coupled reaction diffusion systems (cf. [9]). The construction of the monotone sequences and the definition of upper-lower solutions depend on the quasi-monotone property of the reaction functions. Motivated by the reaction functions given by (1.4), we make the following basic assumptions.

$$\text{(H)} \qquad f(\eta) \geqq 0, \quad g(\eta) \geqq 0, \quad f'(\eta) \leqq 0, \quad g'(\eta) \geqq 0 \quad \text{for } \eta \geqq 0.$$

The main hypothesis in (H) is that $f(v_1)$ is monotone nonincreasing and $g(u_2)$ is monotone nondecreasing on $[0, \infty)$. With this monotone property of f, g we have the following definition of upper-lower solutions. For notational convenience, we write (u_i, v_i) to represent the function $(u_1, v_1; u_2, v_2)$.

DEFINITION 2.1. A pair of smooth functions $(\tilde{u}_i, \tilde{v}_i)$, $(\underset{\sim}{u}_i, \underset{\sim}{v}_i)$ are called upper and lower solutions of (1.1)-(1.3), respectively, if they satisfy

(i) The differential inequalities:

$$\begin{aligned} &L_1[\tilde{u}_1]-[a_1\tilde{u}_2+f(\underset{\sim}{v}_1(x,t-r_1))]\geqq 0\geqq L_1[\underset{\sim}{u}_1]-[a_1\underset{\sim}{u}_2+f(\tilde{v}_1(x,t-r_1))],\\ &\mathscr{L}_1[\tilde{v}_1]-a_2\tilde{v}_2\geqq 0\geqq \mathscr{L}_1[\underset{\sim}{v}_1]-a_2\underset{\sim}{v}_2,\\ &L_2[\tilde{u}_2]\geqq 0\geqq L_2[\underset{\sim}{u}_2],\\ &\mathscr{L}_2[\tilde{v}_2]-g(\tilde{u}_2(x,t-r_2))\geqq 0\geqq \mathscr{L}_2[\underset{\sim}{v}_2]-g(\underset{\sim}{u}_2(x,t-r_2)); \end{aligned}\tag{2.1}$$

(ii) The boundary inequalities:

$$\begin{aligned} &\left.\begin{aligned} &B[\tilde{u}_2]-\beta_1\tilde{u}_1\geqq 0\geqq B[\underset{\sim}{u}_2]-\beta_1\underset{\sim}{u}_1\\ &\mathscr{B}[\tilde{v}_2]-\beta_2\tilde{v}_1\geqq 0\geqq \mathscr{B}[\underset{\sim}{v}_2]-\beta_2\underset{\sim}{v}_1 \end{aligned}\right. &&(t>0,\ x\in\Gamma_1),\\ &\left.\begin{aligned} &\frac{\partial\tilde{u}_2}{\partial\nu}\geqq 0\geqq\frac{\partial\underset{\sim}{u}_2}{\partial\nu}\\ &\frac{\partial\tilde{v}_2}{\partial\nu}\geqq 0\geqq\frac{\partial\underset{\sim}{v}_2}{\partial\nu} \end{aligned}\right. &&(t>0,\ x\in\Gamma_2); \end{aligned}\tag{2.2}$$

(iii) The initial inequalities:

$$\begin{aligned} &\left.\begin{aligned} &\tilde{u}_1(x,0)-\xi_1(x)\geqq 0\geqq\underset{\sim}{u}_1(x,0)-\xi_1(x)\\ &\tilde{v}_1(x,t)-\eta_1(x,t)\geqq 0\geqq\underset{\sim}{v}_1(x,t)-\eta_1(x,t) \end{aligned}\right. &&(x\in\Gamma_1,\ t\in[-r_1,0]),\\ &\left.\begin{aligned} &\tilde{u}_2(x,t)-\xi_2(x,t)\geqq 0\geqq\underset{\sim}{u}_2(x,t)-\xi_2(x,t)\\ &\tilde{v}_2(x,0)-\eta_2(x)\geqq 0\geqq\underset{\sim}{v}_2(x,0)-\eta_2(x) \end{aligned}\right. &&(x\in\Omega,\ t\in[-r_2,0]). \end{aligned}\tag{2.3}$$

Here by a pair of smooth functions we mean $(\tilde{u}_i, \tilde{v}_i)$ and $(\underset{\sim}{u}_i, \underset{\sim}{v}_i)$ are continuous in their respective domains and are once continuously differentiable in t and twice continuously differentiable in x for $t>0, x\in\Omega$; in addition, $(\tilde{u}_2, \tilde{v}_2)$ and $(\underset{\sim}{u}_2, \underset{\sim}{v}_2)$ are differentiable in the outward normal direction on $\partial\Omega$. Notice that in the above definition, upper and lower solutions are interrelated through the first inequality in (2.1).

Let $D_T=\Omega\times(0,T]$, $S_1=\Gamma_1\times(0,T]$, $S_2=\Gamma_2\times(0,T]$ and let $\bar{D}_T$ be the closure of D_T. Assume there exist upper and lower solutions such that $(\tilde{u}_i, \tilde{v}_i)\geqq(\underset{\sim}{u}_i, \underset{\sim}{v}_i)\geqq(0,0)$ on $\bar{D}_T$ (i.e., $\tilde{u}_i\geqq\underset{\sim}{u}_i\geqq 0$, $\tilde{v}_i\geqq\underset{\sim}{v}_i\geqq 0$ on $\bar{D}_T$, $i=1,2$). By using $(\bar{u}_i^{(0)}, \bar{v}_i^{(0)})=(\tilde{u}_i, \tilde{v}_i)$ and $(\underline{u}_i^{(0)}, \underline{v}_i^{(0)})=(\underset{\sim}{u}_i, \underset{\sim}{v}_i)$ as two initial iterations we construct two sequences $\{\bar{u}_i^{(k)}, \bar{v}_i^{(k)}\}$, $\{\underline{u}_i^{(k)}, \underline{v}_i^{(k)}\}$ from the following iterative process:

$$\begin{aligned} &\left.\begin{aligned} &L_1[\bar{u}_1^{(k)}]=a_1\bar{u}_2^{(k-1)}+f(\underline{v}_1^{(k-1)}(x,t-r_1))\\ &\mathscr{L}_1[\bar{v}_1^{(k)}]=a_2\bar{v}_2^{(k-1)} \end{aligned}\right. &&((x,t)\in S_1),\\ &\left.\begin{aligned} &L_2[\bar{u}_2^{(k)}]=0\\ &\mathscr{L}_2[\bar{v}_2^{(k)}]=g(\bar{u}_2^{(k-1)}(x,t-r_2)) \end{aligned}\right. &&((x,t)\in D_T), \end{aligned}\tag{2.4}$$

$$
(2.5)\qquad
\begin{aligned}
&L_1[\underline{u}_1^{(k)}]=a_1\underline{u}_2^{(k-1)}+f(\bar{v}_1^{(k-1)}(x,t-r_1))\\
&\mathscr{L}_1[\underline{v}_1^{(k)}]=a_2\underline{v}_2^{(k-1)}
\end{aligned}
\qquad ((x,t)\in S_1),
$$
$$
\begin{aligned}
&L_2[\underline{u}_2^{(k)}]=0\\
&\mathscr{L}_2[\underline{v}_2^{(k)}]=g(\underline{u}_2^{(k-1)}(x,t-r_2))
\end{aligned}
\qquad ((x,t)\in D_T).
$$

The boundary and initial conditions for both systems (2.4) and (2.5) are given in the form

$$
(2.6)\qquad
\begin{aligned}
&B[u_2^{(k)}]=\beta_1 u_1^{(k-1)}, \qquad \mathscr{B}[v_2^{(k)}]=\beta_2 v_1^{(k-1)} &&((x,t)\in S_1),\\
&\frac{\partial u_2^{(k)}}{\partial \nu}=\frac{\partial v_2^{(k)}}{\partial \nu}=0 &&((x,t)\in S_2),
\end{aligned}
$$

$$
(2.7)\qquad
\begin{aligned}
&u_1^{(k)}(x,0)=\xi_1(x), \qquad v_1^{(k)}(x,t)=\eta_1(x,t) &&(x\in\Gamma_1,\ t\in[-r_1,0]),\\
&u_2^{(k)}(x,t)=\xi_2(x,t), \qquad v_2^{(k)}(x,0)=\eta_2(x) &&(x\in\Omega,\ t\in[-r_2,0)).
\end{aligned}
$$

Notice that the two sequences $\{\bar{u}_i^{(k)},\bar{v}_i^{(k)}\}$, $\{\underline{u}_i^{(k)},\underline{v}_i^{(k)}\}$ are interrelated through the first equation in (2.4) and (2.5). Our aim is to show that these two sequences, called maximal and minimal sequence respectively, are monotone and both converge to a unique solution of the system (1.1)-(1.3). We first establish the monotone property of the sequences.

LEMMA 2.1. *Let* $(\tilde{u}_i,\tilde{v}_i)$, $(\underset{\sim}{u}_i,\underset{\sim}{v}_i)$ *be upper and lower solutions such that* $(\tilde{u}_i,\tilde{v}_i)\geqq(\underset{\sim}{u}_i,\underset{\sim}{v}_i)\geqq(0,0)$ *and let hypothesis* (H) *hold. Then the sequences* $\{\bar{u}_i^{(k)},\bar{v}_i^{(k)}\}$, $\{\underline{u}_i^{(k)},\underline{v}_i^{(k)}\}$ *obtained from* (2.4)-(2.7) *possess the monotone property*

$$
(2.8)\qquad (\underline{u}_i^{(k)},\underline{v}_i^{(k)})\leqq(\underline{u}_i^{(k+1)},\underline{v}_i^{(k+1)})\leqq(\bar{u}_i^{(k+1)},\bar{v}_i^{(k+1)})\leqq(\bar{u}_i^{(k)},\bar{v}_i^{(k)}),\qquad k=0,1,2,\cdots.
$$

Moreover, for each fixed k, the pair $(\bar{u}_i^{(k)},\bar{v}_i^{(k)})$ *and* $(\underline{u}_i^{(k)},\underline{v}_i^{(k)})$ *are also upper and lower solutions of* (1.1)-(1.3).

Proof. The proof of the monotone property follows along the same line as in [8] and we give a sketch as follows: Let $\bar{w}_i=\bar{u}_i^{(0)}-\bar{u}_i^{(1)}=\tilde{u}_i-\bar{u}_i^{(1)}$, $\bar{z}_i=\bar{v}_i^{(0)}-\bar{v}_i^{(1)}=\tilde{v}_i-\bar{v}_i^{(1)}$. Then by (2.4) and (2.1), $(\bar{w}_i,\bar{z}_i)$ satisfies the relation

$$
(2.9)\qquad
\begin{aligned}
&L_1[\bar{w}_1]=L_1[\tilde{u}_1]-(a_1\tilde{u}_2+f(\underset{\sim}{v}_1(x,t-r_1))\geqq 0,\\
&\mathscr{L}_1[\bar{z}_1]=\mathscr{L}_1[\tilde{v}_1]-a_2\tilde{v}_2\geqq 0,\\
&L_2[\bar{w}_2]=L_2[\tilde{u}_2]-L_2[\bar{u}_2^{(1)}]\geqq 0,\\
&\mathscr{L}_2[\bar{z}_2]=\mathscr{L}_2[\tilde{v}_2]-g(\tilde{u}_2(x,t-r_2))\geqq 0.
\end{aligned}
$$

It is easily seen from the boundary and initial requirement on $(\tilde{u}_i,\tilde{v}_i)$ and (2.6) that

$$
(2.10)\qquad
\begin{aligned}
&B[\bar{w}_2]=B[\tilde{u}_2]-\beta_1\tilde{u}_1\geqq 0, \qquad \mathscr{B}[\bar{z}_2]=\mathscr{B}(\tilde{v}_2]-\beta_2\tilde{v}_1\geqq 0,\\
&\frac{\partial \bar{w}_2}{\partial \nu}=\frac{\partial \tilde{u}_2}{\partial \nu}\geqq 0, \qquad \frac{\partial \bar{z}_2}{\partial z}=\frac{\partial \tilde{v}_2}{\partial \nu}\geqq 0
\end{aligned}
$$

and

$$
(2.11)\qquad
\begin{aligned}
&\bar{w}_1(x,0)=\tilde{u}_1(x,0)-\xi_1(x)\geqq 0, \qquad \bar{z}_1(x,t)=\tilde{v}_1(x,t)-\eta_1(x,t)\geqq 0,\\
&\bar{w}_2(x,t)=\tilde{u}(x,t)-\xi_2(x,t)\geqq 0 \qquad \bar{z}_2(x,0)=\tilde{v}_2(x,0)-\eta_2(x)\geqq 0.
\end{aligned}
$$

The relations (2.9)-(2.11) and the maximum principle ensure that $\bar{w}_i\geqq 0$, $\bar{z}_i\geqq 0$. This proves $(\bar{u}_i^{(1)},\bar{v}_i^{(1)})\leqq(\bar{u}_i^{(0)},\bar{v}_i^{(0)})$. By the property of a lower solution and the monotone property of f and g the same argument shows that

$$(\underline{u}_i^{(0)}, \underline{v}_i^{(0)}) \leqq (\underline{u}_i^{(1)}, \underline{v}_i^{(1)}) \leqq (\bar{u}_i^{(1)}, \bar{v}_i^{(1)}) \leqq (\bar{u}_i^{(0)}, \bar{v}_i^{(0)}).$$

Finally, the conclusion in (2.8) follows by induction.

To show that $(\bar{u}_i^{(k)}, \bar{v}_i^{(k)})$ and $(\underline{u}_i^{(k)}, \underline{v}_i^{(k)})$ are upper-lower solutions for each k, we observe from hypothesis (H) and (2.8) that

$$\begin{aligned} &f(\underline{v}_1^{(k-1)}) \geqq f(\underline{v}_1^{(k)}), \quad f(\bar{v}_1^{(k-1)}) \leqq f(\bar{v}_1^{(k)}), \\ &g(\underline{u}_2^{(k-1)}) \leqq g(\underline{u}_2^{(k)}), \quad g(\bar{u}_2^{(k-1)}) \geqq g(\bar{u}_2^{(k)}), \end{aligned} \qquad k = 1, 2, \cdots. \tag{2.12}$$

This relation and (2.4), (2.8) imply that

$$\begin{aligned} &L_1[\bar{u}_1^{(k)}] = a_1 \bar{u}_2^{(k-1)} + f(\underline{v}_1^{(k-1)}(x, t-r_1)) \geqq a_1 \bar{u}_2^{(k)} + f(\underline{v}_1^{(k)}(x, t-r_1)), \\ &\mathscr{L}_1[\bar{v}_1^{(k)}] = a_2 \bar{v}_2^{(k-1)} \geqq a_2 \bar{v}_2^{(k)}, \\ &L_2[\bar{u}_2^{(k)}] = 0, \\ &\mathscr{L}_2[\bar{v}_2^{(k)}] = g(\bar{u}_2^{(k-1)}(x, t-r_2)) \geqq g(\bar{u}_2^{(k)}(x, t-r_2)). \end{aligned} \tag{2.13}$$

By the same reasoning, a similar set of reversed inequalities hold for $(\underline{u}_i^{(k)}, \underline{v}_i^{(k)})$. Since the boundary and initial requirements in (2.2), (2.3) are trivially satisfied, we conclude that $(\bar{u}_i^{(k)}, \bar{v}_i^{(k)})$ and $(\underline{u}_i^{(k)}, \underline{v}_i^{(k)})$ are upper-lower solutions. This proves the lemma.

In view of the relation (2.8) the pointwise limits

$$\lim_{k\to\infty} (\bar{u}_i^{(k)}, \bar{v}_i^{(k)}) = (\bar{u}_i, \bar{v}_i), \qquad \lim_{k\to\infty} (\underline{u}_i^{(k)}, \underline{v}_i^{(k)}) = (\underline{u}_i, \underline{v}_i) \tag{2.14}$$

exist and satisfy the relation

$$(\underline{u}_i^{(k)}, \underline{v}_i^{(k)}) \leqq (\underline{u}_i, \underline{v}_i) \leqq (\bar{u}_i, \bar{v}_i) \leqq (\bar{u}_i^{(k)}, \bar{v}_i^{(k)}), \qquad k = 1, 2, \cdots. \tag{2.15}$$

By letting $k \to \infty$ in (2.4)-(2.7) a standard regularity argument shows that both $(\bar{u}_i, \bar{v}_i)$ and $(\underline{u}_i, \underline{v}_i)$ satisfy all the equations in (1.1)-(1.3) except with the first equation in (1.1) replaced by

$$\begin{aligned} &L_1[\bar{u}_1] = a_1 \bar{u}_2 + f(\underline{v}_1(x, t-r_1)) \quad \text{for } (\bar{u}_i, \bar{v}_i), \\ &L_1[\underline{u}_1] = a_1 \underline{u}_2 + f(\bar{v}_1(x, t-r_1)) \quad \text{for } (\underline{u}_i, \underline{v}_i) \end{aligned} \tag{2.16}$$

(cf. [2], [9]). In order to ensure that $(\bar{u}_i, \bar{v}_i)$ and $(\underline{u}_i, \underline{v}_i)$ are solutions of (1.1)-(1.3), we need to show $(\bar{u}_i, \bar{v}_i) = (\underline{u}_i, \underline{v}_i)$. This is given in the following existence-comparison theorem.

THEOREM 2.1. *Let $(\tilde{u}_i, \tilde{v}_i)$, $(\underset{\sim}{u}_i, \underset{\sim}{v}_i)$ be upper and lower solutions such that $(\tilde{u}_i, \tilde{v}_i) \geqq (\underset{\sim}{u}_i, \underset{\sim}{v}_i) \geqq (0, 0)$ and let hypothesis* (H) *hold. Then there exists a unique global solution (u_i, v_i) to the problem* (1.1)-(1.3). *Moreover, the maximum and minimum sequences $\{\bar{u}_i^{(k)}, \bar{v}_i^{(k)}\}$, $\{\underline{u}_i^{(k)}, \underline{v}_i^{(k)}\}$ converge monotonically to the solution (u_i, v_i) and*

$$(\underset{\sim}{u}_i, \underset{\sim}{v}_i) \leqq (\underline{u}_i^{(k)}, \underline{v}_i^{(k)}) \leqq (u_i, v_i) \leqq (\bar{u}_i^{(k)}, \bar{v}_i^{(k)}) \leqq (\tilde{u}_i, \tilde{v}_i), \qquad k = 1, 2, \cdots. \tag{2.17}$$

Proof. Let $w_i = \bar{u}_i - \underline{u}_i$, $z_i = \bar{v}_i - \underline{v}_i$. Then (w_i, z_i) satisfies the differential equation

$$(2.18)\qquad \begin{aligned} &L_1[w_1] = a_1 w_2 + f(\underline{v}_1(x, t-r_1)) - f(\bar{v}_1(x, t-r_1)) \\ &\mathscr{L}_1[z_1] = a_2 z_2 \end{aligned} \qquad ((x,t)\in S_1),$$
$$\begin{aligned} &L_2[w_2] = 0 \\ &\mathscr{L}_2[z_2] = g(\bar{u}_2(x, t-r_2)) - g(\underline{u}_2(x, t-r_2)) \end{aligned} \qquad ((x,t)\in D_T)$$

the boundary condition

$$(2.19)\qquad \begin{aligned} &B[w_2] = \beta_1 w_1, \qquad \mathscr{B}[z_2] = \beta_2 z_1 \qquad ((x,t)\in S_1), \\ &\frac{\partial w_2}{\partial \nu} = \frac{\partial z_2}{\partial \nu} = 0 \qquad ((x,t)\in S_2) \end{aligned}$$

and the initial condition

$$(2.20)\qquad \begin{aligned} &w_1(x,0) = z_1(x,t) = 0 \qquad (x\in\Gamma_i,\ t\in[-r_1, 0]), \\ &w_2(x,t) = z_2(x,0) = 0 \qquad (x\in\Omega,\ t\in[-r_2, 0]). \end{aligned}$$

In view of Lemma 2.1 and (2.14) it suffices to show that $(w_i, z_i) = (0, 0)$ on $\bar{D}_T$. Clearly the equations for w_1 and z_1 are equivalent to the integral equation

$$(2.21)\qquad \begin{aligned} &w_1(x,t) = \int_0^t e^{-\alpha_1(t-\tau)}[a_1 w_2(x,\tau) + f(\underline{v}_1(x, \tau-r_1)) - f(\bar{v}_1(x,\tau-r_1))]\,d\tau, \\ &z_1(x,t) = a_2\int_0^t e^{-\alpha_2(t+\tau)} z_2(x,\tau)\,d\tau \end{aligned}$$

where $\alpha_i = a_i + b_i$, $i = 1, 2$. To obtain an integral representation for (w_2, z_2), we make use of the Green's function $\hat{G} \equiv \hat{G}(x, t|\xi, \tau)$ which is governed by the equations

$$(2.22)\qquad \begin{aligned} &\hat{L}[G] \equiv G_t - D\nabla^2 G + bG = \delta(x-\xi)\delta(t-\tau) \qquad ((x,t)\in D_T), \\ &\hat{B}[G] \equiv \frac{\partial G}{\partial \nu} + \beta G = 0 \qquad ((x,t)\in S_1), \\ &\frac{\partial G}{\partial \nu} = 0 \qquad ((x,t)\in S_2), \\ &G(x, t|\xi,\tau) = 0 \qquad (x\in\bar{\Omega},\ t<\tau), \end{aligned}$$

where D, b, and β are positive constants, δ is the Dirac δ-function and (ξ, τ) is a fixed point in D_T. This Green's function can be expressed in the form

$$\hat{G}(x, t|\xi,\tau) = \Gamma^*(x, t|\xi,\tau) + V(x, t|\xi, \tau)$$

where Γ^* is the fundamental solution of $\hat{L}$ given by (cf. [2, 10])

$$\Gamma^*(x, t|\xi,\tau) = (4\pi D(t-\tau))^{-n/2}\exp\,[-(bt + |x-\xi|^2/4D|t-\tau|)] \qquad (t>\tau)$$

and V is the solution of the problem

$$(2.23)\qquad \begin{aligned} &\hat{L}[V] = 0 \qquad ((x,t)\in D_T), \\ &\hat{B}[V] = -\hat{B}[\Gamma^*] \quad ((x,t)\in S_1), \quad \frac{\partial V}{\partial \nu} = -\frac{\partial \Gamma^*}{\partial \nu} \quad ((x,t)\in S_2), \\ &V(x, t|\xi,\tau) \equiv 0 \qquad (x\in\Omega,\ t\leqq\tau). \end{aligned}$$

The function Γ^* has a weak singular point at $(x, t) = (\xi, \tau)$ while V is a bounded smooth function in D_T. It is easily seen that by writing Γ^* in the form

$$\Gamma^*(x, t|\xi,\tau) = (\pi^{-n/2}e^{-bt})\frac{(4D(t-\tau)^{-\mu})}{|x-\xi|^{n-2\mu}}\left[\left(\frac{|x-\xi|^2}{4D(t-\tau)}\right)^{(n-2\mu)/2}\exp\left(-\frac{|x-\xi|^2}{4D(t-\tau)}\right)\right],$$

where $\mu \in (0, 1)$ is a fixed constant, there exists a constant K, independent of (x, t), such that

$$\Gamma^*(x, t|\xi, \tau) \leqq K(t-\tau)^{-\mu}|x-\xi|^{-(n-2\mu)}. \tag{2.24}$$

Let G_1, G_2 be the Green's function corresponding to $(\hat{L}, \hat{B}) = (L, B)$ and $(\hat{L}, \hat{B}) = (\mathscr{L}, \mathscr{B})$, respectively. Then the integral representation for (w_2, z_2) is given by

$$\begin{aligned} w_2(x, t) &= \beta_1 D_1 \int_0^t \int_{\Gamma_1} G_1(x, t|\xi, \tau) w_1(\xi, \tau)\, d\xi\, d\tau, \\ z_2(x, t) &= \int_0^t \int_\Omega G_2(x, t|\xi, \tau)[g(\bar{u}_2(\xi, \tau - r_2)) - g(\underline{u}_2(\xi, \tau - r_2))]\, d\xi\, d\tau \\ &\quad + \beta_2 D_2 \int_0^t \int_{\Gamma_1} G_2(x, t|\xi, \tau) z_1(\xi, \tau)\, d\xi\, d\tau. \end{aligned} \tag{2.25}$$

Hence the system (2.18)-(2.20) is reduced to the integral system (2.21) and (2.25). Our aim is to show that the only solution to this integral system is the trivial solution $(w_i, z_i) = (0, 0)$.

Let M_0 be a common upper bound of $f'(v_1)$ and $g'(u_2)$ on $[\underline{v}_1, \bar{v}_1]$ and $[\underline{u}_2, \bar{u}_2]$, respectively. For each fixed $t > 0$ let $\|w\|_t$, $\|z\|_t$ be the respective sup-norm of w and z on $\Omega \times [0, t]$ and $\Gamma_1 \times [0, t]$. In view of $\bar{v}_1(x, t) = \underline{v}_1(x, t)$ on $\Gamma_1 \times [-r_1, 0]$ and $\bar{u}_2(x, t) = \underline{u}_2(x, t)$ on $\bar{\Omega} \times [-r_2, 0]$, f and g satisfy the estimate

$$\begin{aligned} |f(\underline{v}(x, t - r_1)) - f(\bar{v}_1(x, t - r_1))| &\leqq M_0 \|\bar{v}_1 - \underline{v}_1\|_t = M_0 \|z_1\|_t, \\ |g(\bar{u}_2(x, t - r_2)) - g(\underline{u}_2(x, t - r_2))| &\leqq M_0 \|\bar{u}_2 - \underline{u}_2\|_t = M_0 \|w_2\|_t. \end{aligned} \tag{2.26}$$

Since by (2.24)

$$\begin{aligned} |G_i(x, t|\xi, \tau)| &\leqq |\Gamma_i^*(x, t|\xi, \tau)| + |V_i(x, t|\xi, \tau)| \\ &\leqq K_i(t-\tau)^{-\mu}|x-\xi|^{-n+2\mu} + M_i \qquad (i = 1, 2), \end{aligned}$$

where M_i is an upper bound of V_i, an elementary calculation shows that for some constant M_3, independent of $(x, t; \xi, \tau)$,

$$\begin{aligned} &\int_0^t \int_{\Gamma_1} |G_i(x, t|\xi, \tau)|\, d\xi\, d\tau \leqq M_3(t^{1-\mu} + t) \qquad (i = 1, 2), \\ &\int_0^t \int_\Omega |G_2(x, t|\xi, \tau)|\, d\xi\, d\tau \leqq M_3(t^{1-\mu} + t). \end{aligned} \tag{2.27}$$

Using the estimates (2.26), (2.27) in (2.21), (2.25), we obtain

$$\begin{aligned} |w_1(x, t)| &\leqq \alpha_1^{-1}(1 - e^{-\alpha_1 t})[a_1 \|w_2\|_t + M_0 \|z_1\|_t] \\ &\leqq M(1 - e^{-\alpha_1 t})(\|w_2\|_t + \|z_1\|_t), \\ |z_1(x, t)| &\leqq M(1 - e^{-\alpha_2 t}) \|z_2\|_t, \\ |w_2(x, t)| &\leqq M(t^{1-\mu} + t) \|w_1\|_t, \\ |z_2(x, t)| &\leqq M(t^{1-\mu} + t)(\|w_2\|_t + \|z_1\|_t), \end{aligned} \tag{2.28}$$

where M is a constant independent of (x, t). Define

$$\|W\|_t = \|w_1\|_t + \|z_1\|_t + \|w_2\|_t + \|z_2\|_t.$$

Then the relation (2.28) implies that

$$\|W\|_t \leqq M^*(1 - e^{-\alpha t} + t^{1-\mu} + t) \|W\|_t \tag{2.29}$$

for some constants M^*, α. Let $t_1>0$ be any constant such that $M^*(1-e^{-\alpha t_1}+t_1^{1-\mu}+t_1)<1$. Then by (2.29), $\|W\|_{t_1}=0$. But since $\|W\|_t$ is a nondecreasing function of t we conclude that $\|W\|_t=0$ for all $t\in[0,t_1]$. This proves $(\tilde{u}_i,\tilde{v}_i)=(\underline{u}_i,\underline{v}_i)$ on $\bar{\Omega}\times[0,t_1]$. Using $(w_i(t_1),z_i(t_1))=(0,0)$ as the initial condition, a continuation of the above argument leads to the conclusion $(\tilde{u}_i,\tilde{v}_i)=(\underline{u}_i,\underline{v}_i)$ on $\bar{D}_T$. The above argument also shows that $(\tilde{u}_i,\tilde{v}_i)=(\underline{u}_i,\underline{v}_i)$ on $\bar{D}_T$ is the unique solution of (1.1)-(1.3). This completes the proof of the theorem.

It is seen from Theorem 2.1 that the existence of a global solution is ensured if there exists an ordered pair of upper and lower solutions. Such a pair of functions can be obtained in the form $(\tilde{u}_i,\tilde{v}_i)=(\rho,\rho^*)$ and $(\underline{u}_i,\underline{v}_i)=(0,0)$, where ρ and ρ^* are some positive constants. Indeed, these constant functions fulfill all the requirements in Definition 2.1 if ρ and ρ^* satisfy the inequalities

$$
\begin{aligned}
&(a_1+b_1)\rho-(a_1\rho+f(0))\geqq 0\geqq -f(\rho^*),\\
&b_2\rho^*-g(\rho)\geqq 0\geqq -g(0),\\
&\rho-\xi_1\geqq 0,\qquad \rho^*-\eta_1(x,t)\geqq 0\qquad (t\in[-r_1,0]),\\
&\rho-\xi_2(x,t)\geqq 0,\qquad \rho^*\geqq \eta_2(x)\qquad (x\in\Omega,\ t\in[-r_2,0])
\end{aligned}
\tag{2.30}
$$

since all the other inequalities in (2.1)-(2.3) are trivially satisfied. By hypothesis (H), the relations in (2.30) hold if ρ and ρ^* are chosen such that

$$
\rho\geqq\max\left\{\bar{\xi}_1,\bar{\xi}_2,\frac{f(0)}{b_1}\right\},\qquad \rho^*\geqq\max\left\{\bar{\eta}_1,\bar{\eta}_2,\frac{g(\rho)}{b_2}\right\}
\tag{2.31}
$$

where $\bar{\xi}_i$ and $\bar{\eta}_i$ are the least upper bounds of ξ_i, η_i in their respective domains. With this choice of ρ, ρ^*, $(\tilde{u}_i,\tilde{v}_i)=(\rho,\rho^*)$ and $(\underline{u}_i,\underline{v}_i)=(0,0)$ are upper-lower solutions. As an application of Theorem 2.1 we have the following conclusion.

THEOREM 2.2. *Let f, g satisfy hypothesis* (H). *Then the problem* (1.1)-(1.3) *has a unique nonnegative global solution (u_i,v_i). Moreover this solution is uniformly bounded by (ρ,ρ^*), where ρ and ρ^* are any constants satisfying* (2.31).

3. Asymptotic stability. The construction of constant upper-lower solutions in § 2 ensures the existence of a unique bounded solution to (1.1)-(1.3). In order to investigate the asymptotic behavior of the solution for large values of t, we need to find a different pair of upper-lower solutions. In this section we study the stability problem of the system (1.1)-(1.3) for a given steady-state solution. The existence of steady-state solutions will be discussed in the following section. Here by a steady-state solution we mean a time-independent function $(u_i^s,v_i^s)\equiv(u_1^s,v_1^s;u_2^s,v_2^s)$ which satisfies (1.1), (1.2) without the time-derivative term.

THEOREM 3.1. *Let $(u_i^s(x),v_i^s(x))$ be a nonnegative steady-state solution of* (1.1), (1.2) *and let hypothesis (H) hold. If*

$$
-f'(v_1^s)g'(u_2^s)<b_1b_2\qquad (x\in\bar{\Omega})
\tag{3.1}
$$

then (u_i^s,v_i^s) is asymptotically stable. If, in addition,

$$
\sup_{n\geqq 0}[-f'(\eta)]\sup_{\eta\geqq 0}[g'(\eta)]<b_1b_2
\tag{3.2}
$$

then (u_i^s,v_i^s) is globally asymptotically stable.

Proof. Let $p(t)$, $q(t)$ be some positive functions defined in $\mathbb{R}^1$ and let

$$
(\tilde{u}_i,\tilde{v}_i)=(u_i^s+p,\ v_i^s+q),\qquad (\underline{u}_i,\underline{v}_i)=(u_i^s-p,\ v_i^s-q).
\tag{3.3}
$$

Our aim is to find p, q such that $(p(t), q(t)) \to (0,0)$ and the pair given by (3.3) are upper-lower solutions. To achieve this, we observe that (u_i^s+p, v_i^s+q) is an upper solution if (p, q) satisfies the differential inequalities

$$\begin{aligned}
&L_1[u_1^s]+p'+(a_1+b_1)p-a_1(u_2^s+p)-f(v_1^s-q(t-r_1)) \geqq 0,\\
&\mathscr{L}_1[v_1^s]+q'+(a_2+b_2)q-a_2(v_2^s+q) \geqq 0,\\
&L_2[u_2^s]+p'+b_1p \geqq 0,\\
&\mathscr{L}_2[v_2^s]+q'+b_2q \geqq g(u_2^s+p(t-r_2))
\end{aligned}\tag{3.4}$$

and the initial inequalities

$$\begin{aligned}
&u_1^s(x)+p(0) \geqq \xi_1(x), \qquad v_1^s(x)+q(t) \geqq \eta_1(x,t) \qquad (x \in \Gamma_1, t \in [-r_1, 0]),\\
&u_2^s(x)+p(t) \geqq \xi_2(x,t), \qquad v_2^s(x)+q(0) \geqq \eta_2(x) \qquad (x \in \Omega, t \in [-r_2, 0]).
\end{aligned}\tag{3.5}$$

The boundary inequalities are trivially fulfilled since (u_i^s, v_i^s) satisfies the boundary condition (1.2). Using the property of a steady-state solution, relation (3.4) is equivalent to

$$\begin{aligned}
&p'+b_1p \geqq f(v_1^s-q(t-r_1))-f(v_1^s),\\
&q'+b_2q \geqq 0,\\
&p'+b_1p \geqq 0,\\
&q'+b_2q \geqq g(u_2^s+p(t-r_2))-g(u_2^s).
\end{aligned}\tag{3.6}$$

By the same reasoning, (u_i^s-p, v_i^s-q) is a lower solution if all the reversed inequalities in (3.4), (3.5) are satisfied when p, q are replaced by $-p$ and $-q$, respectively. This implies that p and q must satisfy the additional requirements

$$\begin{aligned}
&p'+b_1p \geqq f(v_1^s)-f(v_1^s+q(t-r_1)),\\
&q'+b_2q \geqq g(u_2^s)-g(u_2^s-p(t-r_2)),
\end{aligned}\tag{3.7}$$

$$\begin{aligned}
&u_1^s(x)-p(0) \leqq \xi_1(x), \quad v_1^s(x)-q(t) \leqq \eta_1(x,t) \quad (x \in \Gamma_1, t \in [-r_1, 0]),\\
&u_2^s(x)-p(t) \leqq \xi_2(x,t), \quad v_2^s(x)-q(0) \leqq \eta_2(x) \quad (x \in \Omega, t \in [-r_2, 0]).
\end{aligned}\tag{3.8}$$

For any positive constants ρ_1, ρ_2, define $M_1 = M_1(\rho_1)$, $M_2 = M_2(\rho_2)$ by

$$\begin{aligned}
&M_1 \equiv \max\{-f'(v_1^s+\eta);\ |\eta| \leqq \rho_1,\ v_1^s+\eta \geqq 0,\ x \in \Gamma_1\},\\
&M_2 \equiv \max\{g'(u_2^s+\eta);\ |\eta| \leqq \rho_2,\ u_2^s+\eta \geqq 0,\ x \in \bar{\Omega}\}.
\end{aligned}\tag{3.9}$$

Then all the inequalities in (3.6) and (3.7) are fulfilled if

$$\left.\begin{aligned}
&p'+b_1p \geqq M_1q(t-r_1), \quad q(t-r_1) \leqq \rho_1\\
&q'+b_2q \geqq M_2p(t-r_2), \quad p(t-r_2) \leqq \rho_2
\end{aligned}\right. \quad (t>0).\tag{3.10}$$

We choose ρ_1, ρ_2 such that $M_1M_2 < b_1b_2$. This is possible by virtue of condition (3.1). To satisfy the relation (3.10) we let p, q in the form

$$p(t) = p_0\, e^{-\varepsilon t}, \qquad q(t) = q_0\, e^{-\varepsilon t}$$

for some $\varepsilon > 0$. Then the inequalities in (3.10) hold if

$$\left.\begin{aligned}
&(b_1-\varepsilon)p_0 \geqq M_1q_0\, e^{\varepsilon r_1}, \quad q(t-r_1) \leqq \rho_1\\
&(b_2-\varepsilon)q_0 \geqq M_2p_0\, e^{\varepsilon r_2}, \quad p(t-r_2) \leqq \rho_2
\end{aligned}\right. \quad (t>0).$$

Let $p_0 = \rho_1\, e^{-\varepsilon r_1}$, $q_0 = \rho_2\, e^{-\varepsilon r_2}$ so that $p(t-r_1) \leqq \rho_1$, $q(t-r_2) \leqq \rho_2$ for $t>0$. Then it suffices to find ρ_1, ρ_2 such that

$$\frac{M_1\, e^{\varepsilon(2r_1-r_2)}}{b_1-\varepsilon} \leqq \frac{\rho_1}{\rho_2} \leqq \frac{b_2-\varepsilon}{M_2\, e^{\varepsilon(2r_2-r_1)}}.\tag{3.11}$$

This is clearly possible for a sufficiently small $\varepsilon>0$ since $M_1M_2<b_1b_2$. The above construction shows that the functions given by (3.3) with $p(t)=\rho_1 e^{-\varepsilon(t+r_1)}$, $q(t)=\rho_2 e^{-\varepsilon(t+r_2)}$ are upper and lower solutions whenever the initial functions satisfy the relation

$$\begin{aligned} &|u_1^s-\xi_1|\leqq \rho_1 e^{-\varepsilon r_1}, && |v_1^s-\eta_1|\leqq \rho_2 e^{-\varepsilon(t+r_2)} && (x\in\Gamma_1,\ t\in[-r,0]),\\ &|u_2^s-\xi_2|\leqq \rho_1 e^{-\varepsilon(t+r_1)}, && |v_2^s-\eta_2|\leqq \rho_2 e^{-\varepsilon r_2} && (x\in\Omega,\ t\in[-r_2,0]). \end{aligned}\tag{3.12}$$

It follows from Theorem 2.1 that the time-dependent solution (u_i, v_i) satisfies the relation

$$\begin{aligned} u_i^s(x)-\rho_1 e^{-\varepsilon(t+r_1)}\leqq u_i(t,x)\leqq u_i^s(x)+\rho_1 e^{-\varepsilon(t+r_1)}\\ v_i^s(x)-\rho_2 e^{-\varepsilon(t+r_1)}\leqq v_i(t,x)\leqq v_i^s(x)+\rho_2 e^{-\varepsilon(t+r_2)} \end{aligned}\qquad (t>0,\ x\in\bar{\Omega})\tag{3.13}$$

for a sufficiently small $\varepsilon>0$. The above relation ensures that (u_i^s, v_i^s) is asymptotically stable. Now if condition (3.2) holds then $M_1M_2<b_1b_2$ for all positive constants ρ_1, ρ_2. In this situation the relations (3.11) and (3.12) can be satisfied by choosing ρ_1, ρ_2 sufficiently large. This implies that (u_i^s, v_i^s) is globally asymptotically stable. The proof of the theorem is completed.

Remark 3.1. When ρ_1, ρ_2 are large it is possible that the lower solution (u_i^s-p, v_i^s-q) takes negative values. In this situation we need to define some modified functions $\hat{f}, \hat{g}$ in place of f, g and using the same argument as in [8] to show that all the conclusions in Theorem 3.1 remain true.

4. The steady-state problem. The monotone method for the time-dependent problem (1.1)–(1.3) can be used to construct similar maximal and minimal sequences for the corresponding steady-state problem

$$\begin{aligned} &(a_1+b_1)u_1=a_1u_2+f(v_1)\\ &(a_2+b_2)v_1=a_2v_2 && (x\in\Gamma_1),\\ &-D_1\nabla^2u_2+b_1u_2=0\\ &-D_2\nabla^2v_2+b_2v_2=g(u_2) && (x\in\Omega). \end{aligned}\tag{4.1}$$

By solving the first two equations in (4.1) for (u_1, v_1) and substituting into the boundary condition (1.2), the steady-state solution must satisfy the boundary condition

$$\begin{aligned} &\frac{\partial u_2}{\partial \nu}+(\beta_1b_1/(a_1+b_1))u_2=(\beta_1/(a_1+b_1))f(a_2v_2/(a_2+b_2)),\\ &\frac{\partial v_2}{\partial \nu}+(\beta_2b_2/(a_2+b_2))v_2=0 \qquad (x\in\Gamma_1),\\ &\frac{\partial u_2}{\partial \nu}=\frac{\partial v_2}{\partial \nu}=0 \qquad (x\in\Gamma_2). \end{aligned}\tag{4.2}$$

Let $u=u_2$, $v=v_2$ and define

$$\begin{aligned} &F(v)=(\beta_1/(a_1+b_1))f(a_2v/(a_2+b_2)), \qquad G(u)=\frac{g(u)}{D_2},\\ &c_i=\frac{b_i}{D_i}, \qquad \gamma_i=\frac{\beta_ib_i}{a_i+b_i} \qquad (i=1,2). \end{aligned}\tag{4.3}$$

Then the steady-state problem is reduced to the coupled system

$$-\nabla^2u+c_1u=0, \quad -\nabla^2v+c_2v=G(u) \quad (x\in\Omega),\tag{4.4}$$

$$\begin{aligned}&B_1[u]\equiv\frac{\partial u}{\partial\nu}+\gamma_1 u=F(v),\quad B_2[v]\equiv\frac{\partial v}{\partial\nu}+\gamma_2 v=0\quad(x\in\Gamma_1),\\ &\frac{\partial u}{\partial\nu}=\frac{\partial v}{\partial\nu}=0\qquad(x\in\Gamma_2).\end{aligned}\tag{4.5}$$

It is clear from (4.3) that $F(v)$ and $G(u)$ possess the same monotone property as $f(u)$, $g(v)$ in the hypothesis (H). In view of the quasi-monotone property of F and G we define upper and lower solutions for the problem (4.4)-(4.5) as follows:

DEFINITION 4.1. A pair of smooth functions $(\tilde{u},\tilde{v})$, $(\underset{\sim}{u},\underset{\sim}{v})$ are called upper and lower solutions of (4.4), (4.5) if they satisfy the relations

$$\left.\begin{aligned}&-\nabla^2\tilde{u}+c_1\tilde{u}\geqq 0\geqq-\nabla^2\underset{\sim}{u}+c_1\underset{\sim}{u}\\ &-\nabla^2\tilde{v}+c_2\tilde{v}-G(\tilde{u})\geqq 0\geqq-\nabla^2\underset{\sim}{v}+c_2\underset{\sim}{v}-G(\underset{\sim}{u})\end{aligned}\right.\qquad(x\in\Omega),\tag{4.6}$$

$$\begin{aligned}&B_1[\tilde{u}]-F(\underset{\sim}{v})\geqq 0\geqq B_1[\underset{\sim}{u}]-F(\tilde{v}),\qquad B_2[\tilde{v}]\geqq 0\geqq B_2[\underset{\sim}{v}]\qquad(x\in\Gamma_1).\\ &\frac{\partial\tilde{u}}{\partial\nu}\geqq 0\geqq\frac{\partial\underset{\sim}{u}}{\partial\nu},\quad\frac{\partial\tilde{v}}{\partial\nu}\geqq 0\geqq\frac{\partial\underset{\sim}{v}}{\partial\nu}\qquad(x\in\Gamma_2).\end{aligned}\tag{4.7}$$

In the above definition a smooth function is referred to as a continuous function in $\bar{\Omega}$ which is twice continuously differentiable in Ω and has outward normal derivative on $\partial\Omega$. Notice that upper and lower solutions for the boundary value problem are interrelated through the boundary condition.

Let $(\bar{u}^{(0)},\bar{v}^{(0)})=(\tilde{u},\tilde{v})$, $(\underline{u}^{(0)},\underline{v}^{(0)})=(\underset{\sim}{u},\underset{\sim}{v})$ be initial iterations and construct two sequences $\{\bar{u}^{(k)},\bar{v}^{(k)}\}$, $\{\underline{u}^{(k)},\underline{v}^{(k)}\}$ from the iteration process

$$\begin{aligned}&-\nabla^2\bar{u}^{(k)}+c_1\bar{u}^{(k)}=0,\qquad -\nabla^2\bar{v}^{(k)}+c_2\bar{v}^{(k)}=G(\bar{u}^{(k-1)})\qquad(x\in\Omega),\\ &B_1[\bar{u}^{(k)}]=F(\underline{v}^{(k-1)}),\qquad B_2[\bar{v}^{(k)}]=0\qquad(x\in\Gamma_1)\\ &\frac{\partial\bar{u}^{(k)}}{\partial\nu}=\frac{\partial\bar{v}^{(k)}}{\partial\nu}=0\qquad(x\in\Gamma_2)\end{aligned}\tag{4.8}$$

and

$$\begin{aligned}&-\nabla^2\underline{u}^{(k)}+c_1\underline{u}^{(k)}=0,\qquad -\nabla^2\underline{v}^{(k)}+c_2\underline{v}^{(k)}=G(\underline{u}^{(k-1)})\qquad(x\in\Omega),\\ &B_1[\underline{u}^{(k)}]=F(\bar{v}^{(k-1)}),\qquad B_2[\underline{v}^{(k)}]=0\qquad(x\in\Gamma_1),\\ &\frac{\partial\underline{u}^{(k)}}{\partial\nu}=\frac{\partial\underline{v}^{(k)}}{\partial\nu}=0\qquad(x\in\Gamma_2).\end{aligned}\tag{4.9}$$

We again refer to these two sequences as maximal and minimal sequence, respectively. Just as in the time-dependent problem the maximal and minimal sequences possess the following monotone properties.

LEMMA 4.1. *Let* $(\tilde{u},\tilde{v})$, $(\underset{\sim}{u},\underset{\sim}{v})$ *be upper-lower solutions of* (4.4), (4.5) *such that* $(\tilde{u},\tilde{v})\geqq(\underset{\sim}{u},\underset{\sim}{v})\geqq(0,0)$ *and let hypothesis* (H) *hold. Then the sequences* $\{\bar{u}^{(k)},\bar{v}^{(k)}\}$, $\{\underline{u}^{(k)},\underline{v}^{(k)}\}$ *given by* (4.8), (4.9) *possess the monotone property*

$$\begin{aligned}(\underline{u}^{(k)},\underline{v}^{(k)})&\leqq(\underline{u}^{(k+1)},\underline{v}^{(k+1)})\leqq(\bar{u}^{(k+1)},\bar{v}^{(k+1)})\\ &\leqq(\bar{u}^{(k)},\bar{v}^{(k)}),\qquad k=0,1,2,\cdots.\end{aligned}\tag{4.10}$$

Moreover, for each fixed k, the pair $(\bar{u}^{(k)},\bar{v}^{(k)})$ *and* $(\underline{u}^{(k)},\underline{v}^{(k)})$ *are upper-lower solutions.*

Proof. Since the proof of the lemma follows from the same argument as for Lemma 2.1 we omit the details.

In view of the monotone property (4.10) the pointwise limits

$$\lim_{k\to\infty} (\bar{u}^{(k)}, \bar{v}^{(k)}) = (\bar{u}, \bar{v}), \qquad \lim_{k\to\infty} (\underline{u}^{(k)}, \underline{v}^{(k)}) = (\underline{u}, \underline{v}) \tag{4.11}$$

exist and satisfy the relation

$$(\underline{u}^{(k)}, \underline{v}^{(k)}) \leqq (\underline{u}, \underline{v}) \leqq (\bar{u}, \bar{v}) \leqq (\bar{u}^{(k)}, \bar{v}^{(k)}), \qquad k = 0, 1, 2, \cdots. \tag{4.12}$$

By letting $k \to \infty$ in (4.8), (4.9) the usual regularity argument shows that $(\bar{u}, \bar{v})$ and $(\underline{u}, \underline{v})$ satisfy the equations

$$\begin{aligned}
&-\nabla^2 \bar{u} + c_1 \bar{u} = 0, && B_1[\bar{u}] = f(\underline{v}), && \frac{\partial \bar{u}}{\partial \nu} = 0,\\
&-\nabla^2 \bar{v} + c_2 \bar{v} = G(\bar{u}), && B_2[\bar{v}] = 0, && \frac{\partial \bar{v}}{\partial \nu} = 0,\\
&-\nabla^2 \underline{u} + c_1 \underline{u} = 0, && B_1[\underline{u}] = F(\bar{v}), && \frac{\partial \underline{u}}{\partial \nu} = 0,\\
&-\nabla^2 \underline{v} + c_2 \underline{v} = G(\underline{u}), && B_2[\underline{v}] = 0, && \frac{\partial \underline{v}}{\partial \nu} = 0.
\end{aligned} \tag{4.13}$$

In order to ensure that $(\bar{u}, \bar{v})$ and $(\underline{u}, \underline{v})$ are solutions of (4.4), (4.5), we must show that $(\bar{u}, \bar{v}) = (\underline{u}, \underline{v})$. For this purpose we need to impose some additional conditions on F and G. Specifically, by defining

$$M_f \equiv \max\{-F'(v); \utilde{v} \leqq v \leqq \tilde{v}\}, \qquad M_g \equiv \max\{G'(u); \utilde{u} \leqq u \leqq \tilde{u}\} \tag{4.14}$$

we have the following existence-uniqueness result.

THEOREM 4.1. *Let* $(\tilde{u}, \tilde{v}), (\utilde{u}, \utilde{v})$ *be upper-lower solutions of* (4.4), (4.5) *such that* $(\tilde{u}, \tilde{v}) \geqq (\utilde{u}, \utilde{v}) \geqq (0, 0)$ *and let hypothesis* (H) *hold. If*

$$M_f M_g \leqq c_1 \gamma_2 \tag{4.15}$$

then the steady-state problem (4.4), (4.5) *has a unique solution* (u, v) *such that*

$$(\utilde{u}, \utilde{v}) \leqq (u, v) \leqq (\tilde{u}, \tilde{v}) \qquad (x \in \bar{\Omega}). \tag{4.16}$$

Moreover the maximal and minimal sequences $\{\bar{u}^{(k)}, \bar{v}^{(k)}\}, \{\underline{u}^{(k)}, \underline{v}^{(k)}\}$ *converge monotonically to* (u, v).

Proof. Let $U = \bar{u} - \underline{u} \geqq 0$, $V = \bar{v} - \underline{v} \geqq 0$. Since by the mean value theorem

$$\begin{aligned}
F(\underline{v}) - F(\bar{v}) &= -F'(\hat{v}) V \equiv \sigma_1(x) V\\
G(\bar{u}) - G(\underline{u}) &= G'(\hat{u}) U \equiv \sigma_2(x) U
\end{aligned} \qquad (x \in \Omega), \tag{4.17}$$

where $\hat{u}$ and $\hat{v}$ are some intermediate values in $[\utilde{u}, \tilde{u}]$ and $[\utilde{v}, \tilde{v}]$ respectively, relation (4.13) implies that

$$-\nabla^2 U + c_1 U = 0, \qquad -\nabla^2 V + c_2 V = \sigma_2 U \qquad (x \in \Omega), \tag{4.18}$$

$$\begin{aligned}
&\frac{\partial U}{\partial \nu} + \gamma_1 U = \sigma_1 V, \qquad \frac{\partial V}{\partial \nu} + \gamma_2 V = 0 && (x \in \Gamma_1),\\
&\frac{\partial U}{\partial \nu} = \frac{\partial V}{\partial \nu} = 0 && (x \in \Gamma_2).
\end{aligned} \tag{4.19}$$

Let ρ_1, ρ_2 be any positive constants such that $M_g/c_1 \leqq \rho_1/\rho_2 \leqq \gamma_2/M_f$ and let $W = \rho_1 U + \rho_2 V$. The existence of ρ_1, ρ_2 is ensured by condition (4.15). In view of (4.18) and (4.19), W satisfies the relation

$$\begin{aligned} &-\nabla^2 W+(\rho_1 c_1-\rho_2\sigma_2)U+c_2\rho_2 V=0 && (x\in\Omega),\\ (4.20)\qquad &\frac{\partial W}{\partial \nu}+\rho_1\gamma_1 U+(\rho_2\gamma_2-\rho_1\sigma_1)V=0 && (x\in\Gamma_1),\\ &\frac{\partial W}{\partial \nu}=0 && (x\in\Gamma_2). \end{aligned}$$

Since by the choice of ρ_1, ρ_2,

$$\rho_1 c_1-\rho_2\sigma_2\geqq \rho_1 c_1-\rho_2 M_g\geqq 0, \qquad \rho_2\gamma_2-\rho_1\sigma_1\geqq\rho_2\gamma_2-\rho_1 M_f\geqq 0,$$

the nonnegative property of (U, V) implies that

$$(4.21)\qquad -\nabla^2 W\leqq 0\quad (x\in\Omega), \qquad \frac{\partial W}{\partial \nu}\leqq 0\quad (x\in\partial\Omega).$$

By the maximum principle and the nonnegative property of W we must have $W=0$ on $\bar{\Omega}$. This shows that $(\bar{u}, \bar{v})=(\underline{u}, \underline{v})$. The monotone convergence of the maximal and minimal sequences follows from Lemma 4.1.

To show the uniqueness of the solution, we let (u^*, v^*) be any solution of (4.4), (4.5) such that $(\underset{\sim}{u}, \underset{\sim}{v})\leqq(u^*, v^*)\leqq(\tilde{u}, \tilde{v})$. Then by the iteration process (4.8), (4.9) and hypothesis (H) the functions $(\bar{w}_1, \bar{w}_2)\equiv(\bar{u}^{(1)}-u^*, \bar{v}^{(1)}-v^*)$ and $(\underline{w}_1, \underline{w}_2)\equiv(u^*-\underline{u}_1, v^*-\underline{v}_1)$ satisfy the relations

$$\begin{aligned} &-\nabla^2\bar{w}_1+c_1\bar{w}_1=0, && B_1[\bar{w}_1]=F(\underline{v}^{(0)})-F(v^*)\geqq 0,\\ &-\nabla^2\bar{w}_2+c_2\bar{w}_2=G(\bar{u}^{(0)})-G(u^*)\geqq 0, && B_2[w_2]=0,\\ &-\nabla^2\underline{w}_1+c_1\underline{w}_1=0, && B_1[\underline{w}_1]=F(v^*)-F(\bar{v}^{(0)})\geqq 0,\\ &-\nabla^2\underline{w}_2+c_2\underline{w}_2=G(u^*)-G(\underline{u}^{(0)})\geqq 0, && B_2[\underline{w}_2]=0 \end{aligned}$$

and $\partial\bar{w}_i/\partial\nu=\partial\underline{w}_i/\partial\nu=0$ on Γ_2, $i=1,2$. These inequalities imply that $(\underline{u}^{(1)}, \underline{v}^{(1)})\leqq(u^*, v^*)\leqq(\bar{u}^{(1)}, \bar{v}^{(1)})$. An induction argument shows that

$$(4.22)\qquad (\underline{u}^{(k)}, \underline{v}^{(k)})\leqq(u^*, v^*)\leqq(\bar{u}^{(k)}, \bar{v}^{(k)}), \qquad k=1,2,\cdots.$$

Since the sequences $\{\bar{u}^{(k)}, \bar{v}^{(k)}\}$ and $\{\underline{u}^{(k)}, \underline{v}^{(k)}\}$ both converge to the solution (u, v) we conclude that $(u^*, v^*)=(u, v)$. This completes the proof of the theorem.

Remark 4.1. Unlike systems with both quasi-monotone increasing or quasi-monotone decreasing functions, the pair $(\tilde{u}, \tilde{v})$ and (u^*, v^*) (or (u^*, v^*) and $(\underset{\sim}{u}, \underset{\sim}{v})$) are not upper-lower solutions as defined in Definition 4.1. In fact, if $(\tilde{u}, \tilde{v})$ and (u^*, v^*) are used as the initial iterations in the process (4.8) and (4.9) the corresponding sequences are no longer monotone. Nevertheless each sequence contains a subsequence which consists of the same function (u^*, v^*).

In analogy to the time-dependent problem, the constant functions $(\tilde{u}, \tilde{v})=(\rho, \rho^*)$ and $(\underset{\sim}{u}, \underset{\sim}{v})=(0,0)$ are upper-lower solutions of (4.4), (4.5) if

$$(4.23)\qquad c_2\rho^*-G(\rho)\geqq 0\geqq -G(0), \qquad \gamma_1\rho-F(0)\geqq 0\geqq -F(\rho)$$

because all the other requirements in (4.6) and (4.7) are fulfilled. By hypothesis (H), it suffices to choose ρ, ρ^* such that

$$(4.24)\qquad \rho\geqq\frac{F(0)}{\gamma_1}, \qquad \rho^*\geqq\frac{G(\rho)}{c_2}.$$

This choice of (ρ, ρ^*) and Theorem 4.1 lead to the following.

THEOREM 4.2. *Let hypothesis* (H) *and condition* (4.15) *hold and let* ρ, ρ^* *be any constants satisfying* (4.24). *Then the sequences* $\{\bar{u}^{(k)}, \bar{v}^{(k)}\}$ *and* $\{\underline{u}^{(k)}, \underline{v}^{(k)}\}$ *with* $(\bar{u}^{(0)}, \bar{v}^{(0)}) = (\rho, \rho^*)$, $(\underline{u}^{(0)}, \underline{v}^{(0)}) = (0, 0)$ *converge to a unique solution* (u, v) *and*

$$(0, 0) \leqq (u, v) \leqq (\rho, \rho^*). \tag{4.25}$$

REFERENCES

[1] H. T. BANKS AND J. M. MAHAFFY, *Global asymptotic stability of certain models for protein synthesis and repression*, Quart. Appl. Math., 36 (1978), pp. 209-221.

[2] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[3] B. C. GOODWIN, *Temporal Organization in Cells*, Academic Press, New York, 1963.

[4] ———, *Oscillatory behavior in enzymatic control processes*, Adv. in Enzyme Reg., 3 (1965), pp. 425-439.

[5] N. MCDONALD, *Time lag in a model of a biochemical reaction sequence with end-product inhibition*, J. Theoret. Biol., 67 (1977), pp. 549-556.

[6] J. M. MAHAFFY AND C. V. PAO, *Models of genetic control by repression with time delays and spatial effects*, J. Math. Biol., 20 (1984), pp. 39-57.

[7] H. G. OTHMER, *The qualitative dynamics of a class of biochemical control circuits*, J. Math. Biol., 3 (1976), pp. 53-70.

[8] C. V. PAO AND J. M. MAHAFFY, *Qualitative analysis of a coupled reaction-diffusion model in biology with time delays*, J. Math. Anal. Appl., 109 (1985), pp. 153-169.

[9] C. V. PAO, *Reaction diffusion equations with nonlinear boundary conditions*, J. Nonlinear Anal. Theory Methods Appl., 5 (1981), pp. 1077-1094.

[10] I. STAKGOLD, *Boundary Value Problems of Mathematical Physics*, Vol. II, MacMillan, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

011

CHAOTIC SOLUTIONS OF SYSTEMS OF FIRST ORDER PARTIAL DIFFERENTIAL EQUATIONS*

ROBERT WOLFE† AND HEDLEY C. MORRIS‡

Abstract. We consider the system of partial differential equations

$$\frac{\partial \mathbf{u}}{\partial t}+(\mathbf{c}(\mathbf{x})\cdot\nabla)\mathbf{u}=\mathbf{f}(\mathbf{x},\mathbf{u}),$$

$$\mathbf{u}(t=0,\mathbf{x})=\mathbf{v}(\mathbf{x})$$

where $\mathbf{x}\in\Delta$, a compact connected subset of R^m, $\mathbf{u}\in R^n$, and $t\geqq 0$. This is a generalization of a problem investigated by Lasota [6] and Brunovsky [2], and the methods used are similar to those used by these authors.

Imposing certain conditions on $\mathbf{c}$ and $\mathbf{f}$, we show that for $\mathbf{v}$ in a certain subset of the phase space the solution uniformly approaches a limit set of dimension $n-1$. We also show that in another subset of the phase space there are both (a) dense orbits and (b) periodic orbits of all periods. Each of (a) and (b) leads to a proof of the existence of a type of chaos in this subset.

Key words. first order partial differential equations, semiflow, chaos

AMS(MOS) subject classification. 35B

1. Introduction. This paper is concerned with a particular type of chaotic behaviour that can occur in a function space. In order that the nature of this particular type of stochastic behaviour can be appreciated, it is necessary to present some background information concerning chaos in other types of dynamical systems. There are essentially three types of situation in which chaotic phenomena have been investigated by mathematicians. These are chaos in one dimension involving maps of the interval or the circle, mappings of R^n into itself and chaos in function spaces such as those associated with the equations of fluid mechanics. We consider each in turn.

Chaos in one dimension. "Chaos" is taken loosely to mean a state where nonperiodic motion is observed as a nontransient phenomena. We will give various definitions of chaos, and show how the concept has been developed and generalized in different ways. One of those ways leads to the particular form of chaotic motion considered in this paper.

The first rigorous definition of chaos was that used by Li and Yorke [7] for iterations of an interval map.

THEOREM (Li and Yorke). *Let J be an interval and $F: J\to J$ be continuous. Suppose there is a point $a\in J$ that satisfies either*

$$F^3(a)\leqq a<F(a)<F^2(a) \tag{1.1}$$

or

$$F^3(a)\geqq a>F(a)>F^2(a).$$

Then

1) *For every integer $k>0$ there is a point in J having period k.*

* Received by the editors September 28, 1984; accepted for publication February 25, 1986.

† School of Mathematics, Trinity College, Dublin, Republic of Ireland. The work of this author was supported by the Irish Department of Education and Trinity College.

‡ Department of Mathematics, Claremont Graduate School, Claremont, California 91711. Permanent address, Department of Mathematics and Computer Science, San Jose State University, San Jose, California 95192-0103.

2) *J has an uncountable subset S (called the scrambled set) that contains no periodic points and satisfies the following conditions:*

For each p, q distinct in S we have

$$\limsup_{n\to\infty} |F^n(p) - F^n(q)| > 0 \tag{1.2}$$

and

$$\liminf_{n\to\infty} |F^n(p) - F^n(q)| = 0. \tag{1.3}$$

For each $p \in S$ and each periodic $q \in J$ we have

$$\limsup_{n\to\infty} |F^n(p) - F^n(q)| > 0. \tag{1.4}$$

This "period three implies chaos" theorem makes the definition attractive, and it has been widely used. Li and Yorke's work is partly a special case of a theorem of Sharkovskii [10].

Kloeden et al. [5] have proposed that the existence of orbits of any given period should not be a necessary condition in the definition of chaos. Ott [9] uses a definition of chaos for sequences based on sensitivity to initial conditions and on average correlation functions.

Chaos in R^n. How to characterise the n-dimensional analogue of the chaotic behaviour of one-dimensional systems is not immediately apparent. Marotto [8] has extended the Li-Yorke theorem to R^n by introducing the idea of "snap-back-repellers." If F is a C^1 mapping from R^n to R^n, then $\mathbf{z} \in R^n$ is defined to be a snap-back-repeller if (a) $F(\mathbf{z}) = \mathbf{z}$; (b) there is some $r > 0$ such that all the eigenvalues of $DF(\mathbf{x})$ have norm greater than one; (c) there is some point $\mathbf{x}_0$ with $0 < |\mathbf{z} - \mathbf{x}_0| < r$ such that $F^M(\mathbf{x}_0) = \mathbf{z}$ and $|DF^M(\mathbf{x}_0)| \neq 0$ for some $M > 0$.

Marotto has shown that if a mapping F has a "snap-back-repeller," then it has a scrambled set of the type defined in the Li-Yorke theorem. Marotto's theorem has been further extended by Shiraiwa and Kurata [11].

Chaos in abstract spaces. Auslander and Yorke [1] put forward the following definition, again for discrete dynamical systems.

> If X is a compact metric space and τ a continuous surjection from X to itself then (X, τ) is defined to be a compact system. The point $x \in X$ is said to be stable if for each $\varepsilon > 0$ there is a $\delta > 0$ such that $d(\tau^n(x), \tau^n(y)) < \varepsilon$ for each y with $d(x, y) < \delta$ and each $n \in N$. The compact system (X, τ) is defined to be chaotic if no point $x \in X$ is stable and if there is some $y \in X$ whose orbit is dense in X.

Chaos may also be observed in function spaces. Lasota [6] has proved the existence of chaos, in a sense analogous to that of Auslander and Yorke [1], for the first order partial differential equation

$$\begin{aligned} &\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} = f(u, x), \\ &u(x, 0) = v(x) \end{aligned} \tag{1.5}$$

where $x \in [0, 1]$, $t \geqq 0$, and $c, f,$ and v obey certain specified conditions. Here the phase space is the space of functions $v : [0, 1] \to [0, \infty)$. Lasota's extension of Auslander and Yorke's definition is natural, but, as a result of this extension, the set in which he proves chaos to exist is not compact.

Walther [12], and an der Haiden and Walther [4], examine delay-differential equations. Here the phase space is a function space, $C([0,1], R)$. A Poincaré section is used, and the existence of chaos is proved (in the sense of Li and Yorke) for the first-return map on this section.

In this paper we extend the recent results of Lasota [6] to a higher-dimensional situation. In § 2 we prove a series of preliminary theorems that generalize those of Lasota. In order to produce these results we attempt to extend the conditions that Lasota imposes on scalar valued functions to the vector case. In so doing we also introduce minor modifications in his results in the one-dimensional situation. The main result, confirming a generalized form of chaos for the semiflow defined by (2.1) and (2.2), is proved in § 3. Brunovsky [2] has proved some of Lasota's results by an alternative means. In our final § 4 we generalize this technique to our more general situation. So that this paper can be read independently of Lasota's and so that the differences and extensions we have made can be readily identified it is necessary for us to review the existing results of [6].

Lasota's results. We now review the results of Lasota [6].

If D is some topological space, denote the space of continuous functions from D to the real line by $C(D)$. Denote the space of continuously differentiable functions by $C^1(D)$, and the corresponding spaces of nonnegative functions by $C_+(D)$ and $C^1_+(D)$.

If D is a compact subset of R^n, define the distance between two functions F_1 and F_2 in $C(D)$ to be $\max_{\mathbf{x}\in D} |F_1(\mathbf{x}) - F_2(\mathbf{x})|$.

Consider the partial differential equation (1.5)

$$\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} = f(x, u),$$

$$u(0, x) = v(x),$$

where $t \geqq 0$ and $x \in \Delta = [0, 1]$.

This equation can be used to model a variety of phenomena, in particular the growth of populations of certain types of cells, including red blood cells.

Assume the following:

1) The functions c and f are both C^1;

2) $c(0) = 0$ and $c(x) > 0$ for all $x > 0$;

3) There is some $u_0 \in (0, 1]$ such that

(a) $f(0, u)(u - u_0) < 0$ for all $u > 0$, $u \neq u_0$, and

(b) $f_u(0, u_0) < 0$;

4) There exist $k_1 \geqq 0$, $k_2 \geqq 0$ such that $f(x, u) \leqq k_1 u + k_2$ for all $x \in \Delta$, $u \geqq 0$.

5) $f(x, 0) \geqq 0$ for all $x \in \Delta$, and $f(0, 0) = 0$.

Under these assumptions the following results are proved:

1) For every $v \in C_+(\Delta)$ there is a unique solution to (1.5). Let $\phi(t; t_0, x_0)$ be the unique solution of

$$\frac{dx}{dt} = c(x) \tag{1.6}$$

with $x(t_0) = x_0$; and $\phi(t; x_0, r)$ be the unique solution of

$$\frac{dy}{dt} = f(\phi_{x_0}(t), y) \tag{1.7}$$

with $y(0)=r$, where $\phi_x(t)$ is defined as $\phi(t; 0, x)$. Then the solution $u(t, x)$ of (1.5) can be written as

$$u(t, x)=\psi(t; \phi(0, t, x), v(\phi(0; t, x))). \tag{1.8}$$

2) There is a unique solution $w_0(x)$ to the equation

$$c(x)\frac{du}{dx}=f(x, u) \tag{1.9}$$

for $x \in \Delta$, with $w_0(x)=u_0$. If $v \in C_+(\Delta)$ with $v(0)>0$, then $(S_t v)(x) \to w_0(x)$ as $t \to \infty$, uniformly for $x \in \Delta$.

3) Let

$$V_0=\{v \in C_+(\Delta) \colon v(0)=0\}$$

and

$$V_w=\{v \in C_+(\Delta) \colon v(x)<w_0(x)\ \forall x \in \Delta\}.$$

Define the semidynamical system $\{S_t v\}_{t \geqq 0}$ by

$$(S_t v)(x)=u(t, x) \tag{1.10}$$

where $u(t, x)$ is a solution of (1.5).

Then the sets V_0 and V_w are invariant under S_t; also for each $v \in V_0$ there is a $T_0 \geqq 0$ such that $S_t v \in V_w$ for all $t \geqq T_0$.

If condition 5 is replaced by the stronger condition

5′) $f(x, 0)=0$ for all $x \in \Delta$

(see [2]), then one can show that the semidynamical system $\{S_t\}_{t \geqq 0}$ is chaotic in the set V_w, that is,

(a) no point of V_w is stable;

(b) there is some $v \in V_w$ such that the orbit $\{S_t v \colon t \geqq 0\}$ of v is dense in V_w.

Brunovsky [2] shows that there cannot be chaos in all of V_w if the stronger condition 5′ is not satisfied.

Define a generalized solution of (1.5) to be a function in $C_+(\Delta)$ which is the uniform limit on compact subsets of Δ of solutions of (1.5). We will normally use the word "solution" loosely to refer to a generalized solution.

For consistency we will use notation derived from that of Lasota and Brunovsky where possible.

2. Preliminary theorems. In this section and the next we will generalize the results of Lasota [6], by proving theorems for an initial value problem involving n simultaneous equations in one time variable and m spatial variables.

Lasota deals with the nonlinear initial value problem

$$\begin{gathered}\frac{\partial u}{\partial t}+c(x)\frac{\partial u}{\partial x}=f(x, u),\\ u(t=0, x)=v(x)\end{gathered} \tag{2.1}$$

and the associated semidynamical system

$$(S_t v)(x)=u(t, x). \tag{2.2}$$

Here the results of [6] are extended to the initial value problem for a system of n partial differential of the form

$$\frac{\partial \mathbf{u}}{\partial t}+(\mathbf{c}(\mathbf{x})\cdot\nabla_{\mathbf{x}})\mathbf{u}=\mathbf{f}(\mathbf{x},\mathbf{u}), \qquad \mathbf{u}(t=0,\mathbf{x})=\mathbf{v}(\mathbf{x}) \tag{2.3}$$

where x is now replaced by m independent spatial variables.

Let Δ be a compact simply-connected m-dimensional region in R^m and let Δ have a piecewise C^1 boundary $\partial\Delta$.

Let $\mathbf{c}$ be a function from Δ to R^n satisfying the following conditions:

(C1) $\mathbf{c}$ is a C^1 function of $\mathbf{x}$;

(C2) if the outward unit normal $\hat{\mathbf{n}}(\mathbf{x})$ exists at $\mathbf{x}\in\partial\Delta$ then

$$\hat{\mathbf{n}}(\mathbf{x})\cdot\mathbf{c}(\mathbf{x})>0. \tag{2.4}$$

Thus at all points on Δ the vector field $\mathbf{c}$ points outwards. This is true for the "corners" by continuity of $\mathbf{c}$.

Let $\mathbf{f}$ be a function from $\Delta\times R^m$ to R^n satisfying

(F1) $\mathbf{f}$ is a C^1 function of $\mathbf{x}$ and $\mathbf{u}$.

(F2) We have

$$|\mathbf{f}(\mathbf{x},\mathbf{u})|<k_1|\mathbf{u}|+k_2 \tag{2.5}$$

for all $(\mathbf{x},\mathbf{u})$ in $\Delta\times R^n$ where k_1, k_2 are real numbers, and $|\cdot|$ denotes the usual Euclidean norm.

As in Lasota's original paper the method of characteristics is extensively used. Let $\boldsymbol{\phi}(t;\, t_0,\mathbf{p})$ be the unique solution of

$$\frac{d\mathbf{x}}{dt}=\mathbf{c}(\mathbf{x}), \qquad \mathbf{x}(t_0)=\mathbf{p}. \tag{2.6}$$

Let $\boldsymbol{\phi}_{\mathbf{p}}(t)=\boldsymbol{\phi}(t;\,0,\mathbf{p})$.

Let $\boldsymbol{\psi}(t;\,\mathbf{p},\mathbf{r})$ be the unique solution of

$$\frac{d\mathbf{u}}{dt}=\mathbf{f}(\boldsymbol{\phi}_{\mathbf{p}}(t),\mathbf{u}), \qquad \mathbf{u}(0)=\mathbf{r}. \tag{2.7}$$

Define $T(\mathbf{p})$ to be the value of t at which the curve $\boldsymbol{\phi}_{\mathbf{p}}(t)$ intersects $\partial\Delta$, if this intersection occurs. Let $T(\mathbf{p})=\infty$ if the curve does not intersect the boundary.

Define Δ_0 to be the set $\{p\in\Delta\colon T(\mathbf{p})=\infty\}$.

Define $\partial\Delta_0$ to be $\Delta_0-\operatorname{int}\Delta_0$. Thus, for example, if Δ_0 consists of a single point then $\partial\Delta_0=\Delta_0$.

LEMMA 2.1. *If* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F2), *then the* PDE (2.3) *has a unique solution for all* $\mathbf{x}\in\Delta$, $\mathbf{u}\in R^n$.

Proof. The function $\boldsymbol{\phi}$ is well defined in Δ because Δ is compact and $\mathbf{c}(\mathbf{x})$ is Lipschitz for all $\mathbf{x}$ in Δ. The curves $\boldsymbol{\phi}_{\mathbf{p}}(t)$ are characteristic curves which originate in Δ and cross the boundary $\partial\Delta$ at most once. (The condition $\hat{\mathbf{n}}(\bar{\mathbf{x}})\cdot\mathbf{c}(\bar{\mathbf{x}})>0$ for $\bar{\mathbf{x}}\in\partial\Delta$ ensures that no characteristic curves enter Δ from outside.)

Therefore $\boldsymbol{\phi}_{\mathbf{p}}(t)$ is defined for all $t \leqq T(\mathbf{p})$. Condition (F1) on $\mathbf{f}$ ensures that $\boldsymbol{\psi}(t; \mathbf{p}, \mathbf{r})$ is defined for all t in this range. So letting

$$\text{(2.8)} \qquad \mathbf{u}(t; \mathbf{x}) = \boldsymbol{\psi}(t; \boldsymbol{\phi}(0; t, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t, \mathbf{x})))$$

we have a unique solution to the PDE (2.3).

We will consider, as in Lasota's paper, generalized solutions, which are limits of solutions of (2.3); that is, if $\{\mathbf{u}_n\}_{n=1}^{\infty}$ are solutions of (2.3) and

$$\lim_{n\to\infty} \sup_{\mathbf{x}\in\Delta} |\mathbf{u}_n(\mathbf{x}) - \mathbf{u}(\mathbf{x})| = 0,$$

then $\mathbf{u}$ is said to be a generalized solution of (2.3).

LEMMA 2.2. *T is a continuous function of* $\mathbf{x}$.

Proof. Let $\mathbf{x}_1$ be a point in $\Delta - \Delta_0$. We want to show that for any $\varepsilon > 0$ there is an open set $\mathcal{O}$ in $\Delta - \Delta_0$ containing $\mathbf{x}_1$ such that $|T(\mathbf{x}_1) - T(\mathbf{x})| < \varepsilon$ for all $\mathbf{x}$ in $\mathcal{O}$.

Now $\boldsymbol{\phi}(t_1; t_2, \mathbf{x})$ is a continuous function of $\mathbf{x}$, since the solution of a system of ordinary differential equations is a continuous function of the initial values.

Thus if $t \geqq 0$ then $\phi(0; t, \cdot)$ defines a homeomorphism from $\partial\Delta$ to a subset of $\Delta - \Delta_0$. The sets $\boldsymbol{\phi}(0; T(\mathbf{x}_1) + \varepsilon, \partial\Delta)$ and $\boldsymbol{\phi}(0; T(\mathbf{x}_1) - \varepsilon, \partial\Delta)$ are thus homeomorphic to $\partial\Delta$, and to S^{m-1}. The open region enclosed by these surfaces contains $\mathbf{x}_1$, since the integral curve $\boldsymbol{\phi}(t; 0, \mathbf{x}_1)$ intersects each surface exactly once, and $\mathbf{x}_1$ lies between these intersections. Similarly any point $\mathbf{x}$ in this region satisfies $|T(\mathbf{x}_1) - T(\mathbf{x})| < \varepsilon$. $\mathcal{O}$ can then be chosen to be any open subset of this region that contains $\mathbf{x}_1$.

So T is a continuous function of $\mathbf{x} \in \Delta - \Delta_0$.

LEMMA 2.3. *Δ_0 is nonempty, closed and connected.*

Proof. To show that Δ_0 is not empty, it is sufficient to show that there is some $\mathbf{x}$ in Δ for which $\mathbf{c}(\mathbf{x}) = \mathbf{0}$. For $m = 1$ this follows immediately from the intermediate value theorem. For $m > 1$ Δ can be continuously deformed into B^m, the unit closed ball in m-space, with $\partial\Delta$ becoming S^{m-1}. But any continuous nonvanishing vector field on B^m must contain at least one inward-pointing and at least one outward-pointing normal to S^{m-1}. (See, for example, Dugundji [3].) Here there are no inward-pointing normals, and so it follows that $\mathbf{c}$ vanishes at some point $\mathbf{x}$ in Δ. Note that (2.4) implies that Δ_0 is contained in int Δ.

From every point $\bar{\mathbf{x}}$ of $\partial\Delta$ we can trace back the integral curve by looking at $\boldsymbol{\phi}(0; t, \bar{\mathbf{x}})$ for $t > 0$. We can find in this way points in $\Delta - \Delta_0$ corresponding to any finite nonnegative value of T. Also we have seen that $\boldsymbol{\phi}(0; t, \cdot)$ defines a homeomorphism from $\partial\Delta$ to $T^{-1}(t)$ for any finite $t > 0$.

The complement of Δ_0 is $\bigcup_{n=1}^{\infty} \{\mathbf{x} \in \Delta\colon T(\mathbf{x}) < n\}$. But T is continuous (Lemma 2.2), and $\{\mathbf{x} \in \Delta\colon T(\mathbf{x}) < n\}$ is the inverse image in Δ of an open set in R, and hence is open in Δ. Thus the complement of Δ_0 is a union of sets open in Δ, and so Δ_0 is a closed subset of Δ.

Next suppose that Δ_0 is not connected. Since R^m is normal we can find disjoint open sets U_1 and U_2 such that Δ_0 has nonempty intersection with both U_1 and U_2, and Δ_0 is contained in $U_1 \cup U_2$. Therefore $\Delta - (U_1 \cup U_2)$ is a closed set which does not intersect Δ_0. So there exists

$$T_m = \max_{\mathbf{x}\in\Delta-(U_1\cup U_2)} T(\mathbf{x}) < \infty.$$

Consider the set $T^{-1}(t_0)$ for $t_0 > T_m$. This set $T^{-1}(t_0)$ is contained in $U_1 \cup U_2$ since $t_0 > T_m$. Now since $\boldsymbol{\phi}(0; t_0, \cdot)$ defines a homeomorphism from $\partial\Delta$ to $T^{-1}(t_0)$, it follows that $T^{-1}(t_0)$ is connected. Therefore, $T^{-1}(t_0)$ cannot intersect both U_1 and U_2. Suppose without loss of generality that $T^{-1}(t_0)$ is a subset of U_1, and so $T^{-1}(t_0)$ does not intersect U_2.

Choose ε less than the distance from Δ_0 to $\Delta-(U_1\cup U_2)$. Thus because $\boldsymbol{\phi}(t_0;0,\cdot)$ is continuous we can find $\mathbf{x}_1$ in the intersection of Δ_0 and U_2, and $\mathbf{x}_2\in U_2-\Delta_0$ such that

$$|\boldsymbol{\phi}_{\mathbf{x}_1}(t)-\boldsymbol{\phi}_{\mathbf{x}_2}(t)|<\varepsilon$$

for $0\leqq t\leqq t_0$. Since $\boldsymbol{\phi}_{\mathbf{x}_1}(t_0)\in\Delta_0$ it follows that $\boldsymbol{\phi}_{\mathbf{x}_2}(t_0)\in U_2$. Therefore, $t_0<T(\mathbf{x}_2)<\infty$.

Now consider the curve $\boldsymbol{\phi}_{\mathbf{x}_2}(t)$. We have

$$T(\boldsymbol{\phi}_{\mathbf{x}_2}(0))=T(\mathbf{x}_2)>0,$$

and since $\boldsymbol{\phi}_{\mathbf{x}_2}(T(\mathbf{x}_2))$ lies on $\partial\Delta_0$ we have

$$T(\boldsymbol{\phi}_{\mathbf{x}_2}(T(\mathbf{x}_2)))=0.$$

Therefore there is some point $\mathbf{x}$ on the curve for which $T(\mathbf{x})=t_0$. This point is in U_2, since $t_0>T_m$. But this is a contradiction, since $T^{-1}(t_0)$ does not intersect U_2.

Therefore Δ_0 must be connected.

LEMMA 2.4. *If* $\mathbf{x}_0\in\partial\Delta_0$, *then* $\boldsymbol{\phi}_{\mathbf{x}_0}(t)\in\partial\Delta_0$ *for all* $t\geqq 0$.

Proof. First, suppose that $\boldsymbol{\phi}_{\mathbf{x}_0}(t)\in\Delta-\Delta_0$ for some $t>0$. Then $T(\boldsymbol{\phi}_{\mathbf{x}_0}(t))<\infty$, say $T(\boldsymbol{\phi}_{\mathbf{x}_0}(t))=T_0$. But this implies that $\boldsymbol{\phi}_{\mathbf{x}_0}(t+T_0)\in\Delta_0$, and so $T(\mathbf{x}_0)<\infty$, which contradicts the assumption that $\mathbf{x}_0\in\Delta_0$.

Suppose $\mathbf{x}_0\in\partial\Delta_0$ with $\boldsymbol{\phi}_{\mathbf{x}_0}(t_0)\in\operatorname{int}\Delta_0$ for some $t_0>0$. Choose ε with $0<\varepsilon<d(\boldsymbol{\phi}_{\mathbf{x}_0}(t_0),\partial\Delta_0)$. Let $\delta>0$ be such that for all $\mathbf{x}\in\Delta$ with $|\mathbf{x}-\mathbf{x}_0|<\delta$ we have $|\boldsymbol{\phi}_{\mathbf{x}}(t)-\boldsymbol{\phi}_{\mathbf{x}_0}(t)|<\varepsilon$ for all $t<t_0$. But since $\mathbf{x}_0\in\partial\Delta_0$ we can find $\mathbf{x}\in\Delta-\Delta_0$ with $|\mathbf{x}-\mathbf{x}_0|<\delta$. Thus $|\boldsymbol{\phi}_{\mathbf{x}}(t_0)-\boldsymbol{\phi}_{\mathbf{x}_0}(t_0)|<\varepsilon$, which implies that $\boldsymbol{\phi}_{\mathbf{x}}(t_0)\in\operatorname{int}\Delta_0$. But this is impossible since $T(\mathbf{x})<\infty$.

Therefore $\boldsymbol{\phi}_{\mathbf{x}_0}(t)\in\partial\Delta_0$.

In [6] the function $f(x,u)$ is required to satisfy

$$(u-u_0)f(0,u)<0 \quad \text{for } u>0, u\neq u_0,$$

$$f_u(0,u_0)<0 \quad \text{for some } u_0>0.$$

Thus u_0 is an attracting stationary point for the trajectory $u(0,t)$.

Here this is generalized to an invariant surface of dimension $n-1$, where $n>1$, or two points if $n=1$. This surface is to be attracting in the sense that at least locally $\mathbf{f}(\mathbf{x},\mathbf{u})$ points towards this surface, for each $\mathbf{x}$ in Δ_0. Here, however the point $0\in R$ has been generalized to the set Δ_0. Thus the question arises as to whether the invariant surface can depend on the element $\mathbf{x}$ of Δ_0, or whether it must be independent of $\mathbf{x}$. We find results for the more general case, that is, where the surface depends on $\mathbf{x}\in\Delta_0$.

We require $\mathbf{f}$ to satisfy the following further conditions:

(F3) There is a compact simply-connected n-dimensional subset $W(\mathbf{x})$ of R^n associated (continuously) with each point $\mathbf{x}$ in Δ_0 with $\partial W(\mathbf{x})$ piecewise C^1.

(F4) For each $\mathbf{x}$ in Δ_0 there is an open set $R(\mathbf{x})$ in R^n, such that $\partial W(\mathbf{x})$ is contained in $R(\mathbf{x})$. If $\mathbf{x}\in\Delta_0$, $\mathbf{u}\in R(\mathbf{x})$ and $\mathbf{u}_1\in\partial W(\mathbf{x})$ is the nearest point (in the Euclidean sense) in $\partial W(\mathbf{x})$ to $\mathbf{u}$ (if this is well defined) then

$$(\mathbf{u}-\mathbf{u}_1)\cdot\mathbf{f}(\mathbf{x},\mathbf{u})<0. \tag{2.9}$$

(F5) If $\mathbf{x}_0\in\partial\Delta_0$, $\mathbf{u}_1\in\partial W(\mathbf{x}_0)$, then

$$\mathbf{N}\cdot(Jf(\mathbf{x},\mathbf{u}_1)\mathbf{N})<0 \tag{2.10}$$

where $\mathbf{N}$ is normal to $\partial W(\mathbf{x})$ at $\mathbf{u}_1$ and where Jf denotes the Jacobian of f.

Note that $\mathbf{u}_1(\mathbf{x},\mathbf{u})$ is well defined for almost all $\mathbf{u}\in R^n$. If $\mathbf{u}_1$ is not uniquely defined then $\mathbf{f}(\mathbf{x},\mathbf{u})$ may still satisfy (2.9) for each of the several nearest $\mathbf{u}_1$ to $\mathbf{u}$. However, by an argument similar to that of Lemma 2.3 we see that the vector fields $\mathbf{u}-\mathbf{u}_1$ and $\mathbf{f}$ on

$W(\mathbf{x})$ cannot be well defined, continuous and nonzero everywhere on int $W(\mathbf{x})$. So there is some point $\mathbf{u} \in \operatorname{int} W(\mathbf{x})$ where $\mathbf{u}-\mathbf{u}_1$ is undefined because $\mathbf{u}_1$ is undefined, and at this point it is not possible to define $\mathbf{f}(\mathbf{x}, \mathbf{u})$ such that (2.9) holds for each of the nearest points of $\partial W(\mathbf{x})$. There may be more than one such point, for if the distance $|\mathbf{u}-\mathbf{u}_1|$ has a local maximum at $\mathbf{u}$, then $\mathbf{f}$ is necessarily zero at this point.

Let

$$m = \min_{\mathbf{x}_0 \in \Delta_0} d(\partial R(\mathbf{x}_0), \partial W(\mathbf{x}_0)),$$

and for each $\mathbf{x}_0 \in \Delta_0$ let

$$R'(\mathbf{x}_0) = \{\mathbf{u} \in R^n \colon d(\mathbf{u}, \partial W(\mathbf{x}_0)) \leqq m/2\}.$$

Let $R_0(\mathbf{x}_0)$ be the open subset of R^n defined by

$$R_0(\mathbf{x}_0) = \{\mathbf{u} \in R^n \colon \boldsymbol{\psi}(t; \mathbf{x}_0, \mathbf{u}) \in R'(\boldsymbol{\phi}_{\mathbf{x}_0}(t)) \text{ for some } t \geqq 0\}.$$

$R_0(\mathbf{x}_0)$ is nonempty, as it contains at least $R'(\mathbf{x}_0)$. Also, by an argument similar to that of Lemma 2.3, we see that there is at least one point in int $W(\mathbf{x}_0)$ that is not in $R_0(\mathbf{x}_0)$.

Let $S_0(\mathbf{x}_0) = W(\mathbf{x}_0) - R_0(\mathbf{x}_0)$. Clearly, if $\mathbf{v}(\mathbf{x}_0) \in S_0(\mathbf{x}_0)$ for all $\mathbf{x}_0 \in \Delta_0$, then $(S_t\mathbf{v})(\mathbf{x}_0) \in S_0(\mathbf{x}_0)$ for all $\mathbf{x}_0 \in \Delta_0$ and all $t \geqq 0$.

Let θ be a continuous function from Δ to R with

$$\theta(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \Delta_0,$$
$$\theta(\mathbf{x}) \geqq d(\mathbf{x}, \Delta_0) \quad \forall \mathbf{x} \in \Delta - \Delta_0,$$

such that θ has no stationary points in $\Delta - \operatorname{int} \Delta_0$. Thus if $0 < h \leqq \min_{\mathbf{x} \in \partial\Delta} d(\mathbf{x}, \Delta_0)$ then $\theta^{-1}(h)$ is homeomorphic to S^{m-1}. For each $h > 0$ let

$$A_h = \{\mathbf{x} \in \Delta \colon \theta(\mathbf{x}) \leqq h\}.$$

Then A_h is a closed subset of Δ, containing Δ_0 in its interior with $d(\mathbf{x}, \Delta_0) \leqq h$ for all $\mathbf{x}$ in A_h, and A_{h_1} a subset of A_{h_2} if $0 < h_1 \leqq h_2$. Now let $t_h = \max_{\mathbf{x} \in \partial A_h} T(\mathbf{x})$, and let $\Delta_h = \{\mathbf{x} \in \Delta \colon T(\mathbf{x}) \geqq t_h\}$. Clearly Δ_h is a subset of A_h, containing Δ_0 in its interior. If $0 < h_1 \leqq h_2$, then $t_{h_1} \geqq t_{h_2}$ and Δ_{h_1} is a subset of Δ_{h_2}. Also if $\mathbf{x} \in \Delta_h$ for some $h > 0$ then the characteristic curve $\boldsymbol{\phi}_{\mathbf{x}}(t)$ intersects $\partial\Delta_h$ exactly once.

LEMMA 2.5. *If* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5), *then there is a unique* C^1 *function* $\mathbf{w}_h : \Delta - \operatorname{int} \Delta_h \to R^n$ *satisfying*

$$(2.11) \qquad \begin{aligned} (\mathbf{c}(\mathbf{x}) \cdot \nabla)\mathbf{w}_h &= \mathbf{f}(\mathbf{x}, \mathbf{w}_h) \quad \forall \mathbf{x} \in \Delta - \operatorname{int} \Delta_h, \\ \mathbf{w}_h(\mathbf{x}) &= \mathbf{v}_h(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Delta_h. \end{aligned}$$

Proof. The intersection of $\boldsymbol{\phi}_{\mathbf{x}}$ with $\partial\Delta_h$ is unique for all $\mathbf{x} \in \Delta - \operatorname{int} \Delta_h$. Suppose this intersection takes place at $t = \tau(\mathbf{x}, h) \leqq 0$, and let $\mathbf{x}_h = \boldsymbol{\phi}_{\mathbf{x}}(\tau(\mathbf{x}, h))$. Let $\boldsymbol{\psi}$ satisfy

$$(2.12) \qquad \begin{aligned} &\frac{d\boldsymbol{\psi}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \\ &\boldsymbol{\psi}(\tau(\mathbf{x}, h)) = \mathbf{v}_h(\mathbf{x}_h). \end{aligned}$$

Now $\boldsymbol{\psi}$ is unique for $\tau(\mathbf{x}, h) \leqq t \leqq T(\mathbf{x})$. So $\mathbf{w}_h(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{0})$, giving a unique solution to (2.11).

THEOREM 2.1. *Suppose* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5). *Then we can define* $W(\mathbf{x})$ *for all* $\mathbf{x}$ *in* $\Delta - \text{int}\,\Delta_0$; *that is, to each* $\mathbf{x} \in \Delta - \text{int}\,\Delta_0$ *there corresponds a unique compact simply connected subset* $W(\mathbf{x})$ *of* R^n (*where for* $\mathbf{x}_0 \in \partial\Delta_0$, $W(\mathbf{x}_0)$ *is as previously defined*), *such that if* $\mathbf{w}(\mathbf{x})$ *satisfies*

$$(\mathbf{c}(\mathbf{x}) \cdot \nabla)\mathbf{w} = \mathbf{f}(\mathbf{x}, \mathbf{w}) \tag{2.13}$$

and $\mathbf{w}(\mathbf{x}) \in \partial W(\mathbf{x})$ *for some* $\mathbf{x}$ *in* $\Delta - \Delta_0$, *then* $\mathbf{w}(\mathbf{x}) \in \partial W(\mathbf{x})$ *for all* $\mathbf{x}$ *in* $\Delta - \Delta_0$.

Proof. Let $\mathbf{x} \in \Delta - \Delta_0$. If h is sufficiently small, so that $\mathbf{x} \in \Delta - \Delta_h$, then we can consider the set of points $\{\mathbf{w}_h(\mathbf{x})\}$ arising from different values of $\mathbf{v}_h$ at the point $\mathbf{x}_h$ where the characteristic curve through $\mathbf{x}$ crosses $\partial\Delta_h$. Let $\mathbf{x}_0$ be a point on $\partial\Delta_0$ within a distance h of $\mathbf{x}_h$. Let $\mathbf{v}_h(\mathbf{x}_h)$ range over $\partial W(\mathbf{x}_0)$. Uniqueness of solutions (Lemma 2.5) ensures that the resulting set of points $\{\mathbf{w}_h(\mathbf{x})\}$ is homeomorphic to $\partial W(\mathbf{x}_0)$, that is, homeomorphic to S^{n-1}. Denote by $W_h(\mathbf{x})$ the region bounded by this set, and the set itself by $\partial W(\mathbf{x})$.

We wish to show that as $h \to 0$, $W_h(\mathbf{x})$ approaches a limit.

Let $\mathbf{u}_1 \in \partial W(\mathbf{x}_0)$ and let $\mathbf{N}$ be normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$ with $\hat{\mathbf{N}}$ the corresponding unit normal.

Then (2.9) implies

$$\hat{\mathbf{N}} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}_1 + \alpha\hat{\mathbf{N}}) < 0 \tag{2.14}$$

for $\alpha > 0$ sufficiently small and for all $\mathbf{x}$ in $\partial\Delta_0$. Inequality (2.10) gives

$$\hat{\mathbf{N}} \cdot (Jf(\mathbf{x}, \mathbf{u}_1)\hat{\mathbf{N}}) < 0 \quad \forall \mathbf{x} \text{ in } \partial\Delta_0. \tag{2.15}$$

Choose $h > 0$, $k > 0$ such that if $\mathbf{u}_1 \in \partial W(\mathbf{x}_0)$ and $\mathbf{N}$ is normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$, then the following conditions hold:

(a) If $\mathbf{x} \in \Delta_h$, $|\mathbf{N}| = k$, then

$$\hat{\mathbf{N}} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}_1 + \mathbf{N}) < 0; \tag{2.16}$$

(b) If $\mathbf{x} \in \Delta_h$, $|\mathbf{N}| < k$, then

$$\hat{\mathbf{N}} \cdot (Jf(\mathbf{x}, \mathbf{u}_1 + \mathbf{N})\hat{\mathbf{N}}) < 0; \tag{2.17}$$

(c) If $|\mathbf{N}| \leqq k$ then $\mathbf{u}_1$ is the nearest point on $\partial W(\mathbf{x}_0)$ to $\mathbf{u}_1 + \mathbf{N}$.

Denote by W_h the set of points $\{w_h(\mathbf{x})\colon \mathbf{x} \in \Delta - \text{int}\,\Delta_0\}$, and consider W_{h_1} and W_{h_2} where $h_1 < h_2 < h$. We want to show that W_h approaches a limit as $h \to 0$, and in order to do this we use Cauchy convergence.

Let $\mathbf{x} \in \Delta_h - \Delta_{h_2}$. If $\mathbf{x}_1 \in \partial\Delta_{h_1}$ and $\mathbf{x}_2 \in \partial\Delta_{h_2}$, we have $\mathbf{w}_{h_1}(\mathbf{x}_{h_1}) \in \partial W(\mathbf{x}_0)$ and $\mathbf{w}_{h_1}(\mathbf{x}_{h_1}) \in \partial W(\mathbf{x}_0)$. So $d(\mathbf{w}_{h_i}, W(\mathbf{x}_0)) < k$ because $\mathbf{N} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}_1 + \mathbf{N}) < 0$ for all $\mathbf{u}_1 \in \partial W(\mathbf{x}_0)$, $\mathbf{x} \in \Delta_h$, $|\mathbf{N}| = k$.

Let $z(\mathbf{x}, \mathbf{u}_1)$ be the distance along the normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$ from the intersection with $W_{h_1}(\mathbf{x})$ to the intersection with $W_{h_2}(\mathbf{x})$. (If there is more than one possible value of $z(\mathbf{x}, \mathbf{u}_1)$ due to multiple intersections, take the largest such value.)

So $z(\mathbf{x}, \mathbf{u}_1) < 2k$ if $\mathbf{x} \in \Delta_h$. Let $\mathbf{w}_{h_1}(\mathbf{x}, \mathbf{u}_1)$ and $\mathbf{w}_{h_2}(\mathbf{x}, \mathbf{u}_1)$ be these points of intersection and so

$$\mathbf{z}(\mathbf{x}, \mathbf{u}_1) = \mathbf{w}_{h_1}(\mathbf{x}, \mathbf{u}_1) - \mathbf{w}_{h_2}(\mathbf{x}, \mathbf{u}_1).$$

Therefore $z(\mathbf{x}, \mathbf{u}_1) = \hat{\mathbf{N}} \cdot \mathbf{z}(\mathbf{x}, \mathbf{u}_1)$, where $\hat{\mathbf{N}}$ is the unit normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$ in the direction of $\mathbf{z}(\mathbf{x}, \mathbf{u}_1)$.

Then if s represents the arc-length of the characteristic $\phi_{\mathbf{x}_h}(t)$ we have

$$\begin{aligned}\frac{d}{ds}z(\mathbf{x},\mathbf{u}_1)&=\frac{d}{dt}z(\mathbf{x},\mathbf{u}_1)\Big/\frac{ds}{dt}\\&=\hat{\mathbf{N}}\cdot(\mathbf{f}(\mathbf{x},\mathbf{w}_{h_1})-\mathbf{f}(\mathbf{x},\mathbf{w}_{h_2}))/(|\mathbf{c}(\mathbf{x})|)\\&=\hat{\mathbf{N}}\cdot(Jf(\mathbf{x},\mathbf{w}_{s_2}+\gamma\mathbf{z})\mathbf{z})/(|\mathbf{c}(\mathbf{x})|),\end{aligned}$$

where γ is in $(0,1)$.

So $(d/ds)z(\mathbf{x},\mathbf{u}_1)<0$ because $\mathbf{z}$ is a positive multiple of $\hat{\mathbf{N}}$ and (2.17) holds. So $z(\mathbf{x},\mathbf{u}_1)$ decreases with s for all $\mathbf{u}_1$ in $\partial W(\mathbf{x}_0)$, $\mathbf{x}$ in Δ_h.

Existence of $\lim_{h\to 0} W_h$. Choose $\varepsilon>0$ $(\varepsilon<2k)$. Now there is some $h_0\in(0,h)$ such that

$$\mathbf{N}\cdot\mathbf{f}(\mathbf{x},\mathbf{u}_1+\mathbf{N})<0 \tag{2.18}$$

for $\mathbf{N}=\varepsilon/2$, $\mathbf{x}\in\Delta_{h_0}$ (using (2.7) and (2.15)).

Suppose $h_1, h_2<h_0$. Then $d(\mathbf{w}_{h_i}, W(\mathbf{x}_0))<\varepsilon/2$ for $\mathbf{x}\in\Delta_{h_0}$. Hence if $\mathbf{w}_{h_1}$, $\mathbf{w}_{h_2}$ are on the normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1\in\partial \mathrm{W}(\mathbf{x}_0)$ and

$$\mathbf{z}(\mathbf{x},\mathbf{u}_1)=\mathbf{w}_{h_1}(\mathbf{x},\mathbf{u}_1)-\mathbf{w}_{h_2}(\mathbf{x},\mathbf{u}_1)$$

then $z(\mathbf{x},\mathbf{u}_1)<\varepsilon$ for $\mathbf{x}\in\Delta_{h_0}$. We have just shown that $z(\mathbf{x},\mathbf{u}_1)$ decreases with s if $\mathbf{x}\in\Delta_h$. So $z(\mathbf{x},\mathbf{u}_1)\leqq z(\mathbf{x}_{h_0},\mathbf{u}_1)<\varepsilon$ if $\mathbf{x}_{h_0}\in\Delta_{h_0}$, $\mathbf{x}\in\Delta_h-\Delta_{h_0}$. Therefore $\{W_h(\mathbf{x})\}$ satisfies a Cauchy condition for $\mathbf{x}\in\Delta_h-\Delta_0$. Hence if $\mathbf{x}\in\Delta_h-\Delta_0$ there is a limit subset $W_0(\mathbf{x})$ in R^n satisfying $W_0(\mathbf{x})=\lim_{h\to 0}W_h(\mathbf{x})$. Clearly $W_0(\mathbf{x}_0)=W(\mathbf{x}_0)$ for all $\mathbf{x}_0\in\partial\Delta_0$ for continuity.

By Lemma 2.5 it can be seen that $W_0(\mathbf{x})$ is well defined for all $\mathbf{x}$ in $\Delta-\operatorname{int}\Delta_0$ since elements of $\partial W_0(\mathbf{x})$ clearly lie on integral curves of (2.11) and these curves can be extended until $\partial\Delta$ is reached.

Uniqueness. Suppose there are two limit subsets $W_{01}(\mathbf{x})$ and $W_{02}(x)$ of $W_h(\mathbf{x})$ satisfying $\lim_{h\to 0}W_{0i}(\mathbf{x}_h)=W(\mathbf{x}_0)$. Let $z(\mathbf{x},\mathbf{u}_1)$ be the distance between $W_{01}(\mathbf{x})$ and $W_{02}(\mathbf{x})$ along the normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$. But $z(\mathbf{x},\mathbf{u}_1)$ decreases with s, since both $\mathbf{w}_{01}(\mathbf{x})$ and $\mathbf{w}_{02}(\mathbf{x})$ are generalized solutions of (2.11). $W_{0i}(\mathbf{x})$ is fixed for $\mathbf{x}\in\partial\Delta_0$, giving $\lim_{h\to 0}z(\mathbf{x}_h,\mathbf{u}_1)=0$ for all $\mathbf{u}_1\in W(\mathbf{x}_0)$. So $z(\mathbf{x},\mathbf{u}_1)=0$ for all $\mathbf{x}\in\Delta-\operatorname{int}\Delta_0$, $\mathbf{u}_1\in\partial\mathrm{W}(\mathbf{x}_0)$. Hence $W_0(\mathbf{x})$ is well defined for all $\mathbf{x}\in\Delta-\operatorname{int}\Delta_0$.

THEOREM 2.2. *Suppose* (C1)-(C2) *and* (F1)-(F5) *hold for* $\mathbf{c}$ *and* $\mathbf{f}$ *and that* $\mathbf{v}(\mathbf{x}_0)$ *lies in* $R_0(\mathbf{x}_0)$ *for all* $\mathbf{x}_0$ *in* Δ_0. *Then the solution* $\mathbf{u}(t,\mathbf{x})$ *of* (2.3) *has the property that as* $t\to\infty$ *the distance between the point* $\mathbf{u}(t,\mathbf{x})$ *and the surface* $\partial W(\mathbf{x})$ *tends to zero uniformly for* $\mathbf{x}$ *in* Δ.

First we give three lemmas which will enable us to prove this theorem.

Let $\mathbf{u}(t,\mathbf{x})$ be the solution to

$$\frac{\partial\mathbf{u}}{\partial t}+(\mathbf{c}\cdot\nabla)\mathbf{u}=\mathbf{f}(\mathbf{x},\mathbf{u})$$

with

$$\mathbf{u}(t=0,\mathbf{x})=\mathbf{v}(\mathbf{x})$$

where $\mathbf{c}$ and $\mathbf{f}$ satisfy (C1)-(C2) and (F1)-(F5). Define $w(\mathbf{u}(t,\mathbf{x}))\equiv w(t,\mathbf{x})$ to be the distance from $\mathbf{u}(t,\mathbf{x})$ to the nearest point on $\partial W(\mathbf{x})$. The function $w(\mathbf{u})$ is defined and continuous everywhere. If this nearest point is unique denote it by $\mathbf{u}_1(t,\mathbf{x})$. Now if $\mathbf{u}$ and $\mathbf{u}_1$ are distinct, then $\mathbf{u}-\mathbf{u}_1$ is normal to $\partial W(\mathbf{x})$. Denote $\mathbf{u}-\mathbf{u}_1$ by N, and the corresponding unit vector by $\hat{\mathrm{N}}$.

Let

$$\bar{u}(t,\mathbf{x})=[(w(t,\mathbf{x}))^2+1]^{-1}. \tag{2.19}$$

Note that $0<\bar{u}\leqq 1$. Suppose

$$\frac{\partial \bar{u}}{\partial t}+(\mathbf{c}\cdot\nabla)\bar{u}=\bar{f}(t,\mathbf{x},\mathbf{u}) \tag{2.20}$$

and define

$$f_0(t,\mathbf{x},\mathbf{u})=\frac{\bar{f}(t,x,u)}{\bar{u}-1}. \tag{2.21}$$

LEMMA 2.6. *$\bar{f}$ and f_0 are independent of t.*

Proof. Now

$$\frac{\partial \mathbf{u}}{\partial t}+(\mathbf{c}\cdot\nabla)\mathbf{u}=\mathbf{f}(\mathbf{x},\mathbf{u}),$$

and so

$$\frac{\partial \bar{u}}{\partial t}+(\mathbf{c}\cdot\nabla)\bar{u}=-\frac{2(\mathbf{u}-\mathbf{u}_1)\cdot(\mathbf{f}(\mathbf{x},\mathbf{u})-\partial\mathbf{u}_1/\partial t-(\mathbf{c},\nabla)\mathbf{u}_1)}{(w^2+1)^2}.$$

Now $\mathbf{u}-\mathbf{u}_1$ is normal to the surface $\partial W(\mathbf{x})$, and since $\mathbf{u}_1$ remains on $\partial W(\mathbf{x})$ we have $\partial\mathbf{u}_1/\partial t$ tangential to $\partial W(\mathbf{x})$, giving

$$(\mathbf{u}-\mathbf{u}_1)\cdot\frac{\partial \mathbf{u}_1}{\partial t}=0.$$

Theorem 2.1 states that if $\mathbf{w}(\mathbf{x})$ satisfies $(\mathbf{c}\cdot\nabla)\mathbf{w}=\mathbf{f}(\mathbf{x},\mathbf{w})$, and $\mathbf{w}(\mathbf{x})\in\partial W(\mathbf{x})$ for some $\mathbf{x}\in\Delta-\Delta_0$, then $\mathbf{w}(\mathbf{x})\in W(\mathbf{x})$ for all $\mathbf{x}\in\Delta-\Delta_0$. Suppose that for some $\mathbf{x}\in\Delta-\Delta_0$ and some $t\geqq 0$, $\mathbf{w}(\mathbf{x})=\mathbf{u}_1(t,\mathbf{x})\in\partial W(\mathbf{x})$.

Consider $(\mathbf{c}\cdot\nabla)\mathbf{u}_1$. The function $\mathbf{u}_1(\mathbf{x},\cdot)$ takes values on $\partial W(\mathbf{x})$. Then $(\mathbf{c}\cdot\nabla)(\mathbf{u}_1-\mathbf{w})$ is tangential to $\partial W(\mathbf{x})$, since both $\mathbf{u}_1(\mathrm{t},\mathbf{x})$ and $\mathbf{w}(\mathbf{x})$ remain on $\partial W(\mathbf{x})$. Therefore the components of $(\mathbf{c}\cdot\nabla)\mathbf{u}_1$ and $(\mathbf{c},\nabla)\mathbf{w}$ normal to $\partial W(\mathbf{x})$ must be equal. Therefore

$$(\mathbf{u}-\mathbf{u}_1)\cdot((\mathbf{c}\cdot\nabla)\mathbf{u}_1)=(\mathbf{u}-\mathbf{u}_1)\cdot\mathbf{f}(\mathbf{x},\mathbf{u}_1).$$

In other words $\hat{\mathbf{N}}\cdot\mathbf{f}(\mathbf{x},\mathbf{u}_1)$ is the component of $(\mathbf{c}\cdot\nabla)\mathbf{u}_1$ normal to $\partial W(\mathbf{x})$ at $\mathbf{u}_1$, representing the change in $\mathbf{u}_1$ necessary to keep $\mathbf{u}_1$ on $\partial W(\mathbf{x})$ as $\mathbf{x}$ changes.

Also, if $\mathbf{x}_0\in\Delta_0$, $\mathbf{u}_1\in\partial W(\mathbf{x})$, then $\psi(t;\mathbf{x}_0,\mathbf{u}_1)$ lies on $\partial W(\mathbf{f}_{\mathbf{x}_0}(t))$ for all $t\geqq 0$. So if $\mathbf{u}(t,\mathbf{x})$ satisfies

$$\frac{\partial \mathbf{u}}{\partial t}+(\mathbf{c}\cdot\nabla)\mathbf{u}=\mathbf{f}(\mathbf{x},\mathbf{u}),$$

$$\mathbf{u}(t=0,\mathbf{x})=\mathbf{v}(\mathbf{x}),$$

and $\mathbf{v}(\mathbf{x}_0)\in\partial W(\mathbf{x}_0)$ for each $\mathbf{x}_0\in\Delta_0$, then $\mathbf{u}(t,\mathbf{x}_0)\in\partial W(\mathbf{x}_0)$ for all $\mathbf{x}_0\in\Delta_0$, $t\geqq 0$. Suppose that $\mathbf{v}(\mathbf{x}_0)=\mathbf{u}_1$. Then $\mathbf{N}\cdot(\partial\mathbf{u}/\partial t)=0$ if $\mathbf{N}$ is normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$. So, as before, we have

$$(\mathbf{u}-\mathbf{u}_1)\cdot((\mathbf{c}\cdot\nabla)\mathbf{u}_1)=(\mathbf{u}-\mathbf{u}_1)\cdot\mathbf{f}(\mathbf{x},\mathbf{u}_1).$$

So if $\mathbf{x}$ is in Δ, $\mathbf{u}\in R^n$, then

$$\bar{f}(t,\mathbf{x},\mathbf{u})=\frac{2[(\mathbf{u}-\mathbf{u}_1)\cdot(\mathbf{f}(\mathbf{x},\mathbf{u})-\mathbf{f}(\mathbf{x},\mathbf{u}_1))]}{(w^2+1)^2},$$

which is independent of t for $\mathbf{x}\in\Delta$, $\mathbf{u}\in R^n$. Hence f_0 is also independent of t.

LEMMA 2.7. *If* $\mathbf{x}_0 \in \partial\Delta_0$ *then* $f_0(\mathbf{x}_0, \mathbf{u}_0) < 0$ *for all* $\mathbf{u} \in R_0(\mathbf{x}_0)$.

Proof. Let $\mathbf{x}_0 \in \partial\Delta_0$. Now

$$\frac{1}{(\bar{u}-1)} = \frac{1}{(w^2+1)^{-1}-1} = -\frac{w^2+1}{w^2}.$$

Therefore, if $\mathbf{u}$ is not on $\partial W(\mathbf{x}_0)$

$$\begin{aligned} f_0(\mathbf{x}_0, \mathbf{u}) &= \frac{2[(\mathbf{u}-\mathbf{u}_1)\cdot(\mathbf{f}(\mathbf{x}_0,\mathbf{u})-\mathbf{f}(\mathbf{x}_0,\mathbf{u}_1))]}{w^2(w^2+1)} \\ &= \frac{2[\mathbf{N}\cdot(\mathbf{f}(\mathbf{x}_0,\mathbf{u}_1+\mathbf{N})-\mathbf{f}(\mathbf{x}_0,\mathbf{u}_1))]}{|\mathbf{N}|^2(|\mathbf{N}|^2+1)}. \end{aligned} \tag{2.22}$$

But, since $\mathbf{N}$ is normal to $\partial W(\mathbf{x}_0)$ at $\mathbf{u}_1$, we have $\mathbf{N}\cdot\mathbf{f}(\mathbf{x}_0,\mathbf{u}_1)=0$, and $\mathbf{N}\cdot\mathbf{f}(\mathbf{x}_0,\mathbf{u})<0$, giving

$$f_0(\mathbf{x}_0, \mathbf{u}) < 0. \tag{2.23}$$

For $\mathbf{u}$ close to $\partial W(\mathbf{x}_0)$, we have

$$\begin{aligned} f_0(\mathbf{x}_0, \mathbf{u}) &= \frac{2[\mathbf{N}\cdot((Jf(\mathbf{x}_0,\mathbf{u}_1)\mathbf{N})+O(|\mathbf{N}|)^2)]}{|\mathbf{N}|^2(|\mathbf{N}|^2+1)} \\ &= \frac{[\hat{\mathbf{N}}\cdot(Jf(\mathbf{x}_0,\mathbf{u}_1)\hat{\mathbf{N}})+O(|\mathbf{N}|)]}{|\mathbf{N}|^2+1}. \end{aligned}$$

But condition (F5) implies that $\hat{\mathbf{N}}\cdot(Jf(\mathbf{x}_0,\mathbf{u}_1)\hat{\mathbf{N}})<0$. Also, $f_0(\mathbf{x}_0,\mathbf{u})$ is bounded as $\mathbf{u}$ approaches $\partial W(\mathbf{x}_0)$. So $f_0(\mathbf{x}_0,\mathbf{u})$ is continuous and negative for $\mathbf{u}$ in $R_0(\mathbf{x}_0)$.

LEMMA 2.8. $|f_0(\mathbf{x},\mathbf{u})|$ *is bounded for* $\mathbf{x}\in\Delta-\operatorname{int}\Delta_0$ *and* $\mathbf{u}\in R^n$.

Proof. For any $\mathbf{x}\in\Delta-\operatorname{int}\Delta_0$, $\mathbf{u}$ close to $\partial W(\mathbf{x})$, $\mathbf{N}=\mathbf{u}_1(\mathbf{u})-\mathbf{u}$, we have, similarly to before,

$$f_0(\mathbf{x},\mathbf{u}) = \frac{[\hat{\mathbf{N}}\cdot(Jf(\mathbf{x},\mathbf{u}_1)\hat{\mathbf{N}})+O(|\mathbf{N}|)]}{|\mathbf{N}|^2+1}, \tag{2.24}$$

which is clearly bounded as $|\mathbf{N}|\to 0$.

Also,

$$f_0(\mathbf{x},\mathbf{u}) = 2|\mathbf{N}|^{-2}(|\mathbf{N}|^2+1)^{-1}[\mathbf{N}\cdot(\mathbf{f}(\mathbf{x},\mathbf{u})-\mathbf{f}(\mathbf{x},\mathbf{u}_1))].$$

We know that $|\mathbf{f}(\mathbf{x},\mathbf{u})|<k_1|\mathbf{u}|+k_2$. Let

$$F = \max_{\substack{\mathbf{x}\in\Delta-\operatorname{int}\Delta_0 \\ \mathbf{u}\in\partial W(\mathbf{x})}} |\mathbf{f}(\mathbf{x},\mathbf{u})|. \tag{2.25}$$

Then

$$|f_0(\mathbf{x},\mathbf{u})| < 2|\mathbf{N}|^{-1}(|\mathbf{N}|^2+1)^{-1}[k_1|\mathbf{u}|+k_2+F].$$

Now assume that

$$U = \max_{\substack{\mathbf{x}\in\Delta-\operatorname{int}\Delta_0 \\ \mathbf{u}_1\in\partial W(\mathbf{x})}} |\mathbf{u}_1|. \tag{2.26}$$

If $|\mathbf{u}|>2U$ then $|\mathbf{u}-\mathbf{u}_1|>|\mathbf{u}|/2$. Therefore

$$\begin{aligned} |f_0(\mathbf{x},\mathbf{u})| &< 4|\mathbf{u}|^{-1}(\tfrac{1}{2}|\mathbf{u}|^2+1)^{-1}[k_1|\mathbf{u}|+k_2+F] \\ &< 8|\mathbf{u}|^{-2}[k_1+|\mathbf{u}|^{-1}(k_2+F)], \end{aligned} \tag{2.27}$$

which approaches zero as $|\mathbf{u}|\to\infty$.

So $|f_0(\mathbf{x},\mathbf{u})|$ is bounded in the compact region $\mathbf{x}\in\Delta-\operatorname{int}\Delta_0$, $|\mathbf{u}|<2U$ and $f_0\to 0$ as $|\mathbf{u}|\to\infty$. This implies that f_0 is bounded everywhere.

Proof of Theorem 2.2. First, show that for each $\varepsilon > 0$ there exists $\delta > 0$ such that for any $t_1 > 0$, $\mathbf{p} \in \Delta$, $\mathbf{r} \in R^n$ if

$$1 - \frac{1}{1 + |\boldsymbol{\psi}(t_1; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2} \leqq \delta$$

then

$$1 - \frac{1}{1 + |\boldsymbol{\psi}(t; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2} \leqq \varepsilon$$

for $t_1 \leqq t \leqq T(\mathbf{p})$.

If $\mathbf{p} \in \Delta - \Delta_0$ and $s(t) \geqq 0$ is the arc-length along the characteristic $\boldsymbol{\phi}_{\mathbf{p}}(t)$ from $\mathbf{p}$, then we can define t as $t \equiv t(s)$ because we know that s is strictly increasing with t. Define

$$z(s) = [1 + |\boldsymbol{\psi}(t(s); \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2]^{-1}$$

for $s \geqq 0$.

Then

$$\begin{aligned}\frac{d}{dt} \ln (1 - z) &= \frac{1}{1 - z} \bar{f}(\mathbf{x}, \boldsymbol{\psi}) \\ &= f_0(\mathbf{x}, \boldsymbol{\psi})\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dx} \ln (1 - z) &= f_0(\mathbf{x}, \boldsymbol{\psi}) / |\mathbf{c}(\mathbf{x})| \\ &= g(\mathbf{x}, \boldsymbol{\psi}), \quad \text{say.}\end{aligned} \tag{2.28}$$

Now there are $h > 0$, $k > 0$ such that $g(\mathbf{x}, \mathbf{u}) < 0$ for $\mathbf{x} \in \Delta_h - \operatorname{int} \Delta_0$, $|\mathbf{u} - \mathbf{u}_1(\mathbf{u})| < k$.

So if $1 - z < k$ for $\mathbf{p}$ in $\partial \Delta_0$ then $1 - z$ remains less than k while $\boldsymbol{\phi}_{\mathbf{p}}(t) \in \Delta_h$.

$\Delta - \operatorname{int} \Delta_h$ is a compact set and so there is an $M = \max_{\mathbf{x} \in \Delta - \operatorname{int} \Delta_n} g(\mathbf{x}, \mathbf{u})$ (because $g = f_0 / |\mathbf{c}|$ and $f_0(\mathbf{x}, \mathbf{u})$ is bounded). Therefore $(d/ds) \ln (1 - z) \leqq M$.

Let $m = \max_{\mathbf{x} \in \Delta - \operatorname{int} \Delta_h} s(\mathbf{x})$. Now $1 - z \leqq k$ for $\mathbf{x} \in \Delta_h$, and so

$$\ln (1 - z(s)) \leqq Mm + \ln (1 - z(s_1))$$

for $0 < s_1 \leqq s$. Therefore

$$1 - z(s) \leqq \exp (Mm)(1 - z(s_1)).$$

Putting $\delta = \exp (-Mm) \min (\varepsilon, k)$ we finally obtain $1 - z(s) \leqq \varepsilon$ for $0 < s_1 \leqq s$, giving

$$1 - [1 + |\boldsymbol{\psi}(t; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2]^{-1} \leqq \varepsilon \tag{2.29}$$

if $t_1 \leqq t \leqq T(\mathbf{p})$, and

$$1 - [1 + |\boldsymbol{\psi}(t_1; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2]^{-1} \leqq \delta.$$

Now we must show that this gives $\lim_{t \to \infty} \bar{u}(t, \mathbf{x}) = 1$ uniformly for $\mathbf{x} \in \Delta - \operatorname{int} \Delta_0$. Choose $\varepsilon \in (0, 1)$. Choose δ such that if

$$1 - [1 + |\boldsymbol{\psi}(t_1; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2]^{-1} \leqq \delta$$

then

$$1 - [1 + |\boldsymbol{\psi}(t; \mathbf{p}, \mathbf{r}) - \mathbf{u}_1|^2]^{-1} \leqq \varepsilon/2 \tag{2.30}$$

for all t in the range $[t_1, T(\mathbf{p})]$. Now Lemma 2.4 implies that $\boldsymbol{\phi}_{\mathbf{p}}(t)$ is in $\partial\Delta_0$ for all t if $\mathbf{p}$ is in $\partial\Delta_0$. In other words this characteristic curve lies entirely on $\partial\Delta_0$.

If $\mathbf{x}_0$ is on $\partial\Delta_0$, $\mathbf{r}\in R_0(\mathbf{x}_0)$ and

$$y(t)=[1+|\boldsymbol{\psi}(t;\mathbf{x}_0,\mathbf{r})-\mathbf{u}_1(\boldsymbol{\psi})|^2]^{-1} \tag{2.31}$$

then

$$\frac{dy}{dt}=\bar{f}(\boldsymbol{\phi}_{\mathbf{x}_0}(t),\boldsymbol{\psi}),$$

which, from Lemma 2.7, is positive if $\boldsymbol{\psi}$ is not on $\partial W(\boldsymbol{\phi}_{\mathbf{x}_0}(t))$.

Recall that $R'(\mathbf{x}_0)$ is a closed subset of R^n containing $\partial W(\mathbf{x}_0)$, and if $\mathbf{u}\in R_0(\mathbf{x}_0)$ then there is some $t_0\geqq 0$ such that $\boldsymbol{\psi}(t;\mathbf{x}_0,\mathbf{u})\in R'(\mathbf{x}_0)$ for all $t\geqq t_0$. Let

$$\alpha=\max\{f_0(\mathbf{x}_0,\mathbf{u}):\mathbf{x}_0\in\partial\Delta_0,\mathbf{u}\in R'(\mathbf{x}_0)\}. \tag{2.32}$$

Thus $\alpha<0$. Now $\boldsymbol{\psi}(t;\mathbf{x}_0,\mathbf{r})\in R'$ for all t greater than some $t_0\geqq 0$. So for $t\geqq t_0$

$$\begin{aligned}\frac{d}{dt}\ln(1-y)&=-\frac{1}{(1-y)}\bar{f}(\boldsymbol{\phi}_{\mathbf{x}_0}(t),\boldsymbol{\psi})\\&=f_0(\boldsymbol{\phi}_{\mathbf{x}_0}(t),\boldsymbol{\psi})\\&<\alpha\end{aligned}$$

because $\boldsymbol{\psi}$ approaches $\partial W(\boldsymbol{\phi}_{\mathbf{x}_0}(t))$ and never leaves $R'(\boldsymbol{\phi}_{\mathbf{x}_0}(t))$. So

$$1-y(t)\leqq(1-y(t_0))\,e^{\alpha(t-t_0)}. \tag{2.33}$$

So there exists $t_1>0$ such that $1-y(t)<\delta/2$ for $t\geqq t_1$, $\mathbf{r}\in R_0(\mathbf{x}_0)$. Also, since $\mathbf{v}$ is continuous, there is $h_0>0$ such that

$$1-[1+|\boldsymbol{\psi}(t_1;\mathbf{x},\mathbf{v}(\mathbf{x}))-\mathbf{u}_1|^2]^{-1}<\delta$$

for $\mathbf{x}\in\Delta_{h_0}$, such that $\min_{\mathbf{x}\in\partial\Delta_{h_0}}T(\mathbf{x})\geqq t_1$. Now the conditions for the first part are satisfied, giving

$$1-[1+|\boldsymbol{\psi}(t;\mathbf{x},\mathbf{v}(\mathbf{x}))-\mathbf{u}_1|^2]^{-1}<\varepsilon/2$$

for all $t\geqq t_1$. So

$$|\boldsymbol{\psi}-\mathbf{u}_1|^2<\frac{\varepsilon/2}{1-\varepsilon/2}<\varepsilon,$$

since $\varepsilon<1$.

So if $t\geqq t_1$ we have $\boldsymbol{\phi}(0;t,\mathbf{x})\in\Delta_{h_0}$, and hence

$$|\boldsymbol{\psi}(t;\boldsymbol{\phi}(0;t,\mathbf{x}),\mathbf{v}(\boldsymbol{\phi}(0;t,\mathbf{x})))-\mathbf{u}_1|^2<\varepsilon. \tag{2.34}$$

So $\lim_{t\to\infty}|\mathbf{u}(t,\mathbf{x})-\mathbf{u}_1|=0$.

3. Chaos. Recall that for $\mathbf{x}\in\Delta_0$, $S_0(\mathbf{x})$ was defined as $S_0(\mathbf{x})=W(\mathbf{x})-R_0(\mathbf{x})$. From now on we will assume that for each $\mathbf{x}$ in Δ_0, $S_0(\mathbf{x})$ consists of just one point $\mathbf{u}_0(\mathbf{x})$.

Now we show that chaos exists if $\mathbf{v}(\mathbf{x})=\mathbf{u}_0(\mathbf{x})$ for each $\mathbf{x}\in\Delta_0$.

Let

$$V_+=\{\mathbf{v}\in C(\Delta):\mathbf{v}(\mathbf{x})\in R_0(\mathbf{x})\ \forall\mathbf{x}\in\Delta_0\} \tag{3.1}$$

and

$$V_0=\{\mathbf{v}\in C(\Delta):\mathbf{v}(\mathbf{x})=\mathbf{u}_0(\mathbf{x})\ \forall\mathbf{x}\in\Delta_0\}. \tag{3.2}$$

LEMMA 3.1. *Both V_+ and V_0 are invariant under S_t where S_t is an element of the semidynamical system $\{S_t\}_{t\geqq 0}$ which maps $\mathbf{u}(t_0,\cdot)$ to $\mathbf{u}(t_0+t,\cdot)$.*

Proof. Recall that

$$\begin{aligned}(S_t\mathbf{v})(\mathbf{x})&=\mathbf{u}(t,\mathbf{x})\\&=\boldsymbol{\psi}(t;\mathbf{x}_0,\mathbf{v}(\mathbf{x}_0))\end{aligned}$$

where $\mathbf{x}_0=\boldsymbol{\phi}(0;t,\mathbf{x})$; also $\boldsymbol{\psi}$ satisfies

$$\frac{d\psi}{dt}=\mathbf{f}(\mathbf{x},\boldsymbol{\psi}).$$

Now if $\mathbf{x}\in\Delta_0$ we have $\boldsymbol{\phi}(0;t,\mathbf{x})\in\Delta_0$ for all $t\geqq 0$. Hence

$$\mathbf{v}(\boldsymbol{\phi}(0;t,\mathbf{x}))=\mathbf{u}_0(\boldsymbol{\phi}(0;t,\mathbf{x})),$$

which is in $S_0(\boldsymbol{\phi}(0;t,\mathbf{x}))$. But $S_0(\mathbf{x})$ is defined as the subset of $W(\mathbf{x})$ for which $\boldsymbol{\psi}(t;\mathbf{x},\mathbf{u})$ does not approach $\partial W(\boldsymbol{\phi}_{\mathbf{x}}(t))$ as $t\to\infty$. Therefore if $\mathbf{u}\in S_0(\mathbf{x})$ it follows that $\boldsymbol{\psi}(t;\mathbf{x},\mathbf{u})$ lies in $S_0(\boldsymbol{\phi}_{\mathbf{x}}(t))$ for all $t\geqq 0$. Thus for each $\mathbf{x}\in\Delta_0$ we have $(S_t\mathbf{v})(\mathbf{x})\in S_0(\mathbf{x})$, giving $(S_t\mathbf{v})(\mathbf{x})=\mathbf{u}_0(\mathbf{x})$ for all t. So V_0 is invariant under S_t.

Similarly if $\mathbf{v}(\mathbf{x})\in R_0(\mathbf{x})$ for all $\mathbf{x}$ in Δ_0 then $(S_t\mathbf{v})(\mathbf{x})$ is also in $R_0(\mathbf{x})$ for all $\mathbf{x}$ in Δ_0. So V_+ is invariant under S_t. We have shown that if $\mathbf{v}\in V_+$ then $(S_t\mathbf{v})(\mathbf{x})$ approaches $\partial W(\mathbf{x})$ as $t\to\infty$.

Let

$$V_W=\{\mathbf{v}\in V_0\colon \mathbf{v}(\mathbf{x})\in\operatorname{int} W(\mathbf{x})\ \forall\mathbf{x}\in\Delta\}. \tag{3.3}$$

We demonstrate that this is a global attractor.

LEMMA 3.2. *If* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5), *then V_W is invariant under S_t, and for each $\mathbf{v}\in V_0$ there is some $T_0\geqq 0$ such that $S_t\mathbf{v}\in V_W$ for all $t\geqq T_0$.*

Proof. Now $\partial W(\mathbf{x})$ is invariant under S_t since if $\mathbf{u}(t_0,\mathbf{x})$ lies on $W(\mathbf{x})$ for some $t_0\geqq 0$ and all $\mathbf{x}\in\Delta$ then $\mathbf{u}(t,\mathbf{x})$ lies on $\partial W(\mathbf{x})$ for all $\mathbf{x}$ and all $t\geqq t_0$. Now

$$(S_t\mathbf{v})(\mathbf{x})=\boldsymbol{\psi}(t;\boldsymbol{\phi}(0;t,\mathbf{x}),\mathbf{v}(\boldsymbol{\phi}(0;t,\mathbf{x}))). \tag{3.4}$$

If $\mathbf{v}(\mathbf{x})$ lies in $\partial W(\mathbf{x})$ for all $\mathbf{x}$ in Δ, then $(S_t\mathbf{v})(\mathbf{x})$ remains in $\partial W(\mathbf{x})$ for all $\mathbf{x}\in\Delta$ and all $t\geqq 0$.

The property of uniqueness of solutions ensures that if $\mathbf{v}(\boldsymbol{\phi}(0;t,\mathbf{x}))$ is in $\operatorname{int} W(\boldsymbol{\phi}(0;t,\mathbf{x}))$ then $(S_t\mathbf{v})(\mathbf{x})\in\operatorname{int} W(\mathbf{x})$ for all $t\geqq 0$. Similarly if $\mathbf{v}(\boldsymbol{\phi}(0;t,\mathbf{x}))$ is in $R^n-W(\boldsymbol{\phi}(0;t,\mathbf{x}))$ then $(S_t\mathbf{v})(\mathbf{x})\in R^n-W(\mathbf{x})$ for all $t\geqq 0$. This, together with the invariance of V_0, gives invariance of V_W.

Also if $\mathbf{v}\in V_0$ we know that $\mathbf{v}(\mathbf{x})=\mathbf{u}_0(\mathbf{x})$ for all $\mathbf{x}$ in Δ_0. Since $\partial W(\mathbf{x})$ is contained in $R_0(\mathbf{x})$ for all $\mathbf{x}\in\Delta_0$, it follows that in this case $\mathbf{v}(\mathbf{x})\in\operatorname{int} W(\mathbf{x})$.

So there is some $h_0>0$ such that $\mathbf{v}(\mathbf{x})\in\operatorname{int} W(\mathbf{x})$ for all $\mathbf{x}$ in $\Delta_{h_0}-\operatorname{int}\Delta_0$. So if $T_0=\max_{\mathbf{x}\in\partial\Delta_{h_0}} T(\mathbf{x})$, $\mathbf{x}\in\Delta-\Delta_0$ and $t_0\geqq T_0$, we have $\boldsymbol{\phi}(0;t,\mathbf{x})\in\Delta_{h_0}$ giving $(S_{t_0}\mathbf{v})(\mathbf{x})\in\operatorname{int} W(\mathbf{x})$, because $\mathbf{v}(\boldsymbol{\phi}(0;t_0,\mathbf{x}))$ is in $\operatorname{int} W(\boldsymbol{\phi}(0;t_0,\mathbf{x}))$.

Now extend $\mathbf{u}_0$ to be a continuous function from Δ to R^n with $\mathbf{u}_0(\mathbf{x})\in\operatorname{int} W(\mathbf{x})$ for each $\mathbf{x}\in\Delta$.

LEMMA 3.3. *Suppose* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5), *and that for each $\mathbf{x}$ in Δ_0, $S_0(\mathbf{x})$ consists of the single point $\mathbf{u}_0(\mathbf{x})$. Let $\mathbf{v}_0\in V_W$. Let $\delta<\max_{\mathbf{x}\in\Delta_0} d(\mathbf{u}_0(\mathbf{x}),\partial W(\mathbf{x}))$, and let*

$$V_\delta=\{\mathbf{v}_0+\delta\hat{\mathbf{n}}\colon |\hat{\mathbf{n}}|=1\}, \tag{3.5}$$

with

$$S_tV_\delta=\{S_t\mathbf{v}\colon \mathbf{v}\in V_\delta\}.$$

For $\varepsilon > 0$ denote by $W_\varepsilon(\mathbf{x})$ the subset of $W(\mathbf{x})$ consisting of points $\mathbf{x}$ such that the open ball centered at $\mathbf{x}$ with radius ε lies in $W(\mathbf{x})$. Let $\varepsilon > 0$ be sufficiently small so that for each $\mathbf{x}$ in Δ, $W_\varepsilon(\mathbf{x})$ is nonempty and simply connected.

Then there is some $t_0 > 0$ such that for all $t \geqq t_0$ and all $\mathbf{x} \in \Delta$, $(S_t V_\delta)(\mathbf{x})$ lies in $W(\mathbf{x})$ and encloses $W_\varepsilon(\mathbf{x})$.

Proof. From Theorem 2.2 and Lemma 3.2 it is clear that there is some $t_0 > 0$ such that if $t > t_0$, $\mathbf{x} \in \Delta$ and $\mathbf{v} \in V_\delta$, then $(S_t \mathbf{v})(\mathbf{x}) \in W(\mathbf{x})$ and lies within distance ε of $\partial W(\mathbf{x})$. It remains to show that $(S_t V_\delta)(\mathbf{x})$ encloses all of $W_\varepsilon(\mathbf{x})$.

Consider $\mathbf{x}_0 \in \Delta_0$. Now $\mathbf{v}_0(\mathbf{x}_0)$ is inside $V_\delta(\mathbf{x}_0)$, and so $(S_{t_0}\mathbf{v}_0)(\mathbf{x}_0)$ lies inside $(S_{t_0} V_\delta)(\mathbf{x}_0)$. But $(S_{t_0}\mathbf{v}_0)(\mathbf{x}_0) = \mathbf{u}_0(\mathbf{x}_0)$, whose distance from $\partial W(\mathbf{x}_0)$ is greater than ε. Hence $(S_{t_0} V_\delta)(\mathbf{x}_0)$ encloses at least one point of $W_\varepsilon(\mathbf{x}_0)$. Therefore $(S_{t_0} V_\delta)(\mathbf{x}_0)$ must enclose all of $W_\varepsilon(\mathbf{x}_0)$, since all points of $(S_{t_0} V_\delta)(\mathbf{x}_0)$ lie outside $W_\varepsilon(\mathbf{x}_0)$.

Now $(S_{t_0} V_\delta)(\mathbf{x})$ and $W(\mathbf{x})$ both vary continuously with $\mathbf{x}$. Also, $W_\varepsilon(\mathbf{x})$ is nonempty and connected for each $\mathbf{x} \in \Delta$. So it follows that for each $\mathbf{x} \in \Delta$, $(S_{t_0} V_\delta)(\mathbf{x})$ encloses $W_\varepsilon(\mathbf{x})$.

LEMMA 3.4. *Suppose* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5). *Let* $\mathbf{q}$ *be an element of* V_W. *Let* $\delta > 0$, $\tau \geqq 0$ *and* $h_0 > 0$. *Then there is some* $t_0 > \tau$, *numbers* $\beta \geqq \alpha > 0$ *with* $\beta \leqq h_0$ *and a continuous mapping* $\mathbf{p}$ *from* $\mathbf{x}$ *in* $\Delta_\beta - \operatorname{int} \Delta_\alpha$ *to* $\mathbf{p}(\mathbf{x}) \in R^n$ *with the following properties*:

1) *For all* $\mathbf{x}$ *in* $\Delta_\beta - \operatorname{int} \Delta_\alpha$ *we have* $|\mathbf{p}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \delta$;

2) *For each* $\mathbf{v} \in V_W$ *with*

$$|\mathbf{v}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x} \in \partial\Delta_\alpha} |\mathbf{p}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_\alpha - \operatorname{int} \Delta_0 \tag{3.6}$$

and

$$\mathbf{v}(\mathbf{x}) = \mathbf{p}(\mathbf{x}) \quad \forall \mathbf{x} \in \Delta_\alpha - \operatorname{int} \Delta_\beta$$

we have

$$|(S_{t_0}\mathbf{v})(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x} \in \partial\Delta_{h_0}} |\mathbf{q}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_{h_0} - \operatorname{int} \Delta_0 \tag{3.7}$$

and

$$(S_{t_0}\mathbf{v})(\mathbf{x}) = \mathbf{q}(\mathbf{x}) \quad \forall \mathbf{x} \in \Delta - \operatorname{int} \Delta_{h_0}.$$

Proof. Let

$$q_{h_0} = \max_{\mathbf{x} \in \partial\Delta_{h_0}} |\mathbf{q}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})|. \tag{3.8}$$

Without loss of generality assume that $|\mathbf{q}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq q_{h_0}$ if $\mathbf{x} \in \Delta_{h_0}$.

Let θ be a continuous function from Δ to $[0, 1]$ such that $\theta(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Delta_0$, $0 < \theta(\mathbf{x}) < 1$ for all $\mathbf{x}$ in $\operatorname{int} \Delta_{h_0} - \Delta_0$ and $\theta(\mathbf{x}) = 0$ for all $\mathbf{x}$ outside Δ_{h_0}. Let V_δ be the subset of V_+ consisting of functions $\bar{\mathbf{v}}$ that satisfy

$$\bar{\mathbf{v}}(\mathbf{x}) = \tfrac{1}{2}\delta\theta(\mathbf{x})\mathbf{n} + u_0(\mathbf{x}) \tag{3.9}$$

for all $\mathbf{x}$ in $\Delta - \operatorname{int} \Delta_0$, where $\mathbf{n}$ is some unit vector in R^n. Denote by $V_\delta(\mathbf{x})$ the set of points $\{\bar{\mathbf{v}}(\mathbf{x}) \colon \bar{\mathbf{v}} \in V_\delta\}$. For each $\mathbf{x} \in \operatorname{int} \Delta_{h_0}$, $V_\delta(\mathbf{x})$ forms a surface homeomorphic to S^{n-1}, and $u_0(\mathbf{x})$ lies inside this surface. Denote by $(S_t V_\delta)(\mathbf{x})$ the set of points $\{(S_t\bar{\mathbf{v}})(\mathbf{x}) \colon \bar{\mathbf{v}} \in V_\delta\}$. This is a surface homeomorphic to $V_\delta(\mathbf{x})$.

From Theorem 2.2 we know that for each element $\bar{\mathbf{v}}$ of V_δ, $(S_t\bar{\mathbf{v}})(\mathbf{x})$ approaches $\partial W(\mathbf{x})$ uniformly for $\mathbf{x}$ in Δ. Also since $u_0(\mathbf{x})$ lies inside $V_\delta(\mathbf{x})$ for each $\mathbf{x}$ in $\operatorname{int} \Delta_{h_0}$, and $\mathbf{q} \in V_W$ it follows, as in Lemma 3.3, that there exists $t_0 > \max_{\mathbf{x} \in \partial\Delta_{h_0}} T(\mathbf{x})$ such that $\mathbf{q}(\mathbf{x})$ is inside the surface $(S_{t_0})(\mathbf{x})$ for all $\mathbf{x}$ in Δ. So $\boldsymbol{\phi}(0; t, \mathbf{x}) \in \Delta_{h_0}$ for all $t \geqq t_0$, and for all $\mathbf{x}$ in Δ.

Let $\psi_0(t; \mathbf{x})$ be the solution of

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\boldsymbol{\phi}(t;\, t_0, \mathbf{x}), \mathbf{u}) \tag{3.10}$$

with $\psi_0(t_0, \mathbf{x}) = \mathbf{q}(\mathbf{x})$. Now $\psi_0(t_0, \mathbf{x})$ lies inside $(S_{t_0} V_\delta)(\mathbf{x})$. So $\psi_0(t; \mathbf{x})$ lies inside $(S_t V_\delta)(\mathbf{x})$ for all $t \geqq 0$. Putting $t = 0$ we find that $\psi_0(0; \mathbf{x})$ is inside $V_\delta(\boldsymbol{\phi}(0;\, t_0, \mathbf{x}))$.

So $|\psi_0(0; \mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta$ for all $\mathbf{x}$ in $\Delta - \Delta_0$ since $|\bar{\mathbf{v}}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta$ for all $\mathbf{x}$ in $\Delta - \Delta_0$.

Now if $\mathbf{x} \in \partial\Delta_0$, then $\psi(t_0; \mathbf{x}, \mathbf{u}_0(\mathbf{x})) = \mathbf{u}_0(\mathbf{x})$. So there exist numbers $r_0 > 0$ and $h > 0$ such that if $\mathbf{x} \in \Delta_h - \operatorname{int} \Delta_0$ and $\mathbf{u} \in R^n$ with $|\mathbf{u} - \mathbf{u}_0(\mathbf{x})| < r_0$ then we have $|\psi(t_0; \mathbf{x}, \mathbf{u}) - \mathbf{u}_0(\mathbf{x})| < q_{h_0}$.

Now by the definition of t_0 we have $\boldsymbol{\phi}(0;\, t_0, \mathbf{x}) \in \operatorname{int} \Delta_{h_0}$ for all $\mathbf{x}$ in Δ. Let $\beta < h_0$ such that $\boldsymbol{\phi}(0;\, t_0, \mathbf{x}) \in \Delta_\beta$ for all $\mathbf{x}$ in Δ and choose α such that

1) $0 < \alpha < \beta$,
2) $\boldsymbol{\phi}(0;\, t_0, \mathbf{x}) \in \Delta_\beta - \operatorname{int} \Delta_\alpha$ for all $\mathbf{x}$ in $\Delta - \operatorname{int} \Delta_{h_0}$,
3) $|\psi(0; \boldsymbol{\phi}_{\mathbf{x}}(t_0)) - \mathbf{u}_0(\mathbf{x})| < r_0$ for all $\mathbf{x}$ in Δ_α.

Define

$$\mathbf{p}(\mathbf{x}) = \psi(0; \boldsymbol{\phi}_{\mathbf{x}}(t_0)) \tag{3.11}$$

for all $\mathbf{x}$ in $\Delta_\beta - \operatorname{int} \Delta_\alpha$. Now

$$\psi(t_0; \boldsymbol{\phi}_{\mathbf{x}}(t_0)) = \mathbf{q}(\boldsymbol{\phi}_{\mathbf{x}}(t_0)).$$

We have shown that $|\psi(0; \mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta$, which implies that $|\mathbf{p}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta$ for all $\mathbf{x}$ in $\Delta_\beta - \operatorname{int} \Delta_\alpha$.

Define

$$p_\alpha = \max_{\mathbf{x} \in \partial\Delta_\alpha} |\mathbf{p}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})|. \tag{3.12}$$

Choose $\mathbf{v} \in V_W$ such that

$$|\mathbf{v}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq p_\alpha \quad \forall \mathbf{x} \text{ in } \Delta_\alpha,$$

$$\mathbf{v}(\mathbf{x}) = \mathbf{p}(\mathbf{x}) \quad \forall \mathbf{x} \text{ in } \Delta_\beta - \operatorname{int} \Delta_\alpha.$$

Now $\boldsymbol{\phi}(0;\, t_0, \mathbf{x}) \in \Delta_\beta - \operatorname{int} \Delta_\alpha$ for all $\mathbf{x}$ in $\Delta - \operatorname{int} \Delta_{h_0}$, and so

$$\mathbf{p}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x})) = \mathbf{v}(\boldsymbol{\phi}0;\, t_0, \mathbf{x}))$$

for all $\mathbf{x}$ in $\Delta - \operatorname{int} \Delta_{h_0}$. Therefore

$$\begin{aligned} (S_{t_0}\mathbf{v})(\mathbf{x}) &= \psi(t_0; \boldsymbol{\phi}(0;\, t_0, \mathbf{x}), \mathbf{p}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x}))) \\ &= \psi(t_0; \boldsymbol{\phi}(0;\, \mathrm{t}_0, \mathbf{x}), \psi_0(0; \mathbf{x})) \\ &= \psi_0(t_0; \mathbf{x}) \\ &= \mathbf{q}(\mathbf{x}) \end{aligned} \tag{3.13}$$

if $\mathbf{x}$ is outside Δ_{h_0}.

If $\mathbf{x} \in \Delta_{h_0}$ and $\boldsymbol{\phi}(0;\, t_0, \mathbf{x})$ is not in Δ_α, then

$$(S_{t_0}\mathbf{v})(\mathbf{x}) = \mathbf{q}(\mathbf{x})$$

as before, giving $|(S_{t_0}\mathbf{v})(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < q_{h_0}$.

If $\mathbf{x} \in \Delta_{h_0} - \operatorname{int} \Delta_0$ and $\boldsymbol{\phi}(0;\, t_0, \mathbf{x}) \in \Delta_\alpha$, then

$$|\mathbf{v}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x})) - \mathbf{u}_0(\mathbf{x})| \leqq p_\alpha = \max_{\mathbf{x} \in \partial\Delta_\alpha} |\mathbf{p}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})|.$$

So if $\mathbf{x}_\alpha$ is on $\partial\Delta_\alpha$ we have

$$\begin{aligned}|\mathbf{v}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x})) - \mathbf{u}_0(\mathbf{x})| &\leqq |\mathbf{p}(\mathbf{x}_\alpha) - \mathbf{u}_0(\mathbf{x})| \\ &\leqq |\boldsymbol{\psi}_0(0;\, \boldsymbol{\phi}_{\mathbf{x}_\alpha}(t_0)) - \mathbf{u}_0(\mathbf{x})| \\ &\leqq r_0.\end{aligned}$$

Now

$$(S_{t_0}\mathbf{v})(\mathbf{x}) = \boldsymbol{\psi}(t_0;\, \boldsymbol{\phi}(0;\, t_0, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x}))),$$

and therefore

$$\begin{aligned}|(S_{t_0}\mathbf{v})(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| &= |\boldsymbol{\psi}(t_0;\, \boldsymbol{\phi}(0;\, t_0, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0;\, t_0, \mathbf{x}))) - \mathbf{u}_0(\mathbf{x})| \\ &\leqq \sup_{\substack{\mathbf{x}\in\Delta_{h_0} \\ |\mathbf{r}-\mathbf{u}_0(\mathbf{x})|<r_0}} |\boldsymbol{\psi}(t_0;\, \mathbf{p}, \mathbf{r}) - \mathbf{u}_0(\mathbf{x})| \\ &< q_{h_0} \quad \text{as required.}\end{aligned}$$

THEOREM 3.1. *If* **c** *and* **f** *satisfy* (C1)-(C2) *and* (F1)-(F5) *and* $S_0(\mathbf{x}) = \{\mathbf{u}_0(\mathbf{x})\}$ *for each* $\mathbf{x} \in \Delta_0$, *then the semidynamical system* $\{S_t\}_{t\geqq 0}$ *is chaotic in* V_W.

Proof. First, we show instability of each element of V_W. Let

$$\varepsilon = \tfrac{1}{4} \min_{\substack{\mathbf{x}\in\partial\Delta \\ \mathbf{u}\in\partial W(\mathbf{x})}} |\mathbf{u} - \mathbf{u}_0(\mathbf{x})|. \tag{3.14}$$

Choose $\delta > 0$ with

$$\delta < \min_{\substack{\mathbf{x}\in\Delta-\text{int}\,\Delta_0 \\ \mathbf{u}\in\partial W(\mathbf{x})}} |\mathbf{u} - \mathbf{u}_0(\mathbf{x})|. \tag{3.15}$$

Let $\mathbf{v}_0$ be an element of V_W; that is, $\mathbf{v}_0$ is a continuous function from Δ to R^n with $\mathbf{v}_0(\mathbf{x}) = \mathbf{u}_0(\mathbf{x})$ if $\mathbf{x}$ is in $\partial\Delta_0$, and $\mathbf{v}_0(\mathbf{x})$ is in the interior of $W(\mathbf{x})$ for each $\mathbf{x}$ in $\Delta - \Delta_0$. Let

$$V_\delta = \{\mathbf{v} \in V_W \colon \max_{\mathbf{x}\in\Delta} |\mathbf{v}(\mathbf{x}) - \mathbf{v}_0(\mathbf{x})| < \delta\}. \tag{3.16}$$

Now there exists $h_0 > 0$ such that

$$|\mathbf{v}_0(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta/2 \tag{3.17}$$

for all $\mathbf{x}$ in Δ_{h_0}, since $\mathbf{v}_0(\mathbf{x}) = \mathbf{u}_0(\mathbf{x})$ for all $\mathbf{x}$ in Δ_0.

Choose $\mathbf{q}_1$, $\mathbf{q}_2$ in V_W such that $|\mathbf{q}_1(\mathbf{x}) - \mathbf{q}_2(\mathbf{x})| > 2\varepsilon$ for some $\mathbf{x}$ in Δ.

Applying Lemma 3.4 with $\tau = 0$, we can choose $t_0 > 0$ and $\mathbf{p}_1$, $\mathbf{p}_2$ defined on $\Delta_{\beta_1} - \text{int}\,\Delta_{\alpha_1}$ and $\Delta_{\beta_2} - \text{int}\,\Delta_{\alpha_2}$ respectively, with the following properties:

1) $|\mathbf{p}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta/2$ in the range of definition of $\mathbf{p}_i$;

2) For each $\mathbf{v}_i$ in V_W with

$$|\mathbf{v}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x}\in\partial\Delta_{\alpha_i}} |\mathbf{p}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_{\alpha_i},$$

$$\mathbf{v}_i(\mathbf{x}) = \mathbf{p}_i(\mathbf{x}) \quad \forall \mathbf{x} \in \Delta_{\beta_i} - \text{int}\,\Delta_{\alpha_i}$$

we have

$$|(S_{t_0}\mathbf{v}_i)(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x}\in\partial\Delta_{h_0}} |\mathbf{q}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_{h_0},$$

$$(S_{t_0}\mathbf{v}_i)(\mathbf{x}) = \mathbf{q}_i(\mathbf{x}) \quad \forall \mathbf{x} \in \Delta - \text{int}\,\Delta_{h_0}.$$

Let $\mathbf{v}_i(\mathbf{x})$ be a continuous function from Δ to R^n, with

$$\mathbf{v}_i(\mathbf{x}) = \begin{cases} \mathbf{u}_0(\mathbf{x}) & \forall \mathbf{x} \in \Delta_0, \\ \mathbf{p}_i(\mathbf{x}) & \forall \mathbf{x} \in \Delta_{\beta_i} - \operatorname{int} \Delta_{\alpha_i}, \\ \mathbf{v}_0(\mathbf{x}) & \forall \mathbf{x} \in \Delta - \operatorname{int} \Delta_{h_0}, \end{cases} \tag{3.18}$$

$$|\mathbf{v}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \delta/2 \quad \forall \mathbf{x} \in \Delta_{h_0}$$

and

$$|\mathbf{v}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x} \in \partial\Delta_{\alpha_i}} |\mathbf{p}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_{\alpha_i}.$$

Now for $\mathbf{x}$ outside Δ_{h_0} we have

$$|\mathbf{v}_i(\mathbf{x}) - \mathbf{v}_0(\mathbf{x})| = 0.$$

For $\mathbf{x}$ in $\Delta_{h_0} - \operatorname{int} \Delta_0$, (3.17) implies that $|\mathbf{v}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \delta/2$; and (3.18) implies that $|\mathbf{v}_0(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \delta/2$, which gives

$$|\mathbf{v}_i(\mathbf{x}) - \mathbf{v}_0(\mathbf{x})| \leqq \delta.$$

Also, for $\mathbf{x} \in \Delta_0$ we have

$$\mathbf{v}_i(\mathbf{x}) = \mathbf{v}_0(\mathbf{x}) = \mathbf{u}_0(\mathbf{x}).$$

Hence $|\mathbf{v}_i(\mathbf{x}) - \mathbf{v}_0(\mathbf{x})| \leqq \delta$ for all $\mathbf{x}$ in Δ, and so $\mathbf{v}_i \in V_\delta$.

So

$$\begin{aligned} \mathbf{v}_i(\mathbf{x}) &= \mathbf{p}_i(\mathbf{x}) && \forall \mathbf{x} \in \Delta_{\beta_i} - \operatorname{int} \Delta_{\alpha_i}, \\ |\mathbf{v}_i(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| &\leqq \mathbf{p}_\alpha && \forall \mathbf{x} \in \Delta_{\alpha_i}. \end{aligned}$$

Therefore

$$(S_{t_0}\mathbf{v}_i)(\mathbf{x}) = \mathbf{q}_i(\mathbf{x}) \quad \text{for } \mathbf{x} \text{ in } \Delta - \operatorname{int} \Delta_{h_0},$$

and so

$$\begin{aligned} |(S_{t_0}\boldsymbol{v}_2)(\mathbf{x}) - (S_{t_0}\mathbf{v}_1)(\mathbf{x})| &= |\mathbf{q}_2(\mathbf{x}) - \mathbf{q}_1(\mathbf{x})| \\ &> \varepsilon \end{aligned} \tag{3.19}$$

at some point $\mathbf{x}$ in $\partial\Delta$. This holds for all $\delta > 0$. So there is no uniform convergence, and each element of V_W is unstable under $\{S_t\}_{t \geqq 0}$.

For the second part of the proof it remains to show that some element of V_W has a dense orbit.

Let $\{\mathbf{q}_n\}_{n=1}^{\infty}$ be a dense set in V_W and suppose that $\mathbf{q}_n(\mathbf{x}) \neq \mathbf{u}_0(\mathbf{x})$ for all $\mathbf{x}$ in $\Delta - \Delta_0$. Let

$$b = \min_{\substack{\mathbf{x} \in \Delta \\ \mathbf{u} \in W(\mathbf{x})}} |\mathbf{u} - \mathbf{u}_0(\mathbf{x})| \tag{3.20}$$

and let $\delta = b/2$. Choose $h_1 \in (0, \frac{1}{2})$ such that $|\mathbf{q}_1(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq b/2$ for all $\mathbf{x}$ in Δ_{h_1}. Let $\tau_1 = 0$.

Apply Lemma 3.4 to get a function $\mathbf{p}_1$ defined on $\Delta_{\beta_1} - \operatorname{int} \Delta_{\alpha_1}$ and a number t_{01} such that

1) $|\mathbf{p}_1(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| < \delta_1/2$ in the range of definition of $\mathbf{p}_1$;
2) For each $\mathbf{v}$ in V_W with

$$|\mathbf{v}(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \leqq \max_{\mathbf{x} \in \partial\Delta_{\alpha_1}} |\mathbf{p}_1(\mathbf{x}) - \mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x} \in \Delta_{\alpha_1},$$

$$\mathbf{v}(\mathbf{x}) = \mathbf{p}_1(\mathbf{x}) \quad \forall \mathbf{x} \in \Delta_{\beta_1} - \operatorname{int} \Delta_{\alpha_1}$$

we have

$$|(S_{t_0}\mathbf{v})(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq \max_{\mathbf{x}\in\partial\Delta_{h_0}} |\mathbf{q}_1(\mathbf{x})-\mathbf{u}_0(\mathbf{x})| \quad \forall \mathbf{x}\in\Delta_{h_0},$$

$$(S_{t_0}\mathbf{v})(\mathbf{x})=\mathbf{q}_1(\mathbf{x}) \quad \forall \mathbf{x}\in\Delta-\operatorname{int}\Delta_{h_0}.$$

For each integer $n>1$ let

$$\delta_n=\min\left(\frac{b}{2^n}, \max_{\mathbf{x}\in\Delta_{\alpha_{n-1}}} |\mathbf{p}_{n-1}(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\right)$$

and choose $h_n>0$ with $h_n<\min(2^{-n}, \alpha_{n-1})$ such that $|\mathbf{q}_n(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq(b/2^n)$ for all $\mathbf{x}$ in Δ_{h_n}. Let $\tau_n=t_{0,n-1}$. Again apply Lemma 3.4 to get $\mathbf{p}_n$ defined on $\Delta_{\beta_n}-\operatorname{int}\Delta_{\alpha_n}$ and $t_{0n}\geqq t_{0,n-1}$.

Now

$$0<\cdots<\alpha_n<\beta_n<h_n<\alpha_{n-1}<\cdots.$$

Also, $\lim_{n\to\infty}\delta_n=0$, and $\lim_{n\to\infty}d_n=0$. Let $\mathbf{v}$ be a continuous function from Δ to R^n satisfying

$$\begin{aligned} &\mathbf{v}(\mathbf{x})=\mathbf{u}_0(\mathbf{x}) \quad \forall\mathbf{x}\in\Delta_0,\\ &\mathbf{v}(\mathbf{x})=\mathbf{p}_n(\mathbf{x}) \quad \forall\mathbf{x}\in\Delta_{\beta_n}-\operatorname{int}\Delta_{\alpha_n},\\ &|\mathbf{v}(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq\max_{\mathbf{x}\in\Delta_{\alpha_n}}|\mathbf{p}_n(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|.\end{aligned} \tag{3.21}$$

Now $\mathbf{v}$ is continuous on $\partial\Delta_0$ because $|\mathbf{p}_n(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq\delta_n$ in the area of definition of $\mathbf{p}_n$. Clearly $\mathbf{v}\in V_W$.

Also

$$\begin{aligned}|(S_{t_{0n}}\mathbf{v})(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|&\leqq\max_{\mathbf{x}\in\partial\Delta_{h_n}}|\mathbf{q}_n(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\\ &\leqq\frac{b}{2^n}\end{aligned}$$

for $\mathbf{x}\in\Delta_{h_n}$, and

$$(S_{t_{0n}}\mathbf{v})(\mathbf{x})=\mathbf{q}_n(\mathbf{x})$$

if $\mathbf{x}\in\Delta-\operatorname{int}\Delta_{h_n}$, giving

$$|(S_{t_{0n}}\mathbf{v})(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq\frac{b}{2^n}.$$

So

$$|(S_{t_{0n}}\mathbf{v})(\mathbf{x})-\mathbf{u}_0(\mathbf{x})|\leqq\frac{b}{2^n} \quad \forall\mathbf{x}\in\Delta. \tag{3.22}$$

Therefore

$$\sup_{\mathbf{x}\in\Delta}|(S_{t_{0n}}\mathbf{v})(\mathbf{x})-\mathbf{q}_n(\mathbf{x})|\leqq\frac{2b}{2^n},$$

which implies that $\{S_{t_{0n}}\mathbf{v}\}_{n=1}^{\infty}$ is dense in V_W.

4. Periodic orbits. P. Brunovsky [2] proves the following.

THEOREM (Brunovsky). *Let* $\Delta = [0, 1]$. *Assume that*

1) *The function* c *maps* Δ *to* $[0, \infty)$, *the function* f *maps* $\Delta \times [0, \infty)$ *to* R, *and both* c *and* f *are* C^1;

2) $c(0) = 0$ *and* $c(x) > 0$ *for all* $x > 0$;

3) *There is some* $u_0 \in (0, 1]$ *such that* (a) $f(0, u)(u - u_0) < 0$ *for all* $u > 0$, $u \neq u_0$, *and* (b) $f_u(0, u_0) < 0$;

4) *There exist* $k_1 \geqq 0$, $k_2 \geqq 0$ *such that* $f(x, u) \leqq k_1 u + k_2$ *for all* $x \in \Delta$, $u \geqq 0$;

5) $f(x, 0) = 0$ *for all* $x \in \Delta$.

Then we have

(a) *For each* $\tau > 0$ *there is a continuum of functions* v *in* V_W (*defined as before*) *satisfying* $S_\tau v = v$, *that is, functions* v *that give rise to* τ*-periodic behaviour*;

(b) *The set of all periodic points of* S_t *is dense in* V_W.

We now wish to prove an analogous result for the more general case. In fact this theorem generalizes to an almost identical statement.

THEOREM 4.1. *Suppose* $\mathbf{c}$ *and* $\mathbf{f}$ *satisfy* (C1)-(C2) *and* (F1)-(F5), *and for each* $\mathbf{x}$ *in* Δ_0, $S_0(\mathbf{x})$ *consists of just the one point* $\{\mathbf{u}_0(\mathbf{x})\}$. *Then*

(a) *For each* $\tau > 0$ *there is a continuum of functions* $\mathbf{v}$ *in* V_W (*defined as before*) *satisfying* $S_\tau \mathbf{v} = \mathbf{v}$, *that is, functions* $\mathbf{v}$ *that give rise to* τ*-periodic behaviour*;

(b) *The set of all periodic points of* S_t *is dense in* V_W.

Recall that if $\mathbf{x} \in \Delta - \Delta_0$, then $\boldsymbol{\phi}_{\mathbf{x}}(t)$ is the characteristic curve satisfying

$$\frac{\partial \boldsymbol{\phi}}{\partial t} = \mathbf{c}(\boldsymbol{\phi})$$

and $T(\mathbf{x})$ is the (unique) value of t for which $\boldsymbol{\phi}_{\mathbf{x}}(t)$ intersects $\partial\Delta$. Thus $\boldsymbol{\phi}_{\mathbf{x}}(T(\mathbf{x}))$ is the point of $\partial\Delta$ at which this intersection takes place.

We now consider a one-to-one correspondence between functions $\mathbf{v}\colon \Delta - \Delta_0 \to R^n$ and functions $\mathbf{g}\colon \partial\Delta \times [0, \infty) \to R^n$.

Let $\bar{\mathbf{x}} \in \partial\Delta$. Let $\boldsymbol{\Phi}$ be the mapping from the function space $(C(\Delta - \Delta_0))^n$ to the function space $(C(\partial\Delta \times [0, \infty)))^n$ defined by

$$\begin{aligned} \boldsymbol{\Phi}(\mathbf{v})(\bar{\mathbf{x}}, t) &= \boldsymbol{\psi}(t; \boldsymbol{\phi}(0; t, \bar{\mathbf{x}}), \mathbf{v}(\boldsymbol{\phi}(0; t, \bar{\mathbf{x}}))) \\ &= \mathbf{u}(t, \bar{\mathbf{x}}) \end{aligned} \tag{4.1}$$

if $\mathbf{u}(0, \mathbf{x}) = \mathbf{v}(\mathbf{x})$ for all $\mathbf{x} \in \Delta - \Delta_0$. Thus if $\mathbf{g}$ is a function from $\partial\Delta \times [0, \infty)$ to R^n, we have $\mathbf{g} = \boldsymbol{\Phi}(\mathbf{v})$ iff

$$\mathbf{v}(\mathbf{x}) = \boldsymbol{\psi}(-T(\mathbf{x}); \boldsymbol{\phi}_{\mathbf{x}}(T(\mathbf{x})), \mathbf{g}(\boldsymbol{\phi}_{\mathbf{x}}(T(\mathbf{x})), T(\mathbf{x}))) \tag{4.2}$$

for all $\mathbf{x}$ in $\Delta - \Delta_0$.

Therefore we can state the following.

LEMMA 4.1. *The map* $\boldsymbol{\Phi}$ *defined by* (4.1) *is invertible, with inverse given by* (4.2).

Note that here one is not necessarily able to extend $\mathbf{v}$ to a function from Δ to R^n, as $\mathbf{v}(\mathbf{x})$ need not have a limit as $\mathbf{x}$ approaches $\partial\Delta_0$.

LEMMA 4.2. *Let* $\mathbf{g}$ *be a continuous function from* $\partial\Delta \times [0, \infty)$ *to* R^n, *with* $\mathbf{g}(\bar{\mathbf{x}}, t) \in W(\bar{\mathbf{x}})$ *and*

$$d(\mathbf{g}(\bar{\mathbf{x}}, t), \partial W(\bar{\mathbf{x}})) \geqq \eta \tag{4.3}$$

for some $\eta > 0$ *and all* $t \geqq 0$, *all* $\bar{\mathbf{x}} \in \partial\Delta$. *Then* $\mathbf{g} \in \boldsymbol{\phi}(V_W)$.

Proof. We must show that if $\mathbf{x} \in \Delta - \Delta_0$ approaches $\mathbf{x}_0 \in \partial\Delta_0$, then $\lim_{\mathbf{x} \to \mathbf{x}_0} \boldsymbol{\Phi}^{-1}(\mathbf{g})(\mathbf{x}) = \mathbf{u}_0(\mathbf{x}_0)$. This is sufficient, because then we can define

$$\begin{aligned} \mathbf{v}(\mathbf{x}) &= \boldsymbol{\Phi}^{-1}(\mathbf{g})(\mathbf{x}) \qquad \forall \mathbf{x} \in \Delta - \Delta_0, \\ \mathbf{v}(\mathbf{x}_0) &= \lim_{\mathbf{x} \to \mathbf{x}_0} \boldsymbol{\Phi}^{-1}(\mathbf{g})(\mathbf{x}) \quad \forall \mathbf{x}_0 \in \partial\Delta_0. \end{aligned} \tag{4.4}$$

Let $\bar{\mathbf{x}}$ be a point on $\partial\Delta$. Let $\delta > 0$, $0 < h_0 \leqq \frac{1}{2}$ and define V_δ as in (3.9), that is,

$$V_\delta = \{\mathbf{u}_0 + \tfrac{1}{2}\delta\theta\mathbf{n}: |\mathbf{n}| = 1\}$$

where θ is a continuous function from Δ to R, with $\theta(\mathbf{x}) = 1$ if $\mathbf{x} \in \Delta_0$, $\theta(\mathbf{x}) = 0$ if $\mathbf{x}$ is outside Δ_{h_0}, and $0 < \theta(\mathbf{x}) < 1$ otherwise. Define $S_t V_\delta$ as $\{S_t\mathbf{v}: \mathbf{v} \in V_\delta\}$.

Because $\mathbf{v}(\mathbf{x}_0) \in R_0(\mathbf{x}_0)$ for all $\mathbf{x} \in \Delta_0$, we know from Theorem 2.2 that for each $\mathbf{v}$ in V_δ, $\mathbf{\Phi}(\mathbf{v})(\bar{\mathbf{x}}, t)$ approaches $\partial \mathrm{W}(\bar{\mathbf{x}})$ as $t \to \infty$. Therefore there is some time $t_\delta > 0$ such that for each $t \geqq t_\delta$ and each $\mathbf{v} \in V_\delta$ we have $d(\mathbf{\Phi}(\mathbf{v})(\bar{\mathbf{x}}, t), \partial W(\bar{\mathbf{x}})) < \eta$.

Now $\mathbf{\Phi}(V_\delta)(\bar{\mathbf{x}}, t)$ is homeomorphic to $\partial W(\mathbf{x})$. Also for each $\mathbf{x}_0 \in \partial\Delta_0$, $V_\delta(\mathbf{x}_0)$ encloses $\mathbf{u}_0(\mathbf{x}_0)$. So, by an argument similar to that of Lemma 3.1, if $t \geqq t_\delta$, then $\mathbf{\Phi}(V_\delta)(\bar{\mathbf{x}}, t)$ encloses the subset of $W(\bar{\mathbf{x}})$ which lies at distance greater than η from the boundary. Thus $\mathbf{g}(\bar{\mathbf{x}}, t)$ lies inside $\mathbf{\Phi}(V_\delta)(\bar{\mathbf{x}}, t)$. This implies that $\mathbf{\Phi}^{-1}(\mathbf{g})(\mathbf{x})$ lies inside $V_\delta(\mathbf{x})$ for all $\mathbf{x}$ with $T(\mathbf{x}) \geqq t_\delta$. So there is some $h > 0$ such that $V_\delta(\mathbf{x})$ encloses $\mathbf{\Phi}^{-1}(\mathbf{g})(\mathbf{x})$ for all $\mathbf{x}$ in $\Delta_h - \Delta_0$.

However δ was arbitrary, and for each $\mathbf{x}_0 \in \partial\Delta_0$ the surface $V_\delta(\mathbf{x}_0)$ encloses $\mathbf{u}_0(\mathbf{x}_0)$. Therefore $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{\Phi}^{-1}(\mathbf{g})(\mathbf{x}) = \mathbf{u}_0(\mathbf{x}_0)$, and so $\mathbf{g} \in \mathbf{\Phi}(V_W)$.

LEMMA 4.3. *If $\mathbf{\Phi}(\mathbf{v})(\bar{\mathbf{x}}, t)$ is periodic in t with period τ independent of $\bar{\mathbf{x}}$, then the solution $\mathbf{u}(\mathbf{x}, t)$ is periodic with period τ for all $\mathbf{x} \in \Delta - \Delta_0$.*

Proof. Suppose $(S_t\mathbf{v})(\mathbf{x}) \neq (S_{t+\tau}\mathbf{v})(\mathbf{x})$ for some $\mathbf{x} \in \Delta - \Delta_0$ and some $t \geqq 0$. Therefore

$$\boldsymbol{\psi}(t; \boldsymbol{\phi}(0; t, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t, \mathbf{x}))) \neq \boldsymbol{\psi}(t+\tau; \boldsymbol{\phi}(0; t+\tau, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t+\tau, \mathbf{x}))).$$

Consider the situation after time $T(\mathbf{x})$. Then

$$\boldsymbol{\psi}(t+T(\mathbf{x}); \boldsymbol{\phi}(0; t, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t, \mathbf{x}))) \neq \boldsymbol{\psi}(t+T(\mathbf{x})+\tau; \boldsymbol{\phi}(0; t+\tau, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t+\tau, \mathbf{x}))).$$

But after time $T(\mathbf{x})$ we also have $\boldsymbol{\phi}_{\mathbf{x}}(T(\mathbf{x})) \in \partial\Delta$. Thus

$$\boldsymbol{\psi}(t+T(\mathbf{x}); \boldsymbol{\phi}(0; t, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t, \mathbf{x}))) = \mathbf{\Phi}(\mathbf{v})(\boldsymbol{\phi}(t+T(\mathbf{x}); t, \mathbf{x}), t+T(\mathbf{x}))$$

and

$$\begin{aligned}\boldsymbol{\psi}(t+T(\mathbf{x})+\tau;\ & \boldsymbol{\phi}(0; t+\tau, \mathbf{x}), \mathbf{v}(\boldsymbol{\phi}(0; t+\tau, \mathbf{x}))) \\ &= \mathbf{\Phi}(\mathbf{v})(\boldsymbol{\phi}(t+T(\mathbf{x})+\tau; t+\tau, \mathbf{x}), t+T(\mathbf{x})+\tau).\end{aligned}$$

But since $\mathbf{c}$ is independent of t, we have

$$\boldsymbol{\phi}(t+T(\mathbf{x}); t, \mathbf{x}) = \boldsymbol{\phi}(t+T(\mathbf{x})+\tau; t+\tau, \mathbf{x}),$$

and thus

$$\mathbf{\Phi}(\mathbf{v})(\boldsymbol{\phi}(t+T(\mathbf{x}); t, \mathbf{x}), t+T(\mathbf{x})) \neq \mathbf{\Phi}(\mathbf{v})(\boldsymbol{\phi}(t+T(\mathbf{x}); t, \mathbf{x}), t+T(\mathbf{x})+\tau).$$

But this contradicts the hypothesis that $\mathbf{\Phi}(\mathbf{v})(\bar{\mathbf{x}}, t)$ is periodic in t with period τ. Therefore we must have

$$(S_t\mathbf{v})(\mathbf{x}) = (S_{t+\tau}\mathbf{v})(\mathbf{x})$$

for all $\mathbf{x} \in \Delta - \Delta_0$ and all $t \geqq 0$.

If $\mathbf{v}_0 \in V_W$ and

$$\varepsilon < \tfrac{1}{2} \min_{\mathbf{x}\in\Delta - \mathrm{int}\,\Delta_0} d(\mathbf{v}_0(\mathbf{x}), \partial W(\mathbf{x})), \tag{4.5}$$

denote by $V_\varepsilon(\mathbf{v}_0)$ the set of functions $V_\varepsilon = \{\mathbf{v}_0 + \hat{\mathbf{u}}: |\hat{\mathbf{u}}| = 1\}$, and denote by $S_t V_\varepsilon(\mathbf{v}_0)$ the set of functions $\{S_t\mathbf{v}: \mathbf{v} \in V_\varepsilon(\mathbf{v}_0)\}$.

LEMMA 4.4. *For each $\mathbf{v}_0 \in V_W$ and ε as in* (4.5), *there is some $t_0 > 0$ such that $(S_s V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ lies strictly inside $(S_{s+t}V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ for all $s \geqq 0$, $t \geqq t_0$, $\mathbf{x} \in \Delta$.*

Proof. Each element $\mathbf{v}$ of $V_\varepsilon(\mathbf{v}_0)$ is in V_+, since $\mathbf{v}(\mathbf{x}) \neq \mathbf{u}_0(\mathbf{x})$ for $\mathbf{x}$ in Δ_0. Hence, by Theorem 2.2, $(S_t\mathbf{v})(\mathbf{x})$ approaches $\partial W(\mathbf{x})$ as $t \to \infty$ uniformly for $\mathbf{x} \in \Delta$. Thus there is some $t_0 > 0$ such that $d((S_t\mathbf{v})(\mathbf{x}), \partial W(\mathbf{x})) < \varepsilon$ for each $\mathbf{v} \in V_\varepsilon(\mathbf{v}_0)$, $\mathbf{x} \in \Delta$ and $t \geqq t_0$.

Now, by Lemma 3.3, if $t \geqq t_0$, $\mathbf{x} \in \Delta$, then $(S_t V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ encloses $\{\mathbf{u} \in W(\mathbf{x}) : d(\mathbf{u}, \partial W(\mathbf{x})) > \varepsilon\}$, which in turn contains $V_\varepsilon(\mathbf{v}_0)(\mathbf{x})$. So $V_\varepsilon(\mathbf{v}_0)(\mathbf{x})$ is enclosed by $(S_t V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ for all $\mathbf{x} \in \Delta$ and all $t \geqq t_0$. Hence, using the uniqueness of solutions (Lemma 2.1), it follows that $(S_s V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ is enclosed by $(S_{s+t} V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ for $s \geqq 0$.

Proof of Theorem 4.1. (a) There is a continuum of functions $\mathbf{g}: \partial\Delta \times [0, \infty) \to R^n$ periodic in t with period τ, such that $\mathbf{g}(\bar{\mathbf{x}}, t) \in \operatorname{int} W(\bar{\mathbf{x}})$. Now Lemmas 4.1 and 4.2 imply that each $\mathbf{g}$ is in $\Phi(V_W)$, and from Lemma 4.3 it follows that for each such $\mathbf{g}$ the associated function $\mathbf{u}(t, \mathbf{x})$ is periodic in t with period τ for $\mathbf{x} \in \Delta - \Delta_0$. For $\mathbf{x}_0 \in \Delta_0$, $\mathbf{v} \in V_W$, it follows, by the definition of $S_0(\mathbf{x}_0)$, that $(S_t\mathbf{v})(\mathbf{x}_0) = \mathbf{u}_0(\mathbf{x}_0)$ for all $t \geqq 0$. Hence there is a continuum of initial functions $\mathbf{v} \in V_W$ that give rise to τ-periodic solutions $\mathbf{u}(t, \mathbf{x})$.

(b) Choose $\mathbf{v}_0 \in V_W$, and let

$$\varepsilon < \tfrac{1}{2} \min_{\mathbf{x} \in \Delta - \operatorname{int} \Delta_0} d(\mathbf{v}_0(\mathbf{x}), \partial W(\mathbf{x})).$$

Define $V_\varepsilon(\mathbf{v}_0)$ and $S_t V_\varepsilon(\mathbf{v}_0)$ as before. Let $\mathbf{g} = \Phi(\mathbf{v}_0)$.

By Lemma 4.4 we can choose t_p such that if $t \geqq 0$ and $\mathbf{x} \in \Delta$, then $(S_{t_p+t} V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$ encloses $(S_t V_\varepsilon(\mathbf{v}_0))(\mathbf{x})$.

Now choose $\bar{\mathbf{g}}$ to be a continuous function from $\partial\Delta \times [0, \infty)$ to R^n subject to the following conditions:

1) For all $\bar{\mathbf{x}} \in \partial\Delta$ and $0 \leqq t < t_p$ let $\bar{\mathbf{g}}(\bar{\mathbf{x}}, t)$ be in the open region bounded by $(S_t V_\varepsilon(\mathbf{v}_0))(\bar{\mathbf{x}})$.

2) For all $\bar{\mathbf{x}} \in \partial\Delta$ and all $t \geqq t_p$ let $\bar{\mathbf{g}}(\bar{\mathbf{x}}, t) = \bar{\mathbf{g}}(\bar{\mathbf{x}}, t - nt_p)$, where n is a positive integer, and $0 \leqq t - nt_p < t_p$.

This is possible, since t_p was chosen such that $(S_{t_p} V_\varepsilon(\mathbf{v}_0))(\bar{\mathbf{x}})$ encloses $(S_0 V_\varepsilon(\mathbf{v}_0))(\bar{\mathbf{x}})$, which in turn encloses $\bar{\mathbf{g}}(\bar{\mathbf{x}}, 0)$. Thus $\bar{\mathbf{g}}$ is periodic in t with period t_p.

By Lemma 4.1 we have $\mathbf{g} \in \Phi(\bar{\mathbf{v}})$ for some $\bar{\mathbf{v}}$ which maps Δ to R^n. Also, since $\bar{\mathbf{g}}$ is periodic in t, and $\bar{\mathbf{g}}(\bar{\mathbf{x}}, t) \in \operatorname{int} W(\bar{\mathbf{x}})$ for all $t \geqq 0$ and all $\bar{\mathbf{x}} \in \partial\Delta$, $\bar{\mathbf{g}}$ satisfies the conditions of Lemma 4.2. Therefore $\bar{\mathbf{v}} \in V_W$.

Now $\bar{\mathbf{g}}(\bar{\mathbf{x}}, t)$ lies inside $(S_t V_\varepsilon(\mathbf{v}_0))(\bar{\mathbf{x}})$ for all $t \geqq 0$ and all $\bar{\mathbf{x}} \in \partial\Delta$. Therefore $\bar{\mathbf{v}}(\mathbf{x})$ lies inside $V_\varepsilon(\mathbf{v}_0)(\mathbf{x})$ for all $\mathbf{x}$ in $\Delta - \operatorname{int} \Delta_0$. Also, since $\mathbf{v} \in V_W$, we have

$$\bar{\mathbf{v}}(\mathbf{x}) = \mathbf{v}_0(\mathbf{x}) = \mathbf{u}_0(\mathbf{x})$$

for all $\mathbf{x} \in \Delta_0$. Hence for each $\mathbf{x} \in \Delta$ we have

$$|\bar{\mathbf{v}}(\mathbf{x}) - \mathbf{v}_0(\mathbf{x})| < \varepsilon.$$

Thus for any $\mathbf{v}_0$ in V_W there is some $\bar{\mathbf{v}}$ arbitrarily close to $\mathbf{v}_0$, such that $\{S_t\bar{\mathbf{v}}\}_{t \geqq 0}$ is periodic in t. Therefore the set of periodic points of S_t is dense in V_W.

Hence chaos of a type analogous to that of Li and Yorke [7] is present in V_W.

Acknowledgments. Hedley C. Morris acknowledges the computer and other support facilities made available to him by Claremont McKenna College whilst he was a visiting professor at Claremont Graduate School for the academic year 1983–1984.

REFERENCES

[1] J. AUSLANDER AND J. A. YORKE, *Interval maps, factors of maps and chaos*, Tôhoku Math. J. (2), 32 (1980), pp. 177–188.

[2] P. BRUNOVSKY, *Notes on chaos in the cell population partial differential equation*, Nonlinear Anal., 7 (1983), pp. 167-176.

[3] J. DUGUNDJI, *Topology*, Allyn and Bacon, Boston, 1976.

[4] U. AN DER HEIDEN AND H.-O. WALTHER, *Existence of chaos in control systems with delayed feedback*, J. Differential Equations, 47 (1983), pp. 273-295.

[5] P. K. KLOEDEN, M. A. B. DEAKIN AND Z. TIRKEL, *A precise definition of chaos*, Nature 264, 295 (1976).

[6] A. LASOTA, *Stable and chaotic solutions of a first-order partial differential equation*, Nonlinear Anal., 5 (1981), pp. 1181-1193.

[7] T.-Y. LI AND J. A. YORKE, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985-992.

[8] F. R. MAROTTO, *Snap-back repellers imply chaos in R^n*, J. Math. Anal. Appl., 63 (1978), pp. 199-223.

[9] E. OTT, *Strange attractors and chaotic motions of dynamical systems*, Rev. Modern Phys., 53 (1981), pp. 655-671.

[10] A. N. SHARKOVSKII, *Coexistence of cycles of a continuous mapping of the line onto itself* (*Russian*), Ukrain. Mat. Zh., 16 (1964), pp. 61-71.

[11] K. SHIRAIWA AND M. KURATA, *A generalization of a theorem of Marotto*, Proc. Japan Acad. Ser. A Math. Sci., 55 (1979), pp. 286-289.

[12] H.-O. WALTHER, *Homoclinic solution and chaos in* $\dot{x}=f(x(t-1))$, Nonlinear Anal., 5 (1981), pp. 775-788.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

012

ON THREE-BODY SCATTERING NEAR THRESHOLDS*

F. GESZTESY† AND G. KARNER‡

Abstract. Using exponential decay of the two-body interactions at infinity we discuss analytic expansions of two-cluster scattering operators and scattering amplitudes with respect to channel momenta around (negative) thresholds. An explicit formula for elastic scattering amplitudes in terms of the three-body resolvent is derived.

Key words. three-body scattering, scattering thresholds

AMS(MOS) subject classifications. 46, 47, 81

1. Introduction. This paper continues previous studies of threshold scattering in nonrelativistic two-body systems (cf. [1], [2]) to the three-body problem. We are especially concerned with analytic expansions of two-cluster scattering operators and scattering amplitudes with respect to channel momenta around (negative) thresholds assuming exponential falloff of the two-body interactions at infinity.

In § 2 we introduce our basic assumptions on the two-body interactions, establish the terminology used throughout this paper, and recall the direct integral decomposition of channel Hamiltonians. Faddeev's theory [14] in the Hilbert space version of Ginibre and Moulin [16] is discussed in § 3. Here our presentation essentially follows Amrein, Jauch and Sinha [3] and Amrein and Sinha [5]. Since we are particularly interested in the behaviour of the three-particle resolvent near thresholds, we use stronger decay assumptions on the two-body potentials (of the type $|x|^{-4-\varepsilon}$, $\varepsilon>0$ as $|x|\to\infty$) than in [5]. Section 4, which describes continuity and the threshold behaviour of averaged total cross sections, merely supplements the treatment in [5] (cf. their Remark 3). For two-body potentials decaying like $|x|^{-5-\varepsilon}$, $\varepsilon>0$ at infinity we represent the two-cluster scattering amplitudes as scalar products in $L^2(\mathscr{R}^6)$ (in analogy to the two-body problem) and discuss their continuity properties with respect to energy and angle variables in § 5. As a special case of our threshold considerations of two-cluster scattering amplitudes we mention the definition of elastic scattering lengths and their explicit formulas in terms of the three-body resolvent (cf. also [15]). Finally, using an exponential decay of the two-body potentials in § 6, we derive analytic expansions of two-cluster scattering operators and amplitudes with respect to channel momenta around (negative) thresholds.

2. Notations and basic facts. Greek indices α, β, γ, δ denote pairs of particles, or the corresponding third particle, i.e., (1, 2) or 3 (resp. (2, 3), or 1 and (3, 1) or 2). If all particles are considered simultaneously we denote the triple (1, 2, 3) by $\alpha=0$ ($\beta=0$, etc.). If μ_j, $j=1,2,3$ denote the masses of the spinless and distinguishable particles in $\mathscr{R}^3$, m_α denotes the reduced mass of the pair $\alpha=(j,l)$, i.e.,

$$m_\alpha^{-1}=\mu_j^{-1}+\mu_l^{-1},\qquad \alpha=(j,l) \tag{2.1}$$

and n_α denotes the reduced mass of the pair $\alpha=(j,l)$ and the corresponding third particle, i.e.,

$$n_\alpha^{-1}=(\mu_j+\mu_l)^{-1}+\mu_p^{-1},\qquad \alpha=(j,l)=p. \tag{2.2}$$

* Received by the editors April 18, 1985; accepted for publication (in revised form) May 20, 1986.

† Institut für Theoretische Physik, Universität Graz, Austria. Present address, Zentrum für interdisziplinare Forschung, Universität Bielefeld, D-4800 Bielefeld, West Germany.

‡ Fakultät für Physik, Universität Bielefeld, D-4800 Bielefeld 1, West Germany. Present address, Institut für Theoretische Physik, Universität Wien, Wien, Austria.

Finally $M=\sum_{j=1}^{3}\mu_j$ abbreviates the total mass. If $\xi_j, j=1,2,3$ denote the coordinates of the particles in configuration space the center of mass and Jacobi coordinates are defined by

$$\begin{aligned} x_s &= M^{-1}\sum_{j=1}^{3}\mu_j\xi_j,\\ x_\alpha &= \xi_j-\xi_l,\\ y_\alpha &= x_p-(\mu_j+\mu_l)^{-1}(\mu_j\xi_j+\mu_l\xi_l), \qquad \alpha=(j,l)=p. \end{aligned}\tag{2.3}$$

The associated conjugate momenta are denoted by

$$\begin{aligned} k_s &= \sum_{j=1}^{3}\eta_j,\\ k_\alpha &= (\mu_j+\mu_l)^{-1}(\mu_l\eta_j-\mu_j\eta_l),\\ q_\alpha &= M^{-1}(\mu_j+\mu_l)\eta_p-M^{-1}\mu_p(\eta_j+\eta_l), \qquad \alpha=(j,l)=p \end{aligned}\tag{2.4}$$

where η_j are conjugate momenta of $\xi_j, j=1,2,3$. For later purposes we note that

$$\begin{aligned} x_\alpha &= -x_\beta-x_\gamma,\\ y_\alpha &= -(\mu_\alpha+\mu_\gamma)\mu_\alpha^{-1}y_\beta+\varepsilon(\alpha,\beta,\gamma)(\mu_\beta+\mu_\gamma)^{-1}M(\mu_\gamma/\mu_\alpha)x_\alpha,\\ x_\alpha &= \varepsilon(\alpha,\beta,\gamma)y_\beta-(\mu_\alpha+\mu_\gamma)^{-1}\mu_\alpha x_\beta,\\ y_\alpha &= -(\mu_\beta+\mu_\gamma)^{-1}\mu_\beta y_\beta-\varepsilon(\alpha,\beta,\gamma)[(\mu_\alpha+\mu_\gamma)(\mu_\beta+\mu_\gamma)]^{-1}M\mu_\gamma x_\beta,\\ y_\alpha &= -\varepsilon(\alpha,\beta,\gamma)x_\beta-\varepsilon(\alpha,\beta,\gamma)(\mu_\beta+\mu_\gamma)^{-1}\mu_\beta x_\alpha,\\ y_\alpha &= -\varepsilon(\alpha,\beta,\gamma)(\mu_\beta+\mu_\gamma)^{-1}\mu_\gamma x_\beta+\varepsilon(\alpha,\beta,\gamma)(\mu_\beta+\mu_\gamma)^{-1}\mu_\beta x_\gamma, \end{aligned}\tag{2.5}$$

$$\alpha\neq\beta,\quad \alpha\neq\gamma,\quad \beta\neq\gamma,\quad \varepsilon(\alpha,\beta,\gamma)=\begin{cases}+1 & (\alpha,\beta,\gamma) \text{ an even permutation of } (1,2,3),\\ -1 & (\alpha,\beta,\gamma) \text{ an odd permutation of } (1,2,3).\end{cases}$$

The total kinetic energy (of the relative motion) is given by

$$(2m_\alpha)^{-1}k_\alpha^2+(2n_\alpha)^{-1}q_\alpha^2, \qquad \alpha=1,2,3.\tag{2.6}$$

The self-adjoint free Hamiltonian H_0 in $\mathscr{H}=L^2(\mathscr{R}^6)$ then reads

$$H_0=-(2m_\alpha)^{-1}\Delta_{x_\alpha}\otimes 1-1\otimes(2n_\alpha)^{-1}\Delta_{y_\alpha}, \qquad \alpha=1,2,3.\tag{2.7}$$

We assume that the particles interact via real, local translationally invariant potentials v_α obeying the following hypothesis.

Hypothesis (I). Let $\nu>0$, $u_\alpha\in L^p(\mathscr{R}^3)$ for some $p>3/2$ be real-valued, $\alpha=1,2,3$. Then v_α is defined as

$$v_\alpha(x_\alpha)=(1+|x_\alpha|)^{-2-\nu}u_\alpha(x_\alpha), \qquad \alpha=1,2,3.\tag{2.8}$$

The total Hamiltonian H in $\mathscr{H}$ is then defined as the form sum (cf. [36])

$$H=H_0\dot{+}\sum_{\alpha=1}^{3}v_\alpha\tag{2.9}$$

and the α-cluster Hamiltonian H_α in $\mathscr{H}$ reads

$$H_\alpha=h_\alpha\otimes 1-1\otimes(2n_\alpha)^{-1}\Delta_{y_\alpha}, \qquad \alpha=1,2,3\tag{2.10}$$

where h_α denotes the two-particle Hamiltonian of the subsystem α given by the form sum in $L^2(\mathscr{R}^3)$

$$(2.11)\qquad h_\alpha = h_{0,\alpha} \dotplus v_\alpha, \qquad h_{0,\alpha} = -(2m_\alpha)^{-1}\Delta_{x_\alpha}, \qquad \alpha = 1, 2, 3.$$

The corresponding resolvents are denoted by

$$(2.12)\qquad \begin{aligned} &G_0(z) = (H_0 - z)^{-1}, && z \in \mathscr{C}\backslash[0, \infty), \\ &G(z) = (H - z)^{-1}, && z \in \rho(H), \\ &G_\alpha(z) = (H_\alpha - z)^{-1}, && z \in \rho(H_\alpha), \\ &g_\alpha(z) = (h_\alpha - z)^{-1}, && z \in \rho(h_\alpha), \\ &g_{0,\alpha}(z) = (h_{0,\alpha} - z), && z \in \mathscr{C}\backslash[0, \infty), \qquad \alpha = 1, 2, 3 \end{aligned}$$

where $\rho(\cdot)$ denotes the corresponding resolvent set. Concerning the spectra of the operators involved, we remark that

$$(2.13)\qquad \begin{aligned} &\sigma_{\text{ess}}(h_\alpha) = \sigma_{\text{ess}}(h_{0,\alpha}) = \sigma_{\text{ess}}(H_0) = [0, \infty), \\ &\sigma_{\text{ess}}(H_\alpha) = [E_{0,\alpha}^{(2)}, \infty), \qquad E_{0,\alpha}^{(2)} = \inf \sigma(h_\alpha) > -\infty, \\ &\sigma_{\text{ess}}(H) = [E_0^{(2)}, \infty), \qquad E_0^{(2)} = \inf_{\alpha=1,2,3} \sigma(h_\alpha), \\ &\sigma_d(h_\alpha) \subset [E_{0,\alpha}^{(2)}, 0) \text{ (if } E_{0,\alpha}^{(2)} < 0\text{, otherwise empty) is finite,} \\ &E_0^{(3)} = \inf \sigma(H) > -\infty, \\ &\sigma_d(H) \subset [E_0^{(3)}, E_0^{(2)}) \text{ (if } E_0^{(3)} < E_0^{(2)}\text{, otherwise empty) is finite,} \qquad \alpha = 1, 2, 3 \end{aligned}$$

where $\sigma_{\text{ess}}(\cdot)$, $\sigma_d(\cdot)$ denote the corresponding essential and discrete spectra (for proofs cf. e.g. [11], [31], [39]). For additional results on the point spectrum $\sigma_p(H)$ of H see Lemma 3.8.

Next we introduce the channel structure. Let $E_{(-\infty,0)}(h_\alpha)L^2(\mathscr{R}^3) =$ lin span $\{\psi_\alpha^j | h_\alpha \psi_\alpha^j = e_\alpha^j \psi_\alpha^j, (\psi_\alpha^j, \psi_\alpha^{j'}) = \delta_{jj'}, j, j' = 1, \cdots, N_\alpha\}$, $\alpha = 1, 2, 3$ denote the negative point spectral subspace associated with h_α, where $N_\alpha < \infty$ denotes the number of negative bound states (counting multiplicity) and $e_\alpha^j < 0$ denote the corresponding eigenvalues of h_α, $e_\alpha^1 \leqq e_\alpha^2 \leqq \cdots \leqq e_\alpha^{N_\alpha} < 0$ (not necessarily distinct). The channel subspace $\mathscr{M}_\alpha^j \subset \mathscr{H}$ is then given by

$$(2.14)\qquad \mathscr{M}_\alpha^j = \{\psi_\alpha^j\} \otimes L^2(\mathscr{R}^3; d^3y_\alpha), \qquad j = 1, \cdots, N_\alpha, \qquad \alpha = 1, 2, 3$$

($\psi_\alpha^j = \psi_\alpha^j(x_\alpha)$) and the cluster subspace $\mathscr{M}_\alpha \subset \mathscr{H}$ reads

$$(2.15)\qquad \mathscr{M}_\alpha = \bigoplus_{j=1}^{N_\alpha} \mathscr{M}_\alpha^j = E_{(-\infty,0)}(h_\alpha)L^2(\mathscr{R}^3, d^3x_\alpha) \otimes L^2(\mathscr{R}^3, d^3y_\alpha), \qquad \alpha = 1, 2, 3.$$

The corresponding orthogonal projections onto $\mathscr{M}_\alpha^j$ and $\mathscr{M}_\alpha$ are denoted by E_α^j and E_α:

$$(2.16)\qquad \begin{aligned} &E_\alpha^j = (\psi_\alpha^j, \cdot)\psi_\alpha^j \otimes 1, \qquad E_\alpha = \bigoplus_{j=1}^{N_\alpha} E_\alpha^j, \\ &E_\alpha^j E_\alpha^l = \delta_{jl} E_\alpha^j, \qquad j = 1, \cdots, N_\alpha, \qquad \alpha = 1, 2, 3. \end{aligned}$$

The (α, j)-channel Hamiltonian H_α^j in $\mathscr{H}$ is then defined by

$$(2.17)\qquad H_\alpha^j = H_\alpha E_\alpha^j = \psi_\alpha^j, \cdot)\psi_\alpha^j \otimes [-(2n_\alpha)^{-1}\Delta_{y_\alpha} + e_\alpha^j], \quad j = 1, \cdots, N_\alpha, \quad \alpha = 1, 2, 3.$$

Clearly

$$\sigma_p(H_\alpha^j)=\varnothing, \qquad \sigma_{\text{ess}}(H_\alpha^j)=\sigma_{ac}(H_\alpha^j)=[e_\alpha^j,\infty), \qquad j=1,\cdots,N_\alpha, \quad \alpha=1,2,3. \tag{2.18}$$

($\sigma_{ac}(\cdot)$ denotes the absolutely continuous spectrum.) Finally we introduce the zeroth channel $(0,0)$. Define

$$e_0^0=0, \qquad \mathscr{M}_0^0=\mathscr{H}, \qquad E_0^0=1; \tag{2.19}$$

then $H_0^0=H_0$ is the corresponding cluster (channel) Hamiltonian.

In order to discuss scattering theory, we introduce

Hypothesis (II). Suppose H(I) holds and define

$$\varepsilon_\alpha^\pm=\{z\in\mathscr{C}\,|\,v_\alpha^{1/2}g_{0,\alpha}(z\pm i0)|v_\alpha|^{1/2}\Phi=-\Phi \text{ for some } \Phi\in L^2(\mathscr{R}^3),\ \Phi\neq 0\} \tag{2.20}$$

where

$$v_\alpha^{1/2}(x_\alpha)=\operatorname{sgn}(v_\alpha(x_\alpha))|v_\alpha(x_\alpha)|^{1/2}, \qquad \alpha=1,2,3. \tag{2.21}$$

Then we assume that

$$\varepsilon_\alpha=\{\varepsilon_\alpha^+\cup\varepsilon_\alpha^-\}\cap[0,\infty)=\varnothing, \qquad \alpha=1,2,3. \tag{2.22}$$

Obviously (2.22) implies the absence of zero-energy resonances and nonnegative (embedded) eigenvalues as well as

$$\sigma_{\text{ess}}(h_\alpha)=\sigma_{ac}(h_\alpha)=[0,\infty), \qquad \alpha=1,2,3 \tag{2.23}$$

(cf. [31], [36]). H(II) implies the existence (cf. [30], [37]) and completeness (cf. [12], [16], [17], [20], [21], [25], [26], [30], [34], [35], [38], [40] and [41]) of the wave operators $\Omega_\pm^{\alpha j}$ in $\mathscr{H}$ defined by

$$\Omega_\pm^{\alpha j}=s-\lim_{t\to\pm\infty} e^{itH}e^{-itH_\alpha^j}E_\alpha^j, \qquad \alpha=0,1,2,3. \tag{2.24}$$

More precisely we have

$$\begin{gathered}\operatorname{Ran}\Omega_\pm^{00}\oplus\bigoplus_{\alpha=1}^{3}\bigoplus_{j=1}^{N_\alpha}\operatorname{Ran}\Omega_\pm^{\alpha j}=\mathscr{H}_{ac}(H),\\ (\Omega_\pm^{\beta l})^*\Omega_\pm^{\alpha j}=\delta_{\beta\alpha}^{lj}E_\alpha^j, \qquad \alpha=0,1,2,3.\end{gathered} \tag{2.25}$$

Here and in the following sections we always assume that in a two-cluster channel (α,j), $\alpha\neq 0$, j runs from 1 to N_α whereas for a three-cluster channel (β,l), $\beta=0$, l equals zero. For weaker two-body spectral assumptions cf. [12], [24], [34], [38]. The absence of the singular continuous spectrum of H follows from [26], [29]. The corresponding scattering operator $S_{\beta\alpha}^{lj}$ in $\mathscr{H}$, defined by

$$S_{\beta\alpha}^{lj}=(\Omega_+^{\beta l})^*\Omega_-^{\alpha j}, \qquad \alpha,\beta=0,1,2,3, \tag{2.26}$$

describes scattering from channel (α,j) into channel (β,l): $S_{\alpha\alpha}^{jj}$ describes elastic scattering, $S_{\alpha\alpha}^{lj}$, $l\neq j$ (i.e. $\psi_\alpha^j\neq$ const. ψ_α^l) inelastic scattering, $S_{\beta\alpha}^{lj}$, $\alpha\neq\beta$ rearrangement scattering, $S_{0\alpha}^{0j}$ breakup scattering, S_{00}^{00} the scattering of three free particles. We recall that

$$H_\beta^l S_{\beta\alpha}^{lj}=S_{\beta\alpha}^{lj}H_\alpha^j, \qquad \alpha,\beta=0,1,2,3. \tag{2.27}$$

Clearly $S_{\beta\alpha}^{lj}$ maps $\mathscr{M}_\alpha^j$ onto $\mathscr{M}_\beta^l$ and unitary of the scattering operator is equivalent to

$$S_{\beta 0}^{l0}(S_{\beta' 0}^{l'0})^*+\sum_{\alpha=1}^{3}\sum_{j=1}^{N_\alpha}S_{\beta\alpha}^{lj}(S_{\beta'\alpha}^{l'j})^*=\delta_{\beta\beta'}^{ll'}E_\beta^l, \qquad \beta,\beta'=0,1,2,3. \tag{2.28}$$

Each eigenvalue $e^j_\alpha < 0$ of h_α together with $e^0_0 = 0$ defines a threshold for the three-body system. We introduce

$$(2.29)\qquad \theta_\alpha = \{e^j_\alpha < 0, j = 1, \cdots, N_\alpha\}, \qquad \alpha = 1, 2, 3, \qquad \theta = \{0\} \cup \bigcup_{\alpha=1}^{3} \theta_\alpha.$$

Finally we turn to spectral representations associated with H^j_α, H_0. Define

$$(2.30)\qquad \lambda = (2n_\alpha)^{-1}(q^j_\alpha)^2 + e^j_\alpha, \qquad \omega^j_\alpha = |q^j_\alpha|^{-1} q^j_\alpha \in S^2, \qquad \alpha = 1, 2, 3,$$

S^2 the unit sphere in $\mathscr{R}^3$. The spectral representation (cf. [3]) of H^j_α then reads

$$(2.31)\qquad \mathscr{G}^j_\alpha = L^2((e^j_\alpha, \infty); L^2(S^2))$$

and the corresponding spectral transformation $\mathscr{U}^j_\alpha$ is given by

$$\mathscr{U}^j_\alpha : \mathscr{M}^j_\alpha \to \mathscr{G}^j_\alpha,$$

$$(2.32)\qquad \begin{aligned} \mathscr{U}^j_\alpha(\psi^j_\alpha \otimes \phi)(\lambda, \omega^j_\alpha) &= (2\pi)^{-3/2} n_\alpha^{1/2} [2n_\alpha(\lambda - e^j_\alpha)]^{1/4} \\ &\quad \cdot \operatorname*{l.i.m.}_{R\to\infty} \int_{|y_\alpha| \leqq R} d^3 y_\alpha \exp(-i[2n_\alpha(\lambda - e^j_\alpha)]^{1/2} \omega^j_\alpha y_\alpha) \phi(y_\alpha), \\ &\qquad \phi \in L^2(\mathscr{R}^3), \qquad \alpha = 1, 2, 3. \end{aligned}$$

As a consequence

$$(2.33)\qquad \mathscr{U}^j_\alpha H^j_\alpha \mathscr{U}^{j*}_\alpha = \int^{\oplus}_{(e^j_\alpha, \infty)} d\lambda \lambda 1_{L^2(S^2)}, \qquad \alpha = 1, 2, 3.$$

In addition we define

$$(2.34)\qquad U^j_\alpha = \mathscr{U}^j_\alpha E^j_\alpha, \qquad \alpha = 1, 2, 3$$

and (cf. [3], [5]) $M^j_\alpha(A, \lambda): \mathscr{D}(A) \to L^2(S^2)$,

$$(2.35)\qquad (M^j_\alpha(A, \lambda) g)(\omega^j_\alpha) = (U^j_\alpha E^j_\alpha A g)(\lambda, \omega^j_\alpha), \qquad g \in \mathscr{D}(A), \quad \alpha = 1, 2, 3.$$

Here A denotes a multiplication operator in $\mathscr{H}$ with one of the following functions (cf. (3.4) and (3.5)): $|v_\beta(x_\beta)|^{1/2}$, $v_\beta^{1/2}(x_\beta)$, ${}^{(\hat{}\,)}\rho^j_\alpha(y_\alpha)$, $v_\beta(x_\beta)|{}^{(\hat{}\,)}\rho^j_\beta(y_\beta)|^{-1}$, $v_\beta^{1/2}(x_\beta)|{}^{(\hat{}\,)}\rho^j_\beta(y_\beta)|^{-1}$, $g_\kappa(x_\alpha, x_\beta)$ where $|g_\kappa(x_\alpha, x_\beta)| \leqq \exp[\kappa|x_\alpha|]\phi(x_\beta)$ for some $\kappa < (-2m_\alpha e^j_\alpha)^{1/2}$, $\phi \in L^2(\mathscr{R}^3)$, $\beta = 1, 2, 3$, $\beta \neq \alpha$.

It remains to discuss H_0. Let

$$(2.36)\qquad \begin{aligned} &\lambda = (2m_\gamma)^{-1}|k_\gamma|^2 + (2n_\gamma)^{-1}|q_\gamma|^2, \\ &\omega = |k_\gamma|^{-1} k_\gamma \in S^2, \qquad q = [m_\gamma^{-1} n_\gamma |k_\gamma|^2 + |q_\gamma|^2]^{-1/2} q_\gamma, \qquad \gamma = 1, 2, 3. \end{aligned}$$

Then $\omega^0_0 = (\omega, q)$ parametrizes the ellipsoid E^5

$$(2.37)\qquad E^5 = \{k_\gamma, q_\gamma | (2m_\gamma)^{-1}|k_\gamma|^2 + (2n_\gamma)^{-1}|q_\gamma|^2 = 1\}, \qquad \gamma = 1, 2, 3.$$

The spectral representation of H_0 is

$$(2.38)\qquad \mathscr{G}^0_0 = L^2((0, \infty); L^2(E^5))$$

and the corresponding spectral transformation U_0^0 reads

$$U_0^0: \mathscr{H} \to \mathscr{G}_0^0,$$

$$(U_0^0\phi)(\lambda, \omega, q) = (2\pi)^{-3}[4(n_\gamma m_\gamma)^{3/2}(1-|q|^2)^{1/2}\lambda^2]^{1/2}$$

$$\cdot \operatorname*{l.i.m.}_{R\to\infty} \int_{|x_\gamma|\le R, |y_\gamma|\le R} d^3x_\gamma\, d^3y_\gamma \tag{2.39}$$

$$\cdot \exp(-i[2m_\gamma\lambda(1-|q|^2)]^{1/2}\omega x_\gamma - i(2n_\gamma\lambda)^{1/2}qy_\gamma)\phi(x_\gamma, y_\gamma),$$

$$\phi \in \mathscr{H}, \quad \gamma = 1, 2, 3.$$

Consequently

$$U_0^0 H_0(U_0^0)^* = \int_{(0,\infty)}^{\oplus} d\lambda\, \lambda 1_{L^2(E^5)}. \tag{2.40}$$

Similarly to (2.35) we also define $M_0^0(A, \lambda): \mathscr{D}(A) \to L^2(E^5)$:

$$(M_0^0(A, \lambda)g)(\omega, q) = (U_0^0 Ag)(\lambda, \omega, q), \qquad g \in \mathscr{D}(A) \tag{2.41}$$

where now A denotes a multiplication operator in $\mathscr{H}$ with one of the following functions: $|v_\beta(x_\beta)|^{1/2}$, $\phi(x_\beta)$, where $\phi \in L^p(\mathscr{R}^3)$ for $p=2$ or $p=3$, $\beta = 1, 2, 3$.

3. The three-particle resolvent $G(z)$. In this section we shall discuss the three-particle resolvent $G(z)$. We closely follow [5] (see also [3], [1]). Due to the stronger decay assumptions in H(II) (and in H(III) defined below) when compared to [5] we extend their results to be valid near thresholds contained in θ (cf. (2.29)).

We alternatively use H(II) and the following hypothesis.

Hypothesis (III). Let $b>0$ and $u_\alpha \in L^p(\mathscr{R}^3) + L^\infty(\mathscr{R}^3)$ for some $p > 3/2$ be real-valued, $\alpha = 1, 2, 3$. Then v_α is defined by

$$v_\alpha(x_\alpha) = e^{-b|x_\alpha|}u_\alpha(x_\alpha), \qquad \alpha = 1, 2, 3. \tag{3.1}$$

In addition we assume that

$$\varepsilon_\alpha = \{\varepsilon_\alpha^+ \cup \varepsilon_\alpha^-\} \cap [0, \infty) = \varnothing, \qquad \alpha = 1, 2, 3 \tag{3.2}$$

(cf. (2.20)).

Obviously H(III) implies H(II) and both H(II) and H(III) imply $v_\alpha \in L^1(\mathscr{R}^3) \cap L^{3/2}(\mathscr{R}^3)$, $\alpha = 1, 2, 3$. The exponential decreases of v_α at infinity assumed in H(III) will be used in § 6 to derive analytic expansions of scattering operators and amplitudes with respect to channel momenta near thresholds in $\theta\backslash\{0\}$.

LEMMA 3.1. *Assume* H(II) *and* $h_\alpha\psi_\alpha^j = e_\alpha^j\psi_\alpha^j$, $e_\alpha^j < 0$, $\psi_\alpha^j \in L^2(\mathscr{R}^3)$ *for some* $j = 1, \cdots, N_\alpha$, $\alpha = 1, 2, 3$. *Then* $e^{\kappa|\cdot|}\psi_\alpha^j \in L^2(\mathscr{R}^3)$ *for all* $\kappa < (-2m_\alpha e_\alpha^j)^{1/2}$ *and* $(1+|\cdot|^\beta)\psi_\alpha^j \in H^{2,1}(\mathscr{R}^3)$, $\beta \in [0, 2+\nu)$. *If in addition* H(III) *holds then* $e^{\kappa|\cdot|}\psi_\alpha^j \in H^{2,1}(\mathscr{R}^3)$ *for all* $\kappa < (-2m_\alpha e_\alpha^j)^{1/2}$.

Proof. Using H(II) the result is due to [36] and [5]. Suppose H(III) then $e^{\kappa|\cdot|}\psi_\alpha^j$, $|v_\alpha|^{1/2}e^{\kappa|\cdot|}\psi_\alpha^j \in L^2(\mathscr{R}^3)$ for all $\kappa < \kappa_0 \equiv (-2m_\alpha e_\alpha^j)^{1/2}$ by Theorems VI.6 and VI.7 of [36]. From

$$\psi_\alpha^j(x) = -(4\pi)^{-1}\int_{\mathscr{R}^3} d^3y\, e^{-\kappa_0|x-y|}|x-y|^{-1}v_\alpha(y)\psi_\alpha^j(y)$$

we infer

$$\begin{aligned}|(\nabla\psi_\alpha^j)(x)| &\leqq (4\pi)^{-1}\kappa_0\int_{\mathscr{R}^3} d^3y\, e^{-\kappa_0|x-y|}|x-y|^{-1}|v_\alpha(y)||\psi_\alpha^j(y)| \\ &\quad +(4\pi)^{-1}\int_{\mathscr{R}^3} d^3y\, e^{-\kappa_0|x-y|}|x-y|^{-2}|v_\alpha(y)||\psi_\alpha^j(y)| \\ &\equiv |(\nabla\psi_\alpha^j)_1(x)|+|(\nabla\psi_\alpha^j)_2(x)|.\end{aligned}$$

Clearly $e^{\kappa|\cdot|}|(\nabla\psi_\alpha^j)_1|\in L^2(\mathscr{R}^3)$ for all $\kappa<\kappa_0$ by the above remarks. Next we get

$$\begin{aligned}\|e^{\kappa|\cdot|}|(\nabla\psi_\alpha^j)_2|\,\|_2^2 &\leqq (4\pi)^{-2}\int_{\mathscr{R}^6} d^3y\, d^3z\, e^{(\kappa_0-\kappa)|y|}|v_\alpha(y)|\, e^{(\kappa_0-\kappa)|z|}|v_\alpha(z)| \\ &\quad \cdot e^{\kappa|y|}|\psi_\alpha^j(y)|e^{\kappa|z|}|\psi_\alpha^j(z)| \\ &\quad \cdot\int_{\mathscr{R}^3} d^3x|x-y|^{-2}|x-z|^{-2}\, e^{\kappa|x|}\, e^{-\kappa_0|x-y|}\, e^{-\kappa_0|y|}\, e^{\kappa|x|}\, e^{-\kappa_0|x-z|}\, e^{-\kappa_0|z|} \\ &\leqq \text{const.}\, \|e^{(\kappa_0-\kappa)|\cdot|}v_\alpha\|_R\|\,|v_\alpha|^{1/2}\, e^{\kappa|\cdot|}\psi_\alpha^j\|_2^2<\infty \quad \text{for } 0<\kappa_0-\kappa<b\end{aligned}$$

($\|\cdot\|_R$ the Rollnik norm [36]) using Fubini's Theorem,

$$\int_{\mathscr{R}^3} d^3x|x-y|^{-2}|x-z|^{-2}\leqq \text{const.}\,|y-z|^{-1}, \tag{3.3}$$

and the Schwarz' inequality. Thus $e^{\kappa|\cdot|}|(\nabla\psi_\alpha^j)_2|\in L^2(\mathscr{R}^3)$ for all $\kappa<\kappa_0$ and we only need to note

$$\nabla(e^{\kappa|\cdot|}\psi_\alpha^j)(x)=\kappa\, e^{\kappa|x|}(|x|^{-1}x)\psi_\alpha^j(x)+e^{\kappa|x|}(\nabla\psi_\alpha^j)(x). \qquad \square$$

Next we introduce

$$\begin{aligned}&\rho_\alpha(y_\alpha)=\begin{cases}(1+|y_\alpha|)^{-1-\nu/2} & \text{if H(II) holds,}\\ e^{-a_\alpha|y_\alpha|} & \text{if H(III) holds,}\end{cases} \\ &\qquad 0<a_\alpha<\inf\left\{\frac{b}{2},\left(\frac{\mu_\alpha}{M}\right)\inf_{\beta=1,2,3}[-2m_\beta e_\beta^{N_\beta}]^{1/2}\right\},\end{aligned} \tag{3.4}$$

$$\begin{aligned}&\hat\rho_\alpha(y_\alpha)=\begin{cases}(1+|y_\alpha|)^{-1-\nu/2} & \text{if H(II) holds,}\\ e^{-\hat a|y_\alpha|} & \text{if H(III) holds,}\end{cases} \\ &\qquad 0<\hat a<\inf_{\alpha=1,2,3}\left(\frac{\mu_\alpha}{M}\right)\inf_{\beta=1,2,3}(a_\beta),\end{aligned} \tag{3.5}$$

$$\sigma_\alpha(x_\alpha)=\begin{cases}(1+|x_\alpha|)^{-1-\nu/2} & \text{if H(II) holds,}\\ e^{-a_\alpha|x_\alpha|} & \text{if H(III) holds,}\end{cases} \tag{3.6}$$

$$\hat\sigma_\alpha(x_\alpha)=\begin{cases}(1+|x_\alpha|)^{-1-\nu/2} & \text{if H(II) holds,}\\ e^{-\hat a|x_\alpha|} & \text{if H(III) holds;} \quad \alpha=1,2,3\end{cases} \tag{3.7}$$

and denote by $\rho_\alpha(\hat\rho_\alpha)$ resp. $\sigma_\alpha(\hat\sigma_\alpha)$ the multiplication operators in $\mathscr{H}$

$$(\overset{(\wedge)}{\rho}_\alpha\psi)(x_\alpha,y_\alpha)=\overset{(\wedge)}{\rho}_\alpha(y_\alpha)\psi(x_\alpha,y_\alpha), \qquad (\overset{(\wedge)}{\sigma}_\alpha\psi)(x_\alpha,y_\alpha)=\overset{(\wedge)}{\sigma}_\alpha(x_\alpha)\psi(x_\alpha,y_\alpha),$$

$$\psi\in\mathscr{H}, \quad \alpha=1,2,3.$$

LEMMA 3.2. *Assume* H(II) *or* H(III). *Then the following operators are in* $\mathscr{B}(\mathscr{H})$:

(i) $|u_\alpha|^{1/2}(|\nabla_{x_\beta}|+1)^{-1}, |v_\alpha|^{1/2}(|\nabla_{x_\beta}|+1)^{-1}, |v_\alpha|^{1/2}(H_0+1)^{-1/2},$
$|u_\alpha|^{1/2}E_\beta, |v_\alpha|^{1/2}E_\beta, \quad \alpha, \beta = 1, 2, 3.$

(ii) $\rho_\alpha^{-1}\rho_\beta^{-1}\, e^{-b|x_\delta|}|u_\delta|^{1/2}E_\gamma, \quad \gamma \neq \delta, \quad \rho_\gamma^{-2}E_\gamma\, e^{-b|x_\delta|}|u_\delta|^{1/2}, \quad \gamma \neq \delta,$
$\rho_\gamma^{-1}E_\gamma\, e^{-b|x_\delta|}|u_\delta|^{1/2}\rho_\delta^{-1}, \quad \gamma \neq \delta, \qquad \alpha, \beta, \gamma, \delta = 1, 2, 3$ *if* H(III) *holds.*

If only H(II) *holds replace* $e^{-b|x_\delta|}$ *by* $(1+|x_\delta|)^{-2-\nu}$.

(iii) $|v_\gamma|^{1/2}E_\delta\rho_\delta^{-1}, \quad \gamma \neq \delta, \quad \hat{\rho}_\alpha^{-1}E_\alpha\rho_\beta, \qquad \alpha, \beta, \gamma, \delta = 1, 2, 3.$

(iv) $\sigma_\alpha^{-1}E_\alpha, |v_\alpha|^{1/2}E_\alpha\sigma_\alpha^{-1}, \qquad \alpha = 1, 2, 3.$

(v) $\hat{\sigma}_\beta^{-1}\hat{\rho}_\alpha E_\alpha, \qquad \alpha, \beta = 1, 2, 3.$

Proof. See [5] for the proof under H(II). If H(III) holds one can follow their methods using the second part in Lemma 3.1. □

Introducing for $z \in \mathscr{C}\backslash\mathscr{R}$,

$$
\begin{aligned}
&W_{\gamma\delta}(z) = v_\gamma^{1/2}G_0(z)|v_\delta|^{1/2}, \qquad Z_{\gamma\delta}(z) = v_\gamma^{1/2}G(z)|v_\delta|^{1/2},\\
&X_{\gamma\delta}(z) = v_\gamma^{1/2}G_\gamma(z)|v_\delta|^{1/2}, \qquad \mathbf{X}_{\gamma\delta}(z) = X_{\gamma\gamma}(z)\delta_{\gamma\delta},\\
&V_{\gamma\delta}(z) = v_\gamma^{1/2}G_\gamma(z)\rho_\delta, \qquad Y_{\gamma\delta}(z) = v_\gamma^{1/2}G(z)\rho_\delta,\\
&K_{0,\gamma\delta}(z) = v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)\rho_\delta, \qquad K_{1,\gamma\delta} = \hat{\rho}_\gamma^{-1}E_\gamma\rho_\delta,\\
&L_{\gamma\delta}(z) = v_\gamma^{1/2}E_\gamma G_\gamma(z)\hat{\rho}_\gamma\delta_{\gamma\delta},\\
&D_{00,\gamma\delta}(z) = v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)|v_\delta|^{1/2}(1-\delta_{\gamma\delta}),\\
&D_{01,\gamma\delta}(z) = \sum_{\alpha=1}^{3} D_{00,\gamma\alpha}(z)L_{\alpha\delta}(z)\\
&\qquad = v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)|v_\delta|^{1/2}(1-\delta_{\gamma\delta})v_\delta^{1/2}E_\delta G_\delta(z)\hat{\rho}_\delta,\\
&D_{10,\gamma\delta} = \hat{\rho}_\gamma^{-1}E_\gamma|v_\delta|^{1/2}(1-\delta_{\gamma\delta}),\\
&D_{11,\gamma\delta}(z) = \sum_{\alpha=1}^{3} D_{10,\gamma\alpha}L_{\alpha\delta}(z)\\
&\qquad = \hat{\rho}_\gamma^{-1}E_\gamma|v_\delta|^{1/2}(1-\delta_{\gamma\delta})v_\delta^{1/2}E_\delta G_\delta(z)\hat{\rho}_\delta, \qquad \gamma, \delta = 1, 2, 3.
\end{aligned}
\tag{3.8}
$$

We recall the following lemma ([5]).

LEMMA 3.3. *Assume* H(II) *and* $z \in \mathscr{C}\backslash\mathscr{R}$. *Then all operators defined in* (3.8) *are in* $\mathscr{B}(\mathscr{H})$.

LEMMA 3.4. *Assume* H(II). *Then*

(i) $W_{\gamma\delta}(z)$, $\gamma, \delta = 1, 2, 3$ *converges in norm as* $z \to \lambda \pm i0$, *uniformly in* $\lambda \in \mathscr{R}$. *For* $\gamma \neq \delta$, $W_{\gamma\delta}(z) \in \mathscr{B}_\infty(\mathscr{H})$ *for all* $z \in \mathscr{C}$, $W_{\gamma\gamma}(z) \notin \mathscr{B}_\infty(\mathscr{H})$, $z \in \mathscr{C}$.

(ii) $\rho_\alpha E_\alpha G_\alpha(z)\rho_\alpha$ *converges in norm as* $z \to \lambda \pm i0$, *uniformly in* $\lambda \in \mathscr{R}$, $\rho_\alpha E_\alpha G_\alpha(z)\rho_\alpha \in \mathscr{B}_\infty(\mathscr{H})$ *for all* $z \in \mathscr{C}$, $\alpha = 1, 2, 3$.

(iii) $v_\alpha^{1/2}(1-E_\alpha)G_\alpha(z)|v_\alpha|^{1/2}$ *converges in norm as* $z \to \lambda \pm i0$, *uniformly in* $\lambda \in \mathscr{R}$, $\alpha = 1, 2, 3$.

(iv) $v_\gamma^{1/2}G_\gamma(z)|v_\delta|^{1/2} \in \mathscr{B}_\infty(\mathscr{H})$ *for* $\gamma \neq \delta$, $\operatorname{Im} z \neq 0$, $\gamma, \delta = 1, 2, 3$.

Proof. (i), (iii) and (iv) are due to [5] and based on [22]. Due to our stronger decay properties in ρ_α we infer (ii) for all $\lambda \in \mathscr{R}$ from

$$
\rho_\alpha E_\alpha G_\alpha(z)\rho_\alpha = \bigoplus_{j=1}^{N_\alpha} E_\alpha^j \otimes \rho_\alpha[-(2n_\alpha)^{-1}\Delta_{y_\alpha} - (z - e_\alpha^j)]^{-1}\rho_\alpha
\tag{3.9}
$$

since $N_\alpha < \infty$ and $\rho_\alpha \in L^{3/2}(\mathcal{R}^3)$. We also note that (iii) immediately follows from

$$v_\alpha^{1/2}(1-E_\alpha)G_\alpha(z)|v_\alpha|^{1/2} = i\int_0^\infty d\tau\, e^{iz\tau} v_\alpha^{1/2} E_{ac}(h_\alpha)\, e^{-i\tau h_\alpha}|v_\alpha|^{1/2}$$
$$\otimes \exp(-i\tau(-2n_\alpha)^{-1}\Delta_{y_\alpha}), \qquad \operatorname{Im} z > 0$$

(similar for $\operatorname{Im} z < 0$) and from ([20])

$$\|v_\alpha^{1/2} E_{ac}(h_\alpha)\, e^{-i\tau h_\alpha}|v_\alpha|^{1/2}\| \in L^1(\mathcal{R}, d\tau). \qquad \square$$

For the following it is convenient to define

$$\bar{\mathcal{H}} = \mathcal{H} \oplus \mathcal{H} \oplus \mathcal{H} \tag{3.10}$$

and to represent linear operators $\bar{T}: \bar{\mathcal{H}} \to \bar{\mathcal{H}}$ by $\bar{T} = \{T_{\alpha\beta}\}_{\alpha,\beta=1,2,3}$. Then we have

$$\begin{aligned} \bar{Z}(z) &= \bar{W}(z) - \bar{W}(z)\bar{Z}(z), \\ \bar{Z}(z) &= \bar{X}(z') - [\bar{X}(z) - \bar{\mathbf{X}}(z)]\bar{Z}(z), \\ \bar{Y}(z) &= \bar{V}(z) - [\bar{X}(z) - \bar{\mathbf{X}}(z)]\bar{Y}(z), \qquad \operatorname{Im} z \neq 0. \end{aligned} \tag{3.11}$$

In addition we define

$$\overset{\Delta}{\mathcal{H}} = \bar{\mathcal{H}} \oplus \bar{\mathcal{H}} \tag{3.12}$$

and represent linear operators $\overset{\Delta}{U}: \overset{\Delta}{\mathcal{H}} \to \overset{\Delta}{\mathcal{H}}$ by $\overset{\Delta}{U} = \{\bar{U}_{st}\}_{s,t=0,1}$ such that

$$\overset{\Delta}{U}g = \begin{pmatrix} \bar{U}_{00} & \bar{U}_{01} \\ \bar{U}_{10} & \bar{U}_{11} \end{pmatrix}\begin{pmatrix} g_0 \\ g_1 \end{pmatrix} = \begin{pmatrix} \bar{U}_{00}g_0 + \bar{U}_{01}g_1 \\ \bar{U}_{10}g_0 + \bar{U}_{11}g_1 \end{pmatrix},$$
$$g = \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} \in \mathcal{D}(\overset{\Delta}{U}), \qquad g_0, g_1 \in \bar{\mathcal{H}}.$$

For $\operatorname{Im} z \neq 0$ we introduce

$$\begin{aligned} &K(z): \begin{cases} \bar{\mathcal{H}} \to \overset{\Delta}{\mathcal{H}}, \\ g \to K(z)g = \begin{pmatrix} \bar{K}_0(z)g \\ \bar{K}_1 g \end{pmatrix}, \end{cases} \\ &J(z): \begin{cases} \overset{\Delta}{\mathcal{H}} \to \bar{\mathcal{H}}, \\ \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} \to g_0 + \bar{L}(z)g_1, \end{cases} \end{aligned} \tag{3.13}$$

with $K_{0,\gamma\delta}(z)$, $K_{1,\gamma\delta}$, $L_{\gamma\delta}(z)$ given by (3.8) and

$$\overset{\Delta}{D}(z) = \{\bar{D}_{st}(z)\}_{s,t=0,1} \tag{3.14}$$

with $D_{st,\gamma\delta}(z)$ defined in (3.8).

Then we have the following lemma ([5]).

LEMMA 3.5. *Assume* H(II) *or* H(III) *and* $\operatorname{Im} z \neq 0$. *Then*

(i) $J(z)K(z) = \bar{V}(z)$,

(ii) $J(z)\overset{\Delta}{D}(z) = [\bar{X}(z) - \bar{\mathbf{X}}(z)]J(z)$,

(iii) $[1 + \bar{W}(z)][1 - \bar{Z}(z)] = [1 - \bar{Z}(z)][1 + \bar{W}(z)] = 1$,
i.e., $[1 + \bar{W}(z)]^{-1} \in \mathcal{B}(\bar{H})$,

(iv) $[1+\bar{X}(z)-\bar{\mathbf{X}}(z)]^{-1}\in\mathcal{B}(\bar{\mathcal{H}})$,

(v) $[1+\overset{\Delta}{D}(z)]^{-1}$ *exists*,

$$\bar{Y}(z)=[1+\bar{X}(z)-\bar{\mathbf{X}}(z)]^{-1}J(z)K(z)=J(z)[1+\overset{\Delta}{D}(z)]^{-1}K(z). \tag{3.15)(vi}$$

Since $\bar{D}_{10}\in\mathcal{B}(\bar{\mathcal{H}})$ but $\bar{D}_{10}\notin\mathcal{B}_\infty(\bar{\mathcal{H}})$ we define in $\overset{\Delta}{\mathcal{H}}$ (cf. [3], [5])

$$\overset{\Delta}{N}=\begin{pmatrix}0 & 0\\ \bar{D}_{10} & 0\end{pmatrix},\qquad \overset{\Delta}{A}(z)=(1-\overset{\Delta}{N})[\overset{\Delta}{D}(z)-\overset{\Delta}{N}],\qquad \operatorname{Im} z\neq 0. \tag{3.16}$$

We recall the following lemma ([3], [5]).

LEMMA 3.6. *Assume* H(II) *or* H(III) *and* $\operatorname{Im} z\neq 0$. *Then*

(i) $\overset{\Delta}{N}$ *is nilpotent*, $\overset{\Delta}{N}{}^2=0$, $(1+\overset{\Delta}{N})^{-1}=1-\overset{\Delta}{N}$.

(ii) $1+\overset{\Delta}{A}(z)$ *is invertible if and only if* $1+\overset{\Delta}{D}(z)$ *is since*

$$[1+\overset{\Delta}{A}(z)]^{-1}=[1+\overset{\Delta}{D}(z)]^{-1}(1+\overset{\Delta}{N}),\qquad [1+\overset{\Delta}{D}(z)]^{-1}=[1+\overset{\Delta}{A}(z)]^{-1}(1-\overset{\Delta}{N}).$$

LEMMA 3.7. *Assume* H(II) *or* H(III). *Then*

(i) $\overset{\Delta}{A}(z)\in\mathcal{B}_\infty(\overset{\Delta}{\mathcal{H}})$ *for all* $z\in\mathcal{C}$;

(ii) $\overset{\Delta}{A}(z)$ *converges in norm as* $z\to\lambda\pm i0$, *uniformly in* $\lambda\in\mathcal{R}$ *and*

$$\lim_{|z|\to\infty}\|\overset{\Delta}{A}(z)\|=0,\qquad z\in\mathcal{C}.$$

(iii) $\overset{\Delta}{A}(z)$ *is analytic in* $z\in\mathcal{C}\backslash[E_0^{(2)},\infty)$, $E_0^{(2)}=\inf_{\alpha=1,2,3}\sigma(h_\alpha)$.

Proof. Following [5], we treat $\bar{D}_{00}(z)$, $\bar{D}_{01}(z)$, $\bar{D}_{11}(z)$ in some detail for later purposes (cf., e.g., Lemma 6.2).

$$\begin{aligned}\text{(a)}\quad D_{00,\gamma\delta}(z)&=v_\gamma^{1/2}(1-E_\gamma)[G_0(z)-G_\gamma(z)v_\gamma G_0(z)]|v_\delta|^{1/2}\\ &=[1-v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)|v_\gamma|^{1/2}]v_\gamma^{1/2}G_0(z)|v_\delta|^{1/2}\\ &\quad-[v_\gamma^{1/2}E_\gamma\sigma_\gamma^{-1}]\sigma_\gamma G_0(z)|v_\delta|^{1/2},\qquad \gamma\neq\delta.\end{aligned} \tag{3.17}$$

Lemma 3.4(iii), respectively, Lemma 3.2(iv) apply for the terms $[\cdots]$ in (3.17). Lemma 3.4(i) applies for $v_\gamma^{1/2}G_0(z)|v_\delta|^{1/2}$ and $\sigma_\gamma G_0(z)|v_\delta|^{1/2}$. Clearly $D_{00,\gamma\delta}(z)$ is analytic in $z\in\mathcal{C}\backslash[0,\infty)$.

$$\begin{aligned}\text{(b)}\quad D_{01,\gamma\delta}(z)&=[1-v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)|v_\gamma|^{1/2}]v_\gamma^{1/2}G_0(z)v_\delta E_\delta G_\delta(z)\hat{\rho}_\delta\\ &\quad-[v_\gamma^{1/2}E_\gamma\sigma_\gamma^{-1}]\sigma_\gamma G_0(z)v_\delta E_\delta G_\delta(z)\hat{\rho}_\delta,\qquad \gamma\neq\delta.\end{aligned} \tag{3.18}$$

The terms $[\cdots]$ in (3.18) have already been discussed in (3.17).

$$v_\gamma^{1/2}G_0(z)v_\delta E_\delta G_\delta(z)\hat{\rho}_\delta=[v_\gamma^{1/2}G_0(z)\hat{\sigma}_\delta]\hat{\sigma}_\delta^{-1}E_\delta\hat{\rho}_\delta-[v_\gamma^{1/2}E_\delta\hat{\rho}_\delta^{-1}]\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta. \tag{3.19}$$

Lemma 3.2(iii) and (v) apply for $v_\gamma^{1/2}E_\delta\hat{\rho}_\delta^{-1}$ and $\hat{\sigma}_\delta^{-1}E_\delta\hat{\rho}_\delta$. Lemma 3.4(i) applies for $v_\gamma^{1/2}G_0(z)\hat{\sigma}_\delta$. The operator $\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ is treated in Lemma 3.4(ii). In particular (3.9) shows its analyticity in $z\in\mathcal{C}\backslash[e_\delta^1,\infty)$ (and hence analyticity of $\bar{D}_{01}(z)$ in $z\in\mathcal{C}\backslash[E_0^{(2)},\infty)$). The operator $\sigma_\gamma G_0(z)v_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ can be discussed identically to (3.19) after replacing $v_\gamma^{1/2}$ by σ_γ.

$$\text{(c)}\quad D_{11,\gamma\delta}(z)=[\hat{\rho}_\gamma^{-1}E_\gamma e^{-b|x_\delta|}|u_\delta|^{1/2}\hat{\rho}_\delta^{-1}][u_\delta^{1/2}E_\delta]\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta,\qquad \gamma\neq\delta \tag{3.20}$$

if H(III) holds (if only H(II) holds replace $e^{-b|x_\delta|}$ by $(1+|x_\delta|)^{-2-\nu}$) and Lemma 3.2(i), (ii) and Lemma 3.4(ii) apply. Again $\tilde{D}_{11}(z)$ is analytic in $z \in \mathscr{C}\backslash[E_0^{(2)}, \infty)$.

(d) $\lim_{|z|\to\infty} \|\overset{\Delta}{A}(z)\| = 0$, $z \in \mathscr{C}$ is proved in [16] (for $z \in \mathscr{C}\backslash[0,\infty)$) and in [19]. □

Next we define

$$\overset{\Delta}{\mathscr{E}}{}^{\pm} = \{z \in \mathscr{C} \mid \overset{\Delta}{A}(z \pm i0)\Phi = -\Phi, \Phi \in \overset{\Delta}{\mathscr{H}}, \Phi \neq 0\}, \qquad \overset{\Delta}{\mathscr{E}} = \overset{\Delta}{\mathscr{E}}{}^{+} \cup \overset{\Delta}{\mathscr{E}}{}^{-} \tag{3.21}$$

and note the following lemma.

LEMMA 3.8. *Assume* H(II). *Then*

(i) $\overset{\Delta}{\mathscr{E}}$ *is a compact subset of* $\mathscr{R}$ *of Lebesgue measure zero*;

(ii) $\sigma_d(H) = \overset{\Delta}{\mathscr{E}} \cap (-\infty, E_0^{(2)})$ *is finite,* $\sigma_p(H) \subset \overset{\Delta}{\mathscr{E}}$, $\sigma_p(H) \cap (-\infty, 0) = \overset{\Delta}{\mathscr{E}} \cap (-\infty, 0)$ *consists of eigenvalues of H with finite multiplicity accumulating at most at zero.*

Proof. For (i) cf. [16], [19], [23]; for (ii) see [11], [16], [19].

Lemma 3.7 implies Lemma 3.9 as follows.

LEMMA 3.9. *Assume* H(II) *or* H(III). *Then*

$$[1+\overset{\Delta}{D}(z)]^{-1} \in \mathscr{B}(\overset{\Delta}{\mathscr{H}}) \quad \textit{for all } z \in \mathscr{C}\backslash\overset{\Delta}{\mathscr{E}}.$$

Moreover $\overset{\Delta}{D}(z)$ *converges in norm as* $z \to \lambda \pm i0$, *uniformly in* $\lambda \in \mathscr{R}$ *and* $[1+\overset{\Delta}{D}(z)]^{-1}$ *converges in norm as* $z \to \lambda \pm i0$, $\lambda \notin \overset{\Delta}{\mathscr{E}}$, *uniformly in* $\lambda \in \Delta$, $\Delta \subset \mathscr{R}\backslash\overset{\Delta}{\mathscr{E}}$ *compact if z varies in* $\Delta_\pm = \{z \in \mathscr{C} \mid \operatorname{Re} z \in \Delta, \operatorname{Im} z \in [0, \pm 1]\}$.

Finally we get for the three-particle resolvent $G(z)$

LEMMA 3.10. *Assume* H(II) *or* H(III). *Then*

$$\hat{\rho}_\gamma v_\gamma^{1/2} G(z) |v_\delta|^{1/2} = \hat{\rho}_\gamma \{J(z)[1+\overset{\Delta}{D}(z)]^{-1} K(z)\}_{\gamma\delta} \rho_\delta^{-1} |v_\delta|^{1/2}, \qquad \gamma, \delta = 1, 2, 3$$

is norm continuous in z as $z \to \lambda \pm i0$, $\lambda \notin \overset{\Delta}{\mathscr{E}}$, *uniformly continuous in* $\lambda \in \Delta$, $\Delta \in \mathscr{R}\backslash\overset{\Delta}{\mathscr{E}}$ *compact if z varies in* $\Delta_\pm$ (*cf. Lemma* 3.9).

Proof. The proof follows from

$$\hat{\rho}_\gamma \left\{ J(z) \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} \right\}_\gamma = \hat{\rho}_\gamma g_{0,\gamma} + [v_\gamma^{1/2} E_\gamma] \hat{\rho}_\gamma E_\gamma G_\gamma(z) \hat{\rho}_\gamma g_{1,\gamma}$$

and Lemma 3.2(i), Lemma 3.4(ii), from Lemma 3.9, from

$$K_{0,\gamma\delta}(z) \rho_\delta^{-1} |v_\delta|^{1/2} = v_\gamma^{1/2} (1 - E_\gamma) G_\gamma(z) |v_\delta|^{1/2}$$

which is discussed for $\gamma = \delta$ in Lemma 3.4(iii) and for $\gamma \neq \delta$ in Lemma 3.7 (cf. $D_{00,\gamma\delta}(z)$), and from

$$K_{1,\gamma\delta} \rho_\delta^{-1} |v_\delta|^{1/2} = \hat{\rho}_\gamma^{-1} E_\gamma |v_\delta|^{1/2} \in \mathscr{B}(\mathscr{H})$$

by Lemma 3.2(iii). □

Using the above methods, one can easily prove that $G(z) \in \mathscr{B}_\infty(L^2_{1+\nu/2}(\mathscr{R}^6), L^2_{-1-\nu/2}(\mathscr{R}^6))$, $z \in \mathscr{C}\backslash\overset{\Delta}{\mathscr{E}}$ and converges in $\mathscr{B}(L^2_{1+\nu/2}(\mathscr{R}^6), L^2_{-1-\nu/2}(\mathscr{R}^6))$-norm as $z \to \lambda \pm i0$, $\lambda \in \mathscr{R}\backslash\overset{\Delta}{\mathscr{E}}$, uniformly in $\lambda \in \Delta$, $\Delta \subset \mathscr{R}\backslash\overset{\Delta}{\mathscr{E}}$ compact, if z varies in $\Delta_\pm$, where $L^2_\mu(\mathscr{R}^6) = L^2(\mathscr{R}^6; (1+|x_\alpha|^2+|y_\alpha|^2)^\mu \, d^3x_\alpha \, d^3y_\alpha)$. In fact one can prove Hölder continuity of $G(z)$ for z in compacts not intersecting $\overset{\Delta}{\mathscr{E}}$ (cf. [16]).

4. Continuity of averaged total cross sections. Given the results of § 3 we rederive the continuity results for averaged total cross sections in [5]. Because of our stronger decay assumptions in H(II) when compared to [5] we are particularly able to discuss the threshold behaviour of the scattering operator, resp., the cross section for two-cluster initial channels.

In the following the operator $M_\alpha^j(A, \lambda)$ introduced at the end of § 2 plays a central role. From the definitions (2.35) and (2.41) one infers

$$\begin{aligned} &M_\alpha^j(AB, \lambda) = M_\alpha^j(A, \lambda)B, \qquad \operatorname{Ran}(B) \subset \mathscr{D}(A), \\ (4.1)\quad &M_\alpha^j(A, \lambda) = M_\alpha^j(E_\alpha^j A, \lambda), \qquad \alpha = 0, 1, 2, 3, \\ &M_0^0(A, \lambda) = M_0^0(E_\Lambda(H_0)A, \lambda) = \chi_\Lambda(\lambda) M_0^0(A, \lambda), \qquad \lambda \in \Lambda \subset \mathscr{R} \text{ a Borel set} \end{aligned}$$

where $E_\Lambda(H_0)$ denotes the spectral projection associated with H_0. We recall ([5])

LEMMA 4.1. *Assume* H(II). *Then*

(i) *For $\alpha \neq 0$, $\alpha \neq \beta$ and $|g_\kappa(x_\alpha, x_\beta)| \leqq e^{\kappa |x_\alpha|} \phi(x_\beta)$, $\phi \in L^2(\mathscr{R}^3)$, $\kappa < (-2m_\alpha e_\alpha^j)^{1/2}$, we get $M_\alpha^j(g_\kappa, \lambda) \in \mathscr{B}_2(\mathscr{H}, L^2(S^2))$, $\lambda \geqq e_\alpha^j$ and $M_\alpha^j(g_\kappa, \lambda)$ is continuous in $\mathscr{B}_2$-norm for $\lambda \geqq e_\alpha^j$, $\alpha = 1, 2, 3$.*

(ii) *Let $\alpha \neq 0$, $\alpha \neq \beta$; then $M_\alpha^j(|v_\beta|^{1/2}, \lambda)$, $M_\alpha^j(\rho_\alpha, \lambda) \in \mathscr{B}_4(\mathscr{H}, L^2(S^2))$, $\lambda \geqq e_\alpha^j$ and both operators are continuous in $\mathscr{B}_4$-norm for $\lambda \geqq e_\alpha^j$, $\alpha = 1, 2, 3$.*

(iii) *Assume $\phi \in L^p(\mathscr{R}^3)$, where $p = 2$ or 3 and denote by ϕ_γ the operator of multiplication by $\phi(x_\gamma)$. Then $M_0^0(\phi_\gamma, \lambda) \in \mathscr{B}(\mathscr{H}, L^2(E^5))$, $\lambda \geqq 0$ and $M_0^0(\phi_\gamma, \lambda)$ is strongly continuous in $\lambda \geqq 0$. Moreover $M_0^0(\phi_\gamma, \lambda) \notin \mathscr{B}_\infty(\mathscr{H}, L^2(E^5))$ as long as $\phi_\gamma \not\equiv 0$.*

Introducing the decomposed operators ([3])

$$(4.2)\qquad R_{\beta\alpha}^{lj}(\lambda) = S_{\beta\alpha}^{lj}(\lambda) - \delta_{\beta\alpha}^{lj} 1, \quad \lambda \in [\sup(e_\alpha^j, e_\beta^l), \infty), \quad \alpha \neq 0, \quad \beta = 0, 1, 2, 3,$$

we obtain the splitting ([5])

$$(4.3)\quad R_{\beta\alpha}^{lj}(\lambda) = {}^{(1)}R_{\beta\alpha}^{lj}(\lambda) + {}^{(2)}R_{\beta\alpha}^{lj}(\lambda), \quad \alpha \neq 0, \quad \beta = 0, 1, 2, 3, \quad \lambda \in [\sup(e_\alpha^j, e_\beta^l), \infty) \backslash \overset{\Delta}{\mathscr{E}}$$

where

$$(4.4)\qquad {}^{(1)}R_{\beta\alpha}^{lj}(\lambda) = -2\pi i \sum_{\substack{\gamma=1 \\ \gamma \neq \alpha}}^{3} M_\beta^l(|v_\gamma|^{1/2}, \lambda) M_\alpha^j(v_\gamma^{1/2}, \lambda)^*, \qquad \lambda \in [\sup(e_\alpha^j, e_\beta^l), \infty)$$

represents the Born-term (for $\beta = \gamma$ cf. (6.4)) and

$$\begin{aligned} &{}^{(2)}R_{\beta\alpha}^{lj}(\lambda) = 2\pi i \sum_{\gamma=1}^{3} [1 + \delta_{\beta 0}\delta_{\beta\gamma} - \delta_{\beta\gamma}] \sum_{\substack{\delta=1 \\ \delta \neq \alpha}}^{3} M_\beta^l(|v_\gamma|^{1/2}, \lambda) \\ (4.5)\qquad &\cdot \{J[1 + \overset{\Delta}{D}]^{-1} K\}_{\gamma\delta}(\lambda + i0) M_\alpha^j(v_\delta \rho_\delta^{-1}, \lambda)^*, \qquad \lambda \in [\sup(e_\alpha^j, e_\beta^l), \infty) \backslash \overset{\Delta}{\mathscr{E}}. \end{aligned}$$

For extensive discussions of analyticity properties of $S_{\beta\alpha}^{lj}(\lambda)$ using dilation analytic two-body potentials v_α we refer to [6], [10], [18] and [33]. Trace relations in connection with three-particle systems appeared in [8], [9]. The corresponding Riemann surface and resonances in case of exponentially decaying two-body interactions are studied in great detail in [7].

Going back to Lemma 4.1, we obtain the following.

LEMMA 4.2. *Assume* H(II) *and $\alpha \neq 0$. Then $R_{\beta\alpha}^{lj}(\lambda) \in \mathscr{B}_2(L^2(S^2))$, $\beta \neq 0$, $R_{0\alpha}^{0j}(\lambda) \in \mathscr{B}_2(L^2(S^2), L^2(E^5))$ and $R_{\beta\alpha}^{lj}(\lambda)$ is continuous in $\mathscr{B}_2$-norm for $\lambda \in [\sup(e_\alpha^j, e_\beta^l), \infty) \backslash \overset{\Delta}{\mathscr{E}}$, $\beta = 0, 1, 2, 3$.*

Proof. The only difference to [5] concerns the fact that H(II) allows to include threshold points $\sup(e_\alpha^j, e_\beta^l) \in \theta$. In fact, one can follow [5] step by step, replacing its Proposition 1 (which excludes thresholds) by our Lemma 3.9. □

By (4.2) one obtains analogous results for the decomposed scattering operator $S^{lj}_{\beta\alpha}(\lambda)$.

Next we turn to the concept of averaged total cross sections for initial two-cluster channels as defined, e.g., in [3]-[5]:

$$\bar{\sigma}^{lj}_{\beta\alpha}(\lambda)=\pi[2n_\alpha(\lambda-e^j_\alpha)]^{-1}\|R^{lj}_{\beta\alpha}(\lambda)\|^2_2, \qquad \alpha\neq 0, \quad \beta=0,1,2,3. \tag{4.6}$$

Hence we get the next theorem.

THEOREM 4.3. *Assume* H(II), $\alpha\neq 0$ *and* $\beta=0,1,2,3$. *If* $\sup(e^j_\alpha, e^l_\beta)\notin\overset{\Delta}{\mathscr{E}}$ *then* $\bar{\sigma}^{lj}_{\beta\alpha}(\lambda)$ *is continuous in* $\lambda\in(\sup(e^j_\alpha,e^l_\beta),\infty)\backslash\overset{\Delta}{\mathscr{E}}$. *In addition we obtain the threshold behaviour*

Elastic case: $\alpha=\beta$, $j=l$.

$$\bar{\sigma}^{jj}_{\alpha\alpha}(\lambda)\underset{\lambda\downarrow e^j_\alpha}{=}O(1). \tag{4.7}$$

Inelastic case: $\alpha=\beta$, $j\neq l$ (*i.e.* $\psi^j_\alpha\neq$ const. ψ^l_α).

$$\bar{\sigma}^{lj}_{\alpha\alpha}(\lambda)\underset{\lambda\downarrow\sup(e^j_\alpha,e^l_\alpha)}{=}\begin{cases}O((\lambda-e^j_\alpha)^{-1/2}), & e^j_\alpha>e^l_\alpha,\\ O((\lambda-e^l_\alpha)^{1/2}), & e^j_\alpha<e^l_\alpha,\\ O(1), & e^j_\alpha=e^l_\alpha.\end{cases} \tag{4.8}$$

Rearrangement case: $\alpha\neq\beta$, $\beta\neq 0$.

$$\bar{\sigma}^{lj}_{\beta\alpha}(\lambda)\underset{\lambda\downarrow\sup(e^j_\alpha,e^l_\beta)}{=}\begin{cases}O((\lambda-e^j_\alpha))^{-1/2}, & e^j_\alpha>e^l_\beta,\\ O((\lambda-e^l_\beta))^{1/2}, & e^j_\alpha<e^l_\beta,\\ O(1), & e^j_\alpha=e^l_\beta.\end{cases} \tag{4.9}$$

Breakup case: $\beta=0$.

$$\bar{\sigma}^{0j}_{0\alpha}(\lambda)\underset{\lambda\downarrow 0}{=}O(\lambda^{1/2}). \tag{4.10}$$

Proof. The continuity statement is due to Lemma 4.2 of [5]. The threshold behaviour (4.7)-(4.9) can be read off from

$$\begin{aligned}M^j_\alpha(f_\gamma,\lambda)(\omega^j_\alpha,x_\alpha,y_\alpha)&=(2\pi)^{-3/2}n^{1/2}_\alpha[2n_\alpha(\lambda-e^j_\alpha)]^{1/4}f_\gamma(x_\gamma(x_\alpha,y_\alpha),y_\gamma(x_\alpha,y_\alpha))\\ &\quad\cdot\exp(-i[2n_\alpha(\lambda-e^j_\alpha)]^{1/2}\omega^j_\alpha y_\alpha)\overline{\psi^j_\alpha(x_\alpha)}, \qquad \alpha\neq 0\end{aligned} \tag{4.11}$$

where f_γ denotes the operator of multiplication by $f_\gamma(x_\gamma,y_\gamma)$ (i.e. $f_\gamma=|v_\gamma|^{1/2}$, $v^{1/2}_\gamma$, $v_\gamma\rho^{-1}_\gamma$ according to (4.4) and (4.5)). The estimate (4.10) follows from

$$\begin{aligned}\|M^0_0(|v_\gamma|^{1/2},\lambda)g\|^2_2&=\int_{|q|\leqq 1}d^3q\int_{S^2}d\omega|(U^0_0|v_\gamma|^{1/2}g)(\lambda,\omega,q)|^2\\ &=2m_\gamma(2n_\gamma)^{3/2}\int_{|q|\leqq 1}d^3q\,\lambda^{3/2}\|M(|v_\gamma|^{1/2},2m_\gamma\lambda(1-|q|^2))\\ &\qquad\qquad \hat{g}(\cdot,(2n_\gamma\lambda)^{1/2}q)\|^2_2,\\ &\qquad\qquad\qquad g\in\mathscr{S}(\mathscr{R}^6)\end{aligned}$$

where

$$M(|v_\gamma|^{1/2},\mu):\begin{cases}L^2(\mathscr{R}^3)\to L^2(S^2)\\ f\to(M(|v_\gamma|^{1/2},\mu)f)(\omega)=2^{-1/2}\mu^{1/4}(\overline{|v_\gamma|^{1/2}}f)(\mu^{1/2}\omega), \qquad \omega\in S^2\end{cases}$$

is a "two-particle" operator and

$$\hat{g}(\cdot,p)=(2\pi)^{-3/2}\int_{\mathscr{R}^3}d^3y\,e^{-ipy}g(\cdot,y), \qquad p\in\mathscr{R}^3$$

denotes the partial Fourier transform. Observing

$$\|M(|v_\gamma|^{1/2},\mu)\|_2^2=(2\pi)^{-2}\mu^{1/2}\|v_\gamma\|_1<\infty,$$

we infer

$$\begin{aligned}\|M_0^0(|v_\gamma|^{1/2},\lambda)g\|_2^2&\leqq \text{const. }\|v_\gamma\|_1\int_{|q|\leqq 1}d^3q(1-|q|^2)^{1/2}\lambda^2\|\hat{g}(\cdot,(2n_\gamma\lambda)^{1/2}q)\|_2^2\\&\leqq\text{const. }\|v_\gamma\|_1\lambda^{1/2}\int_{\mathscr{R}^3}d^3p\|\hat{g}(\cdot,p)\|_2^2\\&=\text{const. }\|v_\gamma\|_1\|g\|_2^2\lambda^{1/2}.\end{aligned}\qquad\square$$

5. Continuity of two-cluster scattering amplitudes. In this section we consider two-cluster scattering amplitudes and prove their continuity with respect to energy- and angle variables. Again we discuss their threshold behaviour, in particular we define the concept of elastic scattering lengths and derive an explicit formula in terms of the three-body resolvent.

In order to obtain an expression for the scattering amplitudes as a scalar product in $\mathscr{H}$ (in analogy to the two-body case, cf. e.g. [1], [2]) we introduce

Hypothesis (IV). In addition to H(II) we assume that $u_\alpha\in L^1(\mathscr{R}^3)$, $\alpha=1,2,3$.

According to [3] the scattering amplitude for two-cluster initial channels are proportional to the integral kernel of $R_{\beta\alpha}^{lj}(\lambda)$, more precisely one defines

$$f_{\beta\alpha}^{lj}(\lambda,\omega_\alpha^j\to\omega_\beta^l)=-2\pi i[2n_\alpha(\lambda-e_\alpha^j)]^{-1/2}R_{\beta\alpha}^{lj}(\lambda,\omega_\beta^l,\omega_\alpha^j),\tag{5.1}$$
$$\alpha\neq 0,\quad \beta=0,1,2,3.$$

Thus $f_{\beta\alpha}^{lj}$ splits up into a Born-term and a remainder as can be seen from (4.3)

$$\begin{aligned}&f_{\beta\alpha}^{lj}(\lambda,\omega_\alpha^j\to\omega_\beta^l)={}^{(1)}f_{\beta\alpha}^{lj}(\lambda,\omega_\alpha^j\to\omega_\beta^l)+{}^{(2)}f_{\beta\alpha}^{lj}(\lambda,\omega_\alpha^j\to\omega_\beta^l),\\&\qquad\alpha\neq 0,\quad\beta=0,1,2,3,\quad\lambda\in[\sup(e_\alpha^j,e_\beta^l),\infty)\backslash\overset{\Delta}{\mathscr{E}}.\end{aligned}\tag{5.2}$$

We start with the elastic case $\alpha=\beta\neq 0$, $j=l$.

THEOREM 5.1. *Assume* H(IV), $\alpha\neq 0$ *and* $e_\alpha^j\notin\overset{\Delta}{\mathscr{E}}$. *Then* $f_{\alpha\alpha}^{jj}(\lambda,\omega_\alpha^j\to\hat{\omega}_\alpha^j)$ *is continuous with respect to* $(\lambda,\omega_\alpha^j,\hat{\omega}_\alpha^j)\in\{[e_\alpha^j,\infty)\backslash\overset{\Delta}{\mathscr{E}})\}\times S^2\times S^2$. *In particular*

$$\begin{aligned}&{}^{(1)}f_{\alpha\alpha}^{jj}(\lambda,\omega_\alpha^j\to\hat{\omega}_\alpha^j)\\&\quad=-(2\pi)^{-1}n_\alpha\sum_{\substack{\gamma=1\\\gamma\neq\alpha}}^{3}(|v_\gamma|^{1/2}\exp(i[2n_\alpha(\lambda-e_\alpha^j)]^{1/2}\hat{\omega}_\alpha^jy_\alpha)\psi_\alpha^j,v_\gamma^{1/2}\\&\qquad\cdot\exp(i[2n_\alpha(\lambda-e_\alpha^j)]^{1/2}\omega_\alpha^jy_\alpha)\psi_\alpha^j),\\&\qquad(\lambda,\omega_\alpha^j,\hat{\omega}_\alpha^j)\in[e_\alpha^j,\infty)\times S^2\times S^2,\end{aligned}\tag{5.3}$$

$$\begin{aligned}&{}^{(2)}f_{\alpha\alpha}^{jj}(\lambda,\omega_\alpha^j\to\hat{\omega}_\alpha^j)=(2\pi)^{-1}n_\alpha\sum_{\substack{\gamma=1\\\gamma\neq\alpha}}^{3}\sum_{\substack{\delta=1\\\delta\neq\alpha}}^{3}(\rho_\gamma^{-1}|v_\gamma|^{1/2}\exp(i[2n_\alpha(\lambda-e_\alpha^j)]^{1/2}\hat{\omega}_\alpha^jy_\alpha)\psi_\alpha^j,\\&\quad\rho_\gamma\{J[1+\overset{\Delta}{D}]^{-1}K\}_{\gamma\delta}(\lambda+i0)\rho_\delta^{-1}|v_\delta|^{1/2}v_\delta^{1/2}\exp(i[2n_\alpha(\lambda-e_\alpha^j)]^{1/2}\omega_\alpha^jy_\alpha)\psi_\alpha^j),\\&\qquad(\lambda,\omega_\alpha^j,\hat{\omega}_\alpha^j)\in\{[e_\alpha^j,\infty)\backslash\overset{\Delta}{\mathscr{E}}\}\times S^2\times S^2.\end{aligned}\tag{5.4}$$

Here $(\cdot,\cdot)$ *means the scalar product in* $\mathscr{H}=L^2(\mathscr{R}^6;d^3x_\alpha\,d^3y_\alpha)$, $\psi_\alpha^j=\psi_\alpha^j(x_\alpha)$ *and in obvious notation the plane waves in* (5.3) *and* (5.4) *depend on the variable* y_α. *The* *threshold behaviour reads*

$$f^{jj}_{\alpha\alpha}(\lambda, \omega^j_\alpha \to \hat{\omega}^j_\alpha) \underset{\lambda \downarrow e^j_\alpha}{=} O(1). \tag{5.5}$$

Proof. Clearly

$$|v_\gamma|^{1/2}\, e^{\hat{q}^j_\alpha y_\alpha} \psi^j_\alpha, \quad \gamma \neq \alpha \quad (\hat{q}^j_\alpha = [2n_\alpha(\lambda - e^j_\alpha)]^{1/2} \hat{\omega}^j_\alpha)$$

is strongly continuous in $(\lambda, \hat{\omega}^j_\alpha) \in [e^j_\alpha, \infty) \times S^2$ by dominated convergence since

$$\| |v_\gamma|^{1/2}\, e^{i\hat{q}^j_\alpha y_\alpha} \psi^j_\alpha \|_2^2 = \text{const. } \|v_\gamma\|_1 \|\psi^j_\alpha\|_2^2 < \infty, \qquad \gamma \neq \alpha.$$

Similarly $\rho_\gamma^{-1}|v_\gamma|^{1/2}\, e^{i\hat{q}^j_\alpha y_\alpha}\psi^j_\alpha$, $\gamma \neq \alpha$ is strongly continuous in $(\lambda, \hat{\omega}^j_\alpha) \in [e^j_\alpha, \infty) \times S^2$ since

$$\begin{aligned} \|\rho_\gamma^{-1}|v_\gamma|^{1/2}\, e^{i\hat{q}^j_\alpha y_\alpha}\psi^j_\alpha\|_2^2 &\leqq \text{const.} \int_{\mathscr{R}^6} d^3x_\alpha\, d^3z_\alpha [1+|x_\alpha|]^{2+\nu} |\psi^j_\alpha(x_\alpha)|^2 [1+|z_\alpha|]^{2+\nu} |v_\gamma(z_\alpha)| \\ &\leqq \text{const. } \|(1+|\cdot|)^{1+\nu/2}\psi^j_\alpha\|_2^2 \|u_\gamma\|_1 < \infty, \qquad \gamma \neq \alpha \end{aligned}$$

by H(IV), Lemma 3.1 and

$$\begin{aligned} &|y_\gamma| \leqq |y_\alpha| + |x_\alpha| \quad (\text{cf. } (2.5)), \\ &\rho_\gamma^{-2}(y_\gamma) = [1+|y_\gamma|]^{2+\nu} \leqq [1+|x_\alpha|]^{2+\nu}[1+|y_\alpha|]^{2+\nu}. \end{aligned}$$

By Lemma 3.10, $\rho_\gamma\{(J[1+\overset{\Delta}{D}]^{-1}K)\}_{\gamma\delta}(\lambda + i0)\rho_\delta^{-1}|v_\delta|^{1/2}$ is norm continuous in $\lambda \in [e^j_\alpha, \infty) \backslash \overset{\Delta}{\mathscr{E}}$ completing the proof. □

The inelastic case $\alpha = \beta \neq 0$, $j \neq l$ is contained in the following theorem.

THEOREM 5.2. *Assume* H(IV), $\alpha \neq 0, j \neq l$ (*i.e.* $\psi^j_\alpha \neq$ const. ψ^l_α) *and* $\sup(e^j_\alpha, e^l_\alpha) \notin \overset{\Delta}{\mathscr{E}}$. *Then* $f^{lj}_{\alpha\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\alpha)$ *is continuous with respect to* $(\lambda, \omega^j_\alpha, \omega^l_\alpha) \in \{(\sup(e^j_\alpha, e^l_\alpha), \infty) \backslash \overset{\Delta}{\mathscr{E}}\} \times S^2 \times S^2$. *In particular*

$$\begin{aligned} {}^{(1)}f^{lj}_{\alpha\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\alpha) &= -(2\pi)^{-1} n_\alpha (\lambda - e^j_\alpha)^{-1/4} (\lambda - e^l_\beta)^{1/4} \\ &\quad \cdot \sum_{\substack{\gamma=1 \\ \gamma \neq \alpha}}^{3} (|v_\gamma|^{1/2} \exp(i[2n_\alpha(\lambda - e^l_\alpha)]^{1/2} \omega^l_\alpha y_\alpha)\psi^l_\alpha, v_\gamma^{1/2} \\ &\quad \cdot \exp(i[2n_\alpha(\lambda - e^j_\alpha)]^{1/2}\omega^j_\alpha y_\alpha)\psi^j_\alpha), \\ &\quad (\lambda, \omega^j_\alpha, \omega^l_\alpha) \in (\sup(e^j_\alpha, e^l_\alpha), \infty) \times S^2 \times S^2, \end{aligned} \tag{5.6}$$

$$\begin{aligned} {}^{(2)}f^{lj}_{\alpha\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\alpha) &= (2\pi)^{-1} n_\alpha (\lambda - e^j_\alpha)^{-1/4} (\lambda - e^l_\alpha)^{1/4} \sum_{\substack{\gamma=1 \\ \gamma \neq \alpha}}^{3} \sum_{\substack{\delta=1 \\ \delta \neq \alpha}}^{3} \\ &\quad \cdot (\rho_\gamma^{-1}|v_\gamma|^{1/2} \exp(i[2m_\alpha(\lambda - e^l_\alpha)]^{1/2}\omega^l_\alpha y_\alpha)\psi^l_\alpha, \rho_\gamma \\ &\quad \cdot \{J[1+\overset{\Delta}{D}]^{-1}K\}_{\gamma\delta}(\lambda + i0)\rho_\delta^{-1}|v_\delta|^{1/2} \\ &\quad \cdot v_\delta^{1/2} \exp(i[2n_\alpha(\lambda - e^j_\alpha)]^{1/2}\omega^j_\alpha y_\alpha)\psi^j_\alpha), \\ &\quad (\lambda, \omega^j_\alpha, \omega^l_\alpha) \in \{(\sup(e^j_\alpha, e^l_\alpha), \infty) \backslash \overset{\Delta}{\mathscr{E}}\} \times S^2 \times S^2. \end{aligned} \tag{5.7}$$

The threshold behaviour reads

$$(5.8)\qquad f^{lj}_{\alpha\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\alpha) \underset{\lambda \downarrow \sup(e^j_\alpha, e^l_\alpha)}{=} \begin{cases} O((\lambda - e^j_\alpha)^{-1/4}), & e^j_\alpha > e^l_\alpha, \\ O((\lambda - e^l_\alpha)^{1/4}), & e^j_\alpha < e^l_\alpha, \\ O(1), & e^j_\alpha = e^l_\alpha. \end{cases}$$

Proof. The proof is identical to that of Theorem 5.1.

Finally we discuss the rearrangement case $\alpha \neq \beta$, $\alpha \neq 0$, $\beta \neq 0$.

THEOREM 5.3. *Assume* H(IV), $\alpha \neq 0$, $\beta \neq 0$, $\alpha \neq \beta$ *and* $\sup(e^j_\alpha, e^l_\beta) \notin \overset{\Delta}{\mathscr{E}}$. *Then* $f^{lj}_{\beta\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\beta)$ *is continuous with respect to* $(\lambda, \omega^j_\alpha, \omega^l_\beta) \in \{(\sup(e^j_\alpha, e^l_\beta), \infty) \backslash \overset{\Delta}{\mathscr{E}}\} \times S^2 \times S^2$. *In particular*

$$\begin{aligned} {}^{(1)}f^{lj}_{\beta\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\beta) &= -(2\pi)^{-1}\, n_\alpha^{1/4} n_\beta^{3/4} (\lambda - e^j_\alpha)^{-1/4} (\lambda - e^l_\beta)^{1/4} \\ &\quad \cdot \sum_{\substack{\gamma=1 \\ \gamma \neq \alpha}}^{3} (|v_\gamma|^{1/2} \sigma_\alpha^{3\delta_{\gamma\beta}/2} \exp(i[2n_\beta(\lambda - e^l_\beta)]^{1/2} \omega^l_\beta y_\beta) \\ (5.9)\qquad &\quad \cdot \psi^l_\beta,\, v_\gamma^{1/2} \sigma_\alpha^{-3\delta_{\gamma\beta}/2} \\ &\quad \cdot \exp(i[2n_\alpha(\lambda - e^j_\alpha)]^{1/2} \omega^j_\alpha y_\alpha) \psi^j_\alpha), \\ &\qquad (\lambda, \omega^j_\alpha, \omega^l_\beta) \in (\sup(e^j_\alpha, e^l_\beta), \infty) \times S^2 \times S^2, \end{aligned}$$

$$\begin{aligned} {}^{(2)}f^{lj}_{\beta_\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\beta) &= (2\pi)^{-1} n_\alpha^{1/4} n_\beta^{3/4} (\lambda - e^j_\alpha)^{-1/4} (\lambda - e^l_\beta)^{1/4} \\ &\quad \cdot \sum_{\substack{\gamma=1 \\ \gamma \neq \beta}}^{3} \sum_{\substack{\delta=1 \\ \delta \neq \alpha}}^{3} (\rho_\gamma^{-1} |v_\gamma|^{1/2} \exp(i[2n_\beta(\lambda - e^l_\beta)]^{1/2} \omega^l_\beta y_\beta) \\ (5.10)\qquad &\quad \cdot \psi^l_\beta,\, \rho_\gamma \{J[1 + \overset{\Delta}{D}]^{-1} K\}_{\gamma\delta}(\lambda + i0) \rho_\delta^{-1} |v_\delta|^{1/2} \\ &\quad \cdot v_\delta^{1/2} \exp(i[2n_\alpha(\lambda - e^j_\alpha)]^{1/2} \omega^j_\alpha y_\alpha) \psi^j_\alpha), \\ &\qquad (\lambda, \omega^j_\alpha, \omega^l_\beta) \in \{(\sup(e^j_\alpha, e^l_\beta), \infty) \backslash \overset{\Delta}{\mathscr{E}}\} \times S^2 \times S^2. \end{aligned}$$

The threshold behaviour reads

$$(5.11)\qquad f^{lj}_{\beta\alpha}(\lambda, \omega^j_\alpha \to \omega^l_\beta) \underset{\lambda \downarrow \sup(e^j_\alpha, e^l_\beta)}{=} \begin{cases} O((\lambda - e^j_\alpha)^{-1/4}), & e^j_\alpha > e^l_\beta, \\ O((\lambda - e^l_\beta)^{1/4}, & e^j_\alpha < e^l_\beta, \\ O(1), & e^j_\alpha = e^l_\beta. \end{cases}$$

Proof. Since the discussion of ${}^{(2)}f^{lj}_{\beta\alpha}$ is identical to that of Theorem 5.1 we only consider ${}^{(1)}f^{lj}_{\beta\alpha}$. For $\gamma \neq \beta$ $(\gamma \neq \alpha)$ the corresponding proof of Theorem 5.1 applies. For $\gamma = \beta$ $(\gamma \neq \alpha)$ $e^{iq^j_\alpha \omega^j_\alpha y_\alpha} \sigma_\alpha^{-3/2} \psi^j_\alpha$ is strongly continuous in $(\lambda, \omega^j_\alpha) \in [e^j_\alpha, \infty) \times S^2$ since $e^{\kappa|\cdot|} \psi^j_\alpha \in L^2(\mathscr{R}^3)$ for all $\kappa < (-2m_\alpha e^j_\alpha)^{1/2}$ by Lemma 3.1. In addition $|v_\beta|^{1/2} \sigma_\alpha^{3/2}\, e^{iq^l_\beta \omega^l_\beta y_\beta} \psi^l_\beta$ is strongly continuous in $(\lambda, \omega^l_\beta) \in [e^l_\beta, \infty) \times S^2$ since

$$\begin{aligned} \| |v_\beta|^{1/2} \sigma_\alpha^{3/2}\, e^{iq^l_\beta \omega^l_\beta y_\beta} \psi^l_\beta \|_2^2 &= \text{const.} \int_{\mathscr{R}^6} d^3x_\alpha\, d^3z_\alpha [1 + |x_\alpha|]^{-3-3\nu/2} |v_\beta(z_\alpha)|\, |\psi^l_\beta(z_\alpha)|^2 \\ &= \text{const.}\, \|(1 + |\cdot|)^{-3-3\nu/2}\|_1 \|\, |v_\beta|^{1/2} \psi^l_\beta \|_2^2 < \infty, \qquad \beta \neq \alpha. \qquad \square \end{aligned}$$

Scattering amplitudes have also been discussed in [27] on the basis of Faddeev's original treatment ([14]) (cf. also [32]). Analyticity properties of scattering amplitudes in the case of dilation analytic two-body potentials are studied in [10], [18], [19]. For exponentially decaying two-body interactions their analytic continuation to the corresponding Riemann surface is considered in [7].

Finally we turn to a discussion of elastic scattering lengths.

DEFINITION 5.4. Assume H(IV), $\alpha \neq 0$ and $e_\alpha^j \notin \overset{\Delta}{\mathscr{E}}$. Then the elastic scattering length $A_{\alpha\alpha}^{jj}$ is defined by

$$A_{\alpha\alpha}^{jj} = -\lim_{\lambda \downarrow e_\alpha^j} f_{\alpha\alpha}^{jj}(\lambda, \omega_\alpha^j \to \hat{\omega}_\alpha^j). \tag{5.12}$$

An explicit expression for $A_{\alpha\alpha}^{jj}$ in terms of the three-body resolvent $G(z)$ (in analogy to the two-body situation [1], [2]) reads as follows.

THEOREM 5.5. *Assume* H(IV), $\alpha \neq 0$, $e_\alpha^j \notin \overset{\Delta}{\mathscr{E}}$. *Then*

$$\begin{aligned}A_{\alpha\alpha}^{jj} &= (2\pi)^{-1} n_\alpha \sum_{\substack{\gamma=1\\ \gamma\neq\alpha}}^{3} (|v_\gamma|^{1/2}\psi_\alpha^j, v_\gamma^{1/2}\psi_\alpha^j)\\ &\quad - (2\pi)^{-1} n_\alpha \sum_{\substack{\gamma=1\\ \gamma\neq\alpha}}^{3} \sum_{\substack{\delta=1\\ \delta\neq\alpha}}^{3} (\rho_\gamma^{-1}|v_\gamma|^{1/2}\psi_\alpha^j, \rho_\gamma\{J[1+\overset{\Delta}{D}]^{-1}K\}_{\gamma\delta}\\ &\qquad\qquad (e_\alpha^j + i0)\rho_\delta^{-1}|v_\delta|^{1/2}v_\delta^{1/2}\psi_\alpha^j)\\ &= (2\pi)^{-1} n_\alpha \sum_{\substack{\gamma=1\\ \gamma\neq\alpha}}^{3} (|v_\gamma|^{1/2}\psi_\alpha^j, v_\gamma^{1/2}\psi_\alpha^j)\\ &\quad - (2\pi)^{-1} n_\alpha \sum_{\substack{\gamma=1\\ \gamma\neq\alpha}}^{3} \sum_{\substack{\delta=1\\ \delta\neq\alpha}}^{3} (\rho_\gamma^{-1}|v_\gamma|^{1/2}\psi_\alpha^j, \rho_\gamma v_\gamma^{1/2}\\ &\qquad\qquad \cdot (H - e_\alpha^j - i0)^{-1}|v_\delta|^{1/2}\rho_\delta\rho_\delta^{-1}v_\delta^{1/2}\psi_\alpha^j).\end{aligned} \tag{5.13}$$

Proof. This is obvious by the proof of Theorem 5.1.

Remark 5.6. For $e_{\alpha_0}^{j_0} = E_0^{(2)}$, $E_0^{(2)} \notin \overset{\Delta}{\mathscr{E}}$, $A_{\alpha_0\alpha_0}^{j_0j_0}$ is real.

Proof. It suffices to note that by Lemma 3.10

$$\rho_\gamma|v_\gamma|^{1/2}(H - E_0^{(2)} - i0)^{-1}|v_\delta|^{1/2}\rho_\delta = \rho_\gamma|v_\gamma|^{1/2}(H - E_0^{(2)} + i0)^{-1}|v_\delta|^{1/2}\rho_\delta, \qquad \gamma, \delta = 1, 2, 3.$$

□

Remark 5.7. The analogue of (5.13) in the two-body case reads for h_α (cf. [1], [2])

$$\begin{aligned}a_\alpha &= (2\pi)^{-1} m_\alpha(|v_\alpha|^{1/2}, v_\alpha^{1/2}) - (2\pi)^{-1} m_\alpha(|v_\alpha|^{1/2}, v_\alpha^{1/2}(h_\alpha - i0)^{-1}|v_\alpha|^{1/2}v_\alpha^{1/2})\\ &= (2\pi)^{-1} m_\alpha(|v_\alpha|^{1/2}, [1 + v_\alpha^{1/2}(h_{0,\alpha} - i0)^{-1}|v_\alpha|^{1/2}]^{-1}v_\alpha^{1/2}), \qquad \alpha = 1, 2, 3.\end{aligned} \tag{5.14}$$

For the physical relevance of scattering lengths in three-body systems see, e.g., [28] and the references cited therein.

6. Analytic expansions around thresholds in $\boldsymbol{\theta}\backslash\{0\}$. Finally we use the exponential decay of v_α in H(III) to extend our previous threshold considerations by deriving analytic expansions of two-cluster scattering operators and amplitudes with respect to channel momenta near points in $\theta\backslash\{0\}$.

We start with ($^{(}\hat{\rho}^{)}_\beta$ denotes ρ_β or $\hat{\rho}_\beta$)

LEMMA 6.1. *Assume* H(III) *and* $\alpha \neq 0$. *If* $M_\alpha^j(\lambda)$, $\lambda \geqq e_\alpha^j$ *denotes one of the operators* $M_\alpha^j(|v_\gamma|^{1/2}, \lambda)$, $\gamma \neq \alpha$, $M_\alpha^j(^{(}\hat{\rho}^{)}_\alpha, \lambda)$, $M_\alpha^j(^{(}\hat{\rho}^{)-1}_\gamma|v_\gamma|^{1/2}, \lambda)$, $\gamma \neq \alpha$ *then* $(\lambda - e_\alpha^j)^{-1/4}M_\alpha^j(\lambda)$

is analytic in $(\lambda - e_\alpha^j)^{1/2}$ *near zero in* $\mathcal{B}_2(\mathcal{H}, L^2(S^2))$*-norm. Moreover* $M_\alpha^j(\lambda)$ *is analytic in* $(\lambda - \lambda_0)$ *near zero in* $\mathcal{B}_2(\mathcal{H}, L^2(S^2))$*-norm for all* $\lambda_0 > e_\alpha^j$.

Proof. (a) Taking $f_\gamma = |v_\gamma|^{1/2}$ in (4.11) and observing $(\gamma \neq \alpha)$

$$\int_{\mathcal{R}^6} d^3x_\alpha\, d^3y_\alpha |v_\gamma(cx_\alpha + dy_\alpha)| \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|y_\alpha|)|\psi_\alpha^j(x_\alpha)|^2$$

$$= d^{-3}\int_{\mathcal{R}^6} d^3x_\alpha\, d^3z_\alpha \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|z_\alpha|/d)|v_\gamma(z)|$$

$$\cdot \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}c|x_\alpha|/d)|\psi_\alpha^j(x_\alpha)|^2 < \infty$$

for $|\lambda - e_\alpha^j|$ sufficiently small the assertions follow for $M_\alpha^j(|v_\gamma|^{1/2}, \lambda)$, $\gamma \neq \alpha$ concerning the expansion near e_α^j. Similar for the expansion near $\lambda_0 > e_\alpha^j$. (Note that $e^{\varepsilon|\cdot|}v_\gamma = e^{-(b-\varepsilon)|\cdot|}u_\gamma \in L^1(\mathcal{R}^3)$ for $\varepsilon > 0$ small enough.)

(b) Taking $f_\gamma = {}^{(}\hat{\rho}^{)}_\alpha$ in (4.11), we infer the assertions for $M_\alpha^j({}^{(}\hat{\rho}^{)}, \lambda)$ from

$$\int_{\mathcal{R}^6} d^3x_\alpha\, d^3y_\alpha |{}^{(}\hat{\rho}^{)}_\alpha(y_\alpha)|^2 \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|y_\alpha|)|\psi_\alpha^j(x_\alpha)|^2 < \infty$$

for $|\lambda - e_\alpha^j|$ small enough.

(c) Choosing $f_\gamma = {}^{(}\hat{\rho}^{)-1}_\gamma |v_\gamma|^{1/2}$ in (4.11) we prove the assertions for $M_\alpha^j({}^{(}\hat{\rho}^{)-1}_\gamma |v_\gamma|^{1/2}, \lambda)$ as follows: Noting

$$x_\gamma = cx_\alpha + dy_\alpha, \qquad y_\gamma = c'x_\alpha + d'y_\alpha, \qquad \gamma \neq \alpha,$$

$$|d'/d| \leqq 1, \qquad |c' - d^{-1}\, d'c| = 1$$

(cf. (2.5)) we obtain

$$\int_{\mathcal{R}^6} d^3x_\alpha\, d^3y_\alpha {}^{(}\hat{\rho}^{)-2}_\gamma(c'x_\alpha + d'y_\alpha) \exp(-b|cx_\alpha + dy_\alpha|)|u_\gamma(cx_\alpha + dy_\alpha)|$$

$$\cdot \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|y_\alpha|)|\psi_\alpha^j(x_\alpha)|^2$$

$$\leqq d^{-3}\int_{\mathcal{R}^6} d^3x_\alpha\, d^3z_\alpha {}^{(}\hat{\rho}^{)-2}_\gamma([c' - d^{-1}\, d'c]x_\alpha + d^{-1}\, d'z_\alpha)\, e^{-b|z_\alpha|}|u_\gamma(z_\alpha)|$$

$$\cdot \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|z_\alpha|/d)$$

$$\cdot \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}c|x_\alpha|/d)|\psi_\alpha^j(x_\alpha)|^2$$

$$\leqq d^{-3}\int_{\mathcal{R}^6} d^3x_\alpha\, d^3z_\alpha \exp(2a_\gamma|x_\alpha|) \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}c|x_\alpha|/d)|\psi_\alpha^j(x)|^2$$

$$\cdot \exp(2a_\gamma|z_\alpha|) \exp(2[2n_\alpha|\lambda - e_\alpha^j|]^{1/2}|z_\alpha|/d)\, e^{-b|z_\alpha|}|u_\gamma(z_\alpha)| < \infty$$

for $|\lambda - e_\alpha^j|$ small enough since

$$a_\gamma < \inf_{\beta=1,2,3}[-2m_\beta\, e_\beta^{N_\beta}]^{1/2}, \qquad 2a_\gamma < b. \qquad \square$$

LEMMA 6.2. *Assume* H(III). *Then* $\overset{\Delta}{D}(\lambda+i0)$ *is norm analytic in* $(\lambda-\lambda_0)^{1/2}$ *near zero for all* $\lambda_0<0$. *Thus* $[1+\overset{\Delta}{D}(\lambda+i0)]^{-1}$ *is norm analytic in* $(\lambda-\lambda_0)^{1/2}$ *near zero for all* $\lambda_0<0$, $\lambda_0\notin\overset{\Delta}{\mathscr{E}}$.

Proof. Cf. Lemma 3.7. (a) $D_{00,\gamma\delta}(z)$ is analytic in $z\in\mathscr{C}\backslash[0,\infty)$ and hence in $(\lambda-\lambda_0)$.

(b) $D_{01,\gamma\delta}(z)$ is analytic in $z\in\mathscr{C}\backslash[E_0^{(2)},\infty)$ because of the term $\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$, the rest is analytic in $z\in\mathscr{C}\backslash[0,\infty)$.

$\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta=\bigoplus_{m=1}^{N_\delta}(\psi_\delta^m,\cdot)\psi_\delta^m\otimes e^{-\hat{a}|y_\delta|}[-(2n_\delta)^{-1}\Delta_{y_\delta}-(z-e_\delta^m)]^{-1}e^{-\hat{a}|y'_\delta|}$.

$\lambda_0<e_\delta^1:\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ is norm analytic in $(z-\lambda_0)$ near zero.

$\lambda_0>e_\delta^1$, $\lambda_0\neq e_\delta^m$, $m=2,\cdots,N_\delta:\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ is norm analytic in $(z-\lambda_0)$ near zero.

$\lambda_0\geqq e_\delta^1$, $\lambda_0=e_\delta^m$ for some $m=1,\cdots,N_\delta:\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ is norm analytic in $(z-\lambda_0)^{1/2}$ near zero.

(c) $D_{10,\gamma\delta}$ is independent of z.

(d) $D_{11,\gamma\delta}(z)=[\hat{\rho}_\gamma^{-1}E_\gamma e^{-b|x_\delta|}|u_\delta|^{1/2}\hat{\rho}_\delta^{-1}][u_\delta^{1/2}E_\delta]\hat{\rho}_\delta E_\delta G_\delta(z)\hat{\rho}_\delta$ and hence (b) applies again.

Altogether $\overset{\Delta}{D}(\lambda+i0)$ is norm analytic in $(\lambda-\lambda_0)^{1/2}$ near zero for all $\lambda_0<0$. □

LEMMA 6.3. *Assume* H(III), $\alpha\neq0$, $\beta\neq0$ *and* $\sup(e_\alpha^j,e_\beta^l)\notin\overset{\Delta}{\mathscr{E}}$. *Then* $R_{\beta\alpha}^{lj}(\lambda)\in\mathscr{B}_1(L^2(S^2))$ *and* $R_{\beta\alpha}^{lj}(\lambda)$ *is continuous in* $\mathscr{B}_1$*-norm for* $\lambda\in[\sup(e_\alpha^j,e_\beta^l),\infty)\backslash\overset{\Delta}{\mathscr{E}}$. *Moreover* $R_{\beta\alpha}^{lj}(\lambda)$ *has the following threshold expansions valid in* $\mathscr{B}_1(L^2(S^2))$*-norm when* $|\lambda-\sup(e_\alpha^j,e_\beta^l)|$ *is small enough*:

Elastic case: $\alpha=\beta, j=l$.

$$R_{\alpha\alpha}^{jj}(\lambda)=\sum_{n=1}^{\infty}R_{\alpha\alpha,n}^{jj}(\lambda-e_\alpha^j)^{n/2}. \tag{6.1}$$

Inelastic case: $\alpha=\beta$, $j\neq l$ (*i.e.* $\psi_\alpha^j\neq$ const. ψ_α^l).

$$\begin{aligned}R_{\alpha\alpha}^{lj}(\lambda)&=[\lambda-\sup(e_\alpha^j,e_\alpha^l)]^{1/4}\sum_{n=0}^{\infty}R_{\alpha\alpha,n}^{lj}[\lambda-\sup(e_\alpha^j,e_\alpha^l)]^{n/2}, \qquad e_\alpha^j\neq e_\alpha^l,\\ R_{\alpha\alpha}^{lj}(\lambda)&=\sum_{n=1}^{\infty}\tilde{R}_{\alpha\alpha,n}^{lj}(\lambda-e_\alpha^j)^{n/2}, \qquad e_\alpha^j=e_\alpha^l.\end{aligned} \tag{6.2}$$

Rearrangement case: $\alpha\neq\beta$.

$$\begin{aligned}R_{\beta\alpha}^{lj}(\lambda)&=[\lambda-\sup(e_\alpha^j,e_\beta^l)]^{1/4}\sum_{n=0}^{\infty}R_{\beta\alpha,n}^{lj}[\lambda-\sup(e_\alpha^j,e_\alpha^l)]^{n/2}, \qquad e_\alpha^j\neq e_\beta^l,\\ R_{\beta\alpha}^{lj}(\lambda)&=\sum_{n=1}^{\infty}\tilde{R}_{\beta\alpha,n}^{lj}(\lambda-e_\alpha^j)^{n/2}, \qquad e_\alpha^j=e_\beta^l.\end{aligned} \tag{6.3}$$

Proof. (a) ${}^{(1)}R_{\beta\alpha}^{lj}(\lambda)$: The elastic and inelastic cases directly follow from Lemma 6.1. The rearrangement case is treated as follows:

$${}^{(1)}R_{\beta\alpha}^{lj}(\lambda)=-2\pi i\sum_{\substack{\gamma=1\\ \gamma\neq\alpha}}^{3}M_\beta^l(|v_\gamma|^{1/2},\lambda)M_\alpha^j(v_\gamma^{1/2},\lambda)^*$$

and hence the summand $\gamma\neq\beta$ is treated identical to the (in)-elastic case. For $\gamma=\beta$ $(\gamma\neq\alpha)$ we use

$$M_\gamma^l(|v_\gamma|^{1/2},\lambda)M_\alpha^j(v_\gamma^{1/2},\lambda)^*=M_\gamma^l(\hat{\rho}_\gamma,\lambda)[E_\gamma^l|v_\gamma|^{1/2}]M_\alpha^j(\hat{\rho}_\gamma^{-1}v_\gamma^{1/2},\lambda)^* \tag{6.4}$$

and again Lemma 6.1.

(b) ${}^{(2)}R^{lj}_{\beta\alpha}(\lambda)$: Since $[1+\overset{\Delta}{D}(z)]^{-1}$ has been considered in Lemma 6.2 we concentrate on the "*MJ*" and "*KM*"-terms:

"*MJ*":

$$\sum_{\substack{\gamma=1\\ \gamma\neq\beta}}^{3} M^l_\beta(\hat\rho_\gamma^{-1}|v_\gamma|^{1/2},\lambda)\{\hat\rho_\gamma g_{0,\gamma}+[v_\gamma^{1/2}E_\gamma]\hat\rho_\gamma E_\gamma G_\gamma(\lambda+i0)\hat\rho_\gamma g_{1,\gamma}\},\qquad g=\begin{pmatrix}g_0\\ g_1\end{pmatrix}\in\overset{\Delta}{\mathscr{H}}$$

and we need only to apply Lemma 6.1 and the proof of Lemma 6.2 to obtain the corresponding $\mathscr{B}_2$-expansions.

"*KM*": $\delta\neq\alpha$.

$$K_{1,\gamma\delta}M^j_\alpha(v_\delta\rho_\delta^{-1},\lambda)^*=[\hat\rho_\gamma^{-1}E_\gamma\rho_\delta][E_\gamma|v_\delta|^{1/2}]M^j_\alpha(\rho_\delta^{-1}v_\delta^{1/2},\lambda)^*,$$

$$(K_{0,\gamma\delta}\rho_\delta^{-1}|v_\delta|^{1/2})(\lambda+i0)M^j_\alpha(v_\delta^{1/2},\lambda)^*=v_\gamma^{1/2}(1-E_\gamma)G_\gamma(\lambda+i0)|v_\delta|^{1/2}M^j_\alpha(v_\delta^{1/2},\lambda)^*$$

and we apply again Lemma 6.1 since $v_\gamma^{1/2}(1-E_\gamma)G_\gamma(z)|v_\delta|^{1/2}$ is analytic in $z\in\mathscr{C}\backslash[0,\infty)$. □

Given Lemmas 6.1-6.3, we are able to formulate the following.

THEOREM 6.4. *Assume* H(III), $\alpha\neq 0$, $\beta\neq 0$ *and* $\sup(e^j_\alpha,e^l_\beta)\notin\overset{\Delta}{\mathscr{E}}$. *Then the averaged total cross section* $\bar\sigma^{lj}_{\beta\alpha}(\lambda)$ *is continuous in* $\lambda\in(\sup(e^j_\alpha,e^l_\beta),\infty)\backslash\overset{\Delta}{\mathscr{E}}$ *and has the following Taylor* (*resp. Laurent*) *expansions near the threshold* $\sup(e^j_\alpha,e^l_\beta)$:

Elastic case: $\alpha=\beta, j=l$.

$$\bar\sigma^{jj}_{\alpha\alpha}(\lambda)=\sum_{n=0}^{\infty}\bar\sigma^{jj}_{\alpha\alpha,n}(\lambda-e^j_\alpha)^{n/2},\ \bar\sigma^{jj}_{\alpha\alpha,0}=4\pi|A^{jj}_{\alpha\alpha}|^2.\tag{6.5}$$

Inelastic case: $\alpha=\beta$, $j\neq l$ (*i.e.* $\psi^j_\alpha\neq$ const. ψ^l_α).

$$\begin{aligned}\bar\sigma^{lj}_{\alpha\alpha}(\lambda)&=\sum_{n=-1}^{\infty}\bar\sigma^{lj}_{\alpha\alpha,n}(\lambda-e^j_\alpha)^{n/2}, && e^j_\alpha>e^l_\alpha,\\ \bar\sigma^{lj}_{\alpha\alpha}(\lambda)&=\sum_{n=1}^{\infty}\hat{\bar\sigma}^{lj}_{\alpha\alpha,n}(\lambda-e^l_\alpha)^{n/2}, && e^j_\alpha<e^l_\alpha,\\ \bar\sigma^{lj}_{\alpha\alpha}(\lambda)&=\sum_{n=0}^{\infty}\tilde{\bar\sigma}^{lj}_{\alpha\alpha,n}(\lambda-e^j_\alpha)^{n/2}, && e^j_\alpha=e^l_\alpha.\end{aligned}\tag{6.6}$$

Rearrangement case: $\alpha\neq\beta$.

$$\begin{aligned}\bar\sigma^{lj}_{\beta\alpha}(\lambda)&=\sum_{n=-1}^{\infty}\bar\sigma^{lj}_{\beta\alpha,n}(\lambda-e^j_\alpha)^{n/2}, && e^j_\alpha>e^l_\beta,\\ \bar\sigma^{lj}_{\beta\alpha}(\lambda)&=\sum_{n=1}^{\infty}\hat{\bar\sigma}^{lj}_{\beta\alpha,n}(\lambda-e^l_\beta)^{n/2}, && e^j_\alpha<e^l_\beta,\\ \bar\sigma^{lj}_{\beta\alpha}(\lambda)&=\sum_{n=0}^{\infty}\tilde{\bar\sigma}^{lj}_{\beta\alpha,n}(\lambda-e^j_\alpha)^{n/2}, && e^j_\alpha=e^l_\beta.\end{aligned}\tag{6.7}$$

Proof. All expansions directly follow from Lemma 6.3. Since without loss of generality we may assume $u_\alpha\in L^1(\mathscr{R}^3)$, $\alpha=1,2,3$ (otherwise $v_\alpha(x_\alpha)=e^{-b'|x_\alpha|}e^{-(b-b')|x_\alpha|}u_\alpha(x_\alpha)\equiv e^{-b'|x_\alpha|}u'_\alpha(x_\alpha)$, $0<b'<b$, $u'_\alpha\in L^1(\mathscr{R}^3)\cap L^p(\mathscr{R}^3)$ for some $p>3/2$ and we need only to replace (b,u_α) by (b',u'_α), $\alpha=1,2,3$) Definition 5.4 applies and yields the assertion for $\bar\sigma^{jj}_{\alpha\alpha,0}$ in terms of $A^{jj}_{\alpha\alpha}$. □

Our final result reads

THEOREM 6.5. *Assume* H(III), $\alpha \neq 0$, $\beta \neq 0$ *and* $\sup(e_\alpha^j, e_\beta^l) \notin \overset{\Delta}{\mathscr{E}}$. *Then the scattering amplitude* $f_{\beta\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\beta^l)$ *is continuous in* $(\lambda, \omega_\alpha^j, \omega_\beta^l) \in \{(\sup(e_\alpha^j, e_\beta^l), \infty) \backslash \overset{\Delta}{\mathscr{E}}\} \times S^2 \times S^2$ *and has the following Taylor* (*resp. Laurent*) *expansion near the threshold* $\sup(e_\alpha^j, e_\beta^l)$:

Elastic case: $\alpha = \beta, j = l$.

$$(6.8)\quad f_{\alpha\alpha}^{jj}(\lambda, \omega_\alpha^j \to \hat{\omega}_\alpha^j) = \sum_{n=0}^{\infty} f_{\alpha\alpha,n}^{jj}(\omega_\alpha^j \to \hat{\omega}_\alpha^j)(\lambda - e_\alpha^j)^{n/2}, \qquad f_{\alpha\alpha,0}^{jj}(\omega_\alpha^j \to \hat{\omega}_\alpha^j) = -A_{\alpha\alpha}^{jj}.$$

Inelastic case: $\alpha = \beta$, $j \neq l$ (*i.e.* $\psi_\alpha^j \neq \text{const.}\ \psi_\alpha^l$).

$$f_{\alpha\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\alpha^l) = (\lambda - e_\alpha^j)^{-1/4} \sum_{n=0}^{\infty} f_{\alpha\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\alpha^l)(\lambda - e_\alpha^j)^{n/2}, \qquad e_\alpha^j > e_\alpha^l,$$

$$(6.9)\quad f_{\alpha\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\alpha^l) = (\lambda - e_\alpha^l)^{1/4} \sum_{n=0}^{\infty} \hat{f}_{\alpha\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\alpha^l)(\lambda - e_\alpha^l)^{n/2}, \qquad e_\alpha^j < e_\alpha^l,$$

$$f_{\alpha\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\alpha^l) = \sum_{n=0}^{\infty} \tilde{f}_{\alpha\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\alpha^l)(\lambda - e_\alpha^j)^{n/2}, \qquad e_\alpha^j = e_\alpha^l.$$

Rearrangement case: $\alpha \neq \beta$.

$$f_{\beta\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\beta^l) = (\lambda - e_\alpha^j)^{-1/4} \sum_{n=0}^{\infty} f_{\beta\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\beta^l)(\lambda - e_\alpha^j)^{n/2}, \qquad e_\alpha^j > e_\beta^l,$$

$$(6.10)\quad f_{\beta\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\beta^l) = (\lambda - e_\beta^l)^{1/4} \sum_{n=0}^{\infty} \hat{f}_{\beta\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\beta^l)(\lambda - e_\beta^l)^{n/2}, \qquad e_\alpha^j < e_\beta^l,$$

$$f_{\beta\alpha}^{lj}(\lambda, \omega_\alpha^j \to \omega_\beta^l) = \sum_{n=0}^{\infty} \tilde{f}_{\beta\alpha,n}^{lj}(\omega_\alpha^j \to \omega_\beta^l)(\lambda - e_\alpha^j)^{n/2}, \qquad e_\alpha^j = e_\beta^l.$$

Proof. The proof is a direct consequence of Theorems 5.1-5.3 and Lemmas 6.1-6.3.

Given the expansion (6.8) one can define higher-order low energy parameters such as the elastic effective range parameter similar to the two-body case (cf. [1]). We omit the details.

Acknowledgments. We are indebted to Profs. D. Bollé, L. Streit and H. Zingl for many helpful suggestions. We gratefully acknowledge the kind hospitality and stimulating atmosphere at Zentrüm fur interdisziplinäre Forschung, Universität Bielefeld, during Project No. 2: Mathematics and Physics.

REFERENCES

[1] S. ALBEVERIO, D. BOLLÉ, F. GESZTESY, R. HØEGH-KROHN AND L. STREIT, *Low-energy parameters in nonrelativistic scattering theory*, Ann. Physics, 148 (1983), pp. 308-326.

[2] S. ALBEVERIO, F. GESZTESY AND R. HØEGH-KROHN, *The low energy expansion in nonrelativistic scattering theory*, Ann. Inst. H. Poincaré: A37 (1982), pp. 1-28.

[3] W. O. AMREIN, J. M. JAUCH AND K. B. SINHA, *Scattering Theory in Quantum Mechanics*, W. A. Benjamin, Reading, MA, New York, 1977.

[4] W. O. AMREIN, D. B. PEARSON AND K. B. SINHA, *Bounds on the total scattering cross-section for N-body systems*, Nuovo Cimento, 52A (1979), pp. 115-131.

[5] W. O. Amrein and K. B. Sinha, *On three-body scattering cross sections*, J. Phys. A, 15 (1982), pp. 1567-1586.

[6] E. Balslev, *Analytic scattering theory of quantum mechanical three-body systems*, Ann. Inst. H. Poincaré. A32 (1980), pp. 125-160, and Aarhus Univ. Preprint Series No. 26, 1978/79.

[7] ———, *Resonances in three-body scattering theory*, Adv. in Appl. Math., 5 (1984), pp. 260-285.

[8] D. Bollé and T. A. Osborn, *Spectral sum rules for the three-body problem*, Phys. Rev. A , 26 (1982), pp. 3062-3072.

[9] V. S. Buslaev and S. P. Merkuriev, *Trace equation for a three particle system*, Soviet Phys. Dokl., 14 (1970), pp. 1055-1057.

[10] J. M. Combes, *Analytic perturbation approach to N-particle quantum systems*, in Scattering Theory in Mathematical Physics, J. A. La Vita and J. P. Marchand, eds., Reidel, Dordrecht, 1974, pp. 243-272.

[11] M. Combescure and J. Ginibre, *On the negative point spectrum of quantum mechanical three-body systems*, Ann. Physics, 101 (1976), pp. 355-379.

[12] V. Enss, *Completeness of three body quantum scattering*, in Bielefeld Encounters in Physics and Mathematics III, Springer Lecture Notes in Mathematics 1031, Ph. Blanchard and L. Streit, eds., Springer-Verlag, Berlin, New York, 1983, pp. 62-88.

[13] ———, *Scattering and spectral theory for three particle systems*, Proc. of the International Conference on Differential Equations held at the Univ. of Alabama, Birmingham, I. Knowles and R. Lewis, eds., North-Holland, Amsterdam.

[14] L. D. Faddeev, *Mathematical Aspects of the Three-Body Problem in the Quantum Scattering Theory*, Israel Program for Scientific Translations, Jerusalem, 1965.

[15] F. Gesztesy and G. Karner, *Scattering lengths in nonrelativistic three-body systems*, in Few Body Problems in Physics II, B. Zeitnitz, ed., Elsevier, Amsterdam, 1984, pp. 375-376.

[16] J. Ginibre and M. Moulin, *Hilbert space approach to the quantum mechanical three-body problem*, Ann. Inst. H. Poincaré. A21 (1974), pp. 97-145.

[17] G. A. Hagedorn, *Asymptotic completeness for classes of two, three and four particle Schrödinger operators*, Trans. Amer. Math. Soc., 258 (1980), pp. 1-75.

[18] ———, *A link between scattering resonances and dilation analytic resonances in few body quantum mechanics*, Comm. Math. Phys., 65 (1979), pp. 181-188.

[19] ———, *Born series for* (2 *cluster*) → (2 *cluster*) *scattering of two-, three- and four-particle Schrödinger operators*, Comm. Math. Phys., 66 (1979), pp. 77-94.

[20] G. A. Hadedorn and P. A. Perry, *Asymptotic completeness for certain three-body Schrödinger operators*, Comm. Pure Appl. Math., 36 (1983), pp. 213-232.

[21] J. S. Howland, *Abstract stationary theory of multichannel scattering*, J. Funct. Anal., 22 (1976), pp. 250-282.

[22] T. Kato, *Wave operators and similarity for some non-selfadjoint operators*, Math. Ann., 162 (1966), pp. 258-279.

[23] S. T. Kuroda, *An abstract stationary approach to perturbation of continuous spectra and scattering theory*, J. Anal. Math., 20 (1967), pp. 57-117.

[24] M. Loss and I. M. Sigal, *The three-body problem with theshold singularities*, ETH-Zürich, preprint, 1982.

[25] E. Mourre, *Applications de la méthode de Lavine au problème à trois corps*, Ann. Inst. H. Poincaré. A26 (1977), pp. 219-262.

[26] ———, *Absence of singular continuous spectrum for certain self-adjoint operators*, Comm. Math. Phys., 78 (1981), pp. 391-408.

[27] T. A. Osborn and D. Bollé, *Primary singularities, asymptotic wave functions and unitarity in the three-body problem*, Phys. Rev. C8 (1973), pp. 1198-1210.

[28] G. L. Payne, J. L. Friar and B. F. Gibson, *Configuration space Faddeev continuum calculations*. I. *n-d scattering length*, Phys. Rev. C26 (1982), pp. 1385-1398.

[29] P. Perry, I. M. Sigal and B. Simon, *Spectral analysis of N-body Schrödinger operators*, Ann. Math., 114 (1981), pp. 519-567.

[30] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* III: *Scattering Theory*, Academic Press, New York, 1979.

[31] ———, *Methods of Modern Mathematical Physics* IV: *Analysis of Operators*, Academic Press, New York, 1978.

[32] F. Riahi, *On the analyticity properties of the N-body scattering amplitude in non-relativistic quantum mechanics*, Helv. Phys. Acta, 42 (1969), pp. 299-329.

[33] I. M. Sigal, *Scattering theory in many-body quantum systems. Analyticity of the scattering matrix*, in Quantum Mechanics in Mathematics, Chemistry and Physics, K. Gustafson and W. P. Reinhardt, eds., Plenum, New York, 1981, pp. 307-336.

[34] I. M. SIGAL, *Scattering Theory for Many-Body Quantum Mechanical Systems*, Lecture Notes in Math., 1011, Springer, Berlin, New York, 1983.

[35] ———, *Asymptotic completeness of many-body short-range systems*, Lett. Math. Phys., 8 (1984), pp. 181-188.

[36] B. SIMON, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms*, Princeton Univ. Press, Princeton, NJ, 1971.

[37] ———, *Scattering theory and quadratic forms: On a theorem of Schechter*, Comm. Math. Phys., 53 (1977), pp. 151-153.

[38] K. B. SINHA, M. KRISHNA AND PL. MUTHURAMALINGAM, *On the completeness in three body scattering*, Ann. Inst. H. Poincaré. A41 (1984), pp. 79-101.

[39] W. THIRRING, *A Course in Mathematical Physics* 3: *Quantum Mechanics of Atoms and Molecules*, Springer, New York, 1981.

[40] L. E. THOMAS, *Asymptotic completeness in two- and three-particle quantum mechanical scattering*, Ann. Physics, 90 (1975), pp. 127-165.

[41] K. YAJIMA, *An abstract stationary approach to three-body scattering*, J. Fac. Sci. Univ. Tokyo, Sect. 1A Math., 25 (1978), pp. 109-132.

LIE THEORY OF SOLUTIONS OF CERTAIN DIFFERINTEGRAL EQUATIONS*

M. A. AL-BASSAM† AND H. L. MANOCHA‡

Abstract. A new Lie algebraic technique based on the idea of fractional differentiation is evolved for constructing new models of representations of the Lie algebra $sl(2, C)$. Some of the Lie algebra elements in the new models are differintegral (integral-differential) operators while the basis vectors turn out to be solutions of differintegral equations

Key words. Lie algebra $sl(2, C)$, multiplier representation, generating functions, fractional derivative, Lie group $SL(2, C)$

AMS(MOS) subject classification. 33A75

1. Introduction. In [7], irreducible representations of the Lie algebras $sl(2, C)$, the oscillator algebra and the algebra of the Euclidean group in the plane have been determined, and models of these representations have been constructed in terms of first order differential operators (the operator types A, B, C', C'', D') acting on spaces of functions of two complex variables z, t. The basis vectors $f_m(z, t) = Z_m(z)t^m$ of such irreducible representations have turned out to be such that $Z_m(z)$ are functions of hypergeometric type. This connection between Lie algebras and special functions has led to recurrence relations, differential equations, generating functions and addition theorems for the functions of hypergeometric type. Later, in [8], the type A, $\cdots$, D$'$ operators have been used as building blocks to construct more complicated models of irreducible representations of these Lie algebras. Some of the Lie algebra elements in these models are second order differential operators. Furthermore, the models have been constructed in such a way that the basis vectors turn out to be special functions satisfying second order nonhomogeneous differential equations.

In this paper we propose to use type A and type B operators as building blocks for constructing new models of irreducible representations of the Lie algebras $sl(2, C)$. The method that we employ for this construction will be based on the idea of fractional differentiation. These new models will be such that somc of the Lie algebra elements in them are differintegral operators, while the basis vectors turn out to be solutions of differintegral equations.

In § 2, we define the fractional (or generalized) derivative of order α of an analytic function $f(z)$, written as $D_z^\alpha f(z)$, and then make use of this to express fractional derivative representations of some special functions which we need in our discussion.

In § 3, we discuss the generalized Leibnitz rule for the fractional derivative of the product of two analytic functions. We introduce a new operator $\mathscr{D}$, defined as $\mathscr{D}f(z) = z^{1-\lambda}D_z^{\mu-\lambda}f(z)$, and thereafter make use of the rule for obtaining operator expressions for $\mathscr{D}^{-1}L\mathscr{D}$, $L = z$, $z(d/dz)$, $z^2(d/dz)$. The new expressions turn out to be differintegral operators.

In §§ 4-6, we consider the type A and type B representations of $sl(2, C)$ [7, Chap. 5]. The basis functions corresponding to these representations are in terms of ${}_2F_1$ and ${}_1F_1$, respectively. Through $\mathscr{D}^{-1}L\mathscr{D}$, the type A,B operators induce new sets of Lie algebra operators, the corresponding basis vectors turning out in terms of ${}_3F_2$ and ${}_2F_2$, respectively. This leads to the construction of new models of representations of $sl(2, C)$.

* Received by the editors May 29, 1985; accepted for publication (in revised form) June 3, 1986.

† Department of Mathematics, University of Kuwait, Kuwait.

‡ Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India.

2. Fractional derivative representation. In 1731, L. E. Euler considered the concept of fractional differentiation when he extended the familiar formula

$$\frac{d^n z^p}{dz^n} = p(p-1)(p-2)\cdots(p-n+1)z^{p-n} = \frac{p!}{(p-n)!}z^{p-n}$$

to $n=\alpha$, where α is as usual arbitrary, by writing

$$D_z^\alpha z^p = \frac{\Gamma(p+1)}{\Gamma(p-\alpha+1)} z^{p-\alpha}. \tag{2.1}$$

In fact, it was this formula which led Euler to invent the gamma function for fractional values of the factorial: $\Gamma(p+1)=p!$.

It immediately follows that if

$$g(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad |z|<R, \tag{2.2}$$

then, for $0<|z|<R$,

$$\begin{aligned} D_z^\alpha(z^p g(z)) &= \sum_{n=0}^{\infty} a_n D_z^\alpha z^{p+n} \\ &= \sum_{n=0}^{\infty} a_n \frac{\Gamma(p+n+1)}{\Gamma(p+n-\alpha+1)} z^{p+n-\alpha}. \end{aligned} \tag{2.3}$$

Relation (2.3) fails to have meaning when p is a negative integer. Using (2.3), we list the following results:

$$D_z^{\lambda-\mu}\left\{z^{\lambda-1}\,{}_1F_1\left[\begin{matrix}\alpha; \\ \beta;\end{matrix}\; z\right]\right\} = \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1}\,{}_2F_2\left[\begin{matrix}\alpha, & \lambda; \\ \beta, & \mu;\end{matrix}\; z\right], \qquad 0<|z|<\infty, \tag{2.4}$$

$$D_z^{\lambda-\mu}\left\{z^{\lambda-1}\,{}_2F_1\left[\begin{matrix}\alpha, \beta; \\ \gamma;\end{matrix}\; z\right]\right\} = \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1}\,{}_3F_2\left[\begin{matrix}\alpha, \beta, \lambda; \\ \gamma, \mu;\end{matrix}\; z\right], \qquad 0<|z|<1, \tag{2.5}$$

$$\begin{aligned} D_z^{\lambda-\mu}\left\{z^{\lambda-1}e^{az}\,{}_1F_1\left[\begin{matrix}\alpha; \\ \beta;\end{matrix}\; bz\right]\right\} &= \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1} F^{1:1;0}_{1:1;0}\left(\begin{matrix}\lambda: \alpha; -; \\ \mu: \beta; -;\end{matrix}\; az, bz\right), \\ &\qquad 0<|az|, \qquad |bz|<\infty, \end{aligned} \tag{2.6}$$

$$\begin{aligned} &D_z^{\lambda-\mu}\left\{z^{\lambda-1}(1-bz)^{-\alpha}2F_1\left[\begin{matrix}\alpha, \beta; \\ \gamma;\end{matrix}\; \frac{az}{1-bz}\right]\right\} \\ &= \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1} F^{2:1;0}_{1:1;0}\left(\begin{matrix}\alpha, \lambda: \beta; —; \\ \mu: \gamma; —;\end{matrix}\; az, bz\right), \\ &\qquad\qquad 0<(|a|+|b|)|z|<1, \end{aligned} \tag{2.7}$$

$$\begin{aligned} D_z^{\lambda-\mu}\left\{z^{\lambda-1}F_2(\alpha; \beta, \beta'; \gamma, \gamma'; z, t)\right\} &= \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1} F^{1:2;1}_{0:2;1}\left(\begin{matrix}\alpha: \beta, \lambda; \beta'; \\ -: \gamma, \mu; \gamma';\end{matrix}\; z, t\right), \\ &\qquad 0<|z|+|t|<1, \end{aligned} \tag{2.8}$$

$$\begin{aligned} &D_z^{\lambda-\mu}\left\{z^{\lambda-1}e^z\psi_2[\alpha; \gamma, \gamma'; -z, t]\right\} \\ &= \frac{\Gamma(\lambda)}{\Gamma(\mu)} z^{\mu-1} G^{1:1;2}_{0:2;1}\left(\begin{matrix}\gamma-\alpha: \lambda; \alpha, 1+\alpha-\gamma; \\ -: \mu, \gamma; \gamma';\end{matrix}\; z, -t\right), \\ &\qquad\qquad 0<|z|<\infty, \quad |t|<\infty, \end{aligned} \tag{2.9}$$

$$D_z^{\lambda-\mu}\left\{z^{\lambda-1}e^z\psi_2[\alpha;\alpha,\beta;-z,t]\right\}=\frac{\Gamma(\lambda)}{\Gamma(\mu)}z^{\mu-1}F^{1:1;0}_{1:2;0}\left(\begin{matrix}\alpha:\lambda;—;\\ \beta:\alpha,\mu;—;\end{matrix}\quad -zt,t\right),\tag{2.10}$$

$$0<|z|<\infty,\quad |t|<\infty,$$

$$\begin{aligned}&D_z^{\lambda-\mu}\left\{z^{\lambda-1}(1-z)^{-\beta}F_2\left[\alpha;\beta,\beta';\alpha,\gamma;\frac{-z}{1-z},t\right]\right\}\\&=\frac{\Gamma(\lambda)}{\Gamma(\mu)}z^{\mu-1}F^{2:2;0}_{1:2;0}\left(\begin{matrix}\alpha,\alpha':\beta,\lambda;—;\\ \beta':\alpha,\mu;—;\end{matrix}\quad -zt,t\right),\end{aligned}\tag{2.11}$$

$$0<|z|<1,\quad |zt|+|t|<1,$$

where [10],

$${}_pF_q\left[\begin{matrix}\alpha_1,\cdots,\alpha_p;\\ \beta_1,\cdots,\beta_q;\end{matrix}\quad z\right]=\sum_{n=0}^{\infty}\frac{(\alpha_1)_n\cdots(\alpha_p)_n}{(\beta_1)_n\cdots(\beta_q)_n}\frac{z^n}{n!},\tag{2.12}$$

$$|z|<\infty\quad \text{if } p\leqq q;\quad |z|<1\quad \text{if } p=q+1,$$

$$\psi_2(\alpha;\beta,\gamma;z,t)=\sum_{m,n=0}^{\infty}\frac{(\alpha)_{m+n}}{(\beta)_m(\gamma)_n}\frac{z^mt^n}{m!n!},\qquad 0\leqq|z|<\infty,\quad 0\leqq|t|<\infty,\tag{2.13}$$

$$F_2[\alpha;\beta,\beta';\gamma,\gamma';z,t]=\sum_{m,n=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_m(\beta')_n}{(\gamma)_m(\gamma')_n}\frac{z^mt^n}{m!n!},\qquad |z|+|t|<1,\tag{2.14}$$

$$\begin{aligned}&F^{A:B;B'}_{C:E;E'}\left(\begin{matrix}(a):(b);(b');\\ (c):(e);(e');\end{matrix}\quad z,t\right)\\&=\sum_{m,n=0}^{\infty}\frac{\prod_{j=1}^{A}(a_j)_{m+n}\prod_{j=1}^{B}(b_j)_m\prod_{j=1}^{B'}(b_j')_n}{\prod_{j=1}^{C}(c_j)_{m+n}\prod_{j=1}^{E}(e_j)_m\prod_{j=1}^{E'}(e_j')_n}\frac{z^mt^n}{m!n!},\end{aligned}\tag{2.15}$$

$$\begin{aligned}&G^{A:B;B'}_{C:E;E'}\left(\begin{matrix}(a):(b);(b');\\ (c):(e);(e');\end{matrix}\quad z,t\right)\\&=\sum_{m,n=0}^{\infty}\frac{\prod_{j=1}^{A}(a_j)_{m-n}\prod_{j=1}^{B}(b_j)_m\prod_{j=1}^{B'}(b_j')_n}{\prod_{j=1}^{C}(c_j)_{m-n}\prod_{j=1}^{E}(e_j)_m\prod_{j=1}^{E'}(e_j')_n}\frac{z^mt^n}{m!n!}.\end{aligned}\tag{2.16}$$

3. Generalized Leibnitz rule. Consider the Leibnitz rule from elementary calculus for the derivative of the product of two functions $u(z)$ and $v(z)$:

$$D_z^Nuv=\sum_{n=0}^{N}\binom{N}{n}D_z^{N-n}uD_z^nv.$$

A reasonable guess for the generalization of this result to fractional derivatives is

$$D_z^\alpha uv=\sum_{n=0}^{\infty}\binom{\alpha}{n}D_z^{\alpha-n}uD_z^nv.\tag{3.1}$$

This guess is indeed correct and it was given as early as 1867 by A. K. Grunwald [2]. It has been proved by Al-Bassam [1] and Osler [9] in a manner different from that of Grunwald.

In order to construct new models of representations, we introduce, for convenience's sake, operators $\mathcal{D}$ and $\mathcal{D}^{-1}$ defined as

$$\mathcal{D}f(z)=z^{1-\lambda}D_z^{\mu-\lambda}f(z),\tag{3.2}$$

$$\mathcal{D}^{-1}f(z)=D_z^{\lambda-\mu}z^{\lambda-1}f(z).\tag{3.3}$$

Indeed

$$\mathscr{D}\mathscr{D}^{-1}f(z)=\mathscr{D}^{-1}\mathscr{D}f(z)=f(z). \tag{3.4}$$

Using (3.1), one arrives at the following:

$$\mathscr{D}^{-1}z\mathscr{D}=z+(\lambda-\mu)\frac{d^{-1}}{dz^{-1}}, \tag{3.5}$$

$$\mathscr{D}^{-1}\left(z\frac{d}{dz}\right)\mathscr{D}=z\frac{d}{dz}+1-\mu, \tag{3.6}$$

$$\mathscr{D}^{-1}\left(z^2\frac{d}{dz}\right)\mathscr{D}=z^2\frac{d}{dz}+(1+\lambda-2\mu)z-\mu(\lambda-\mu)\frac{d^{-1}}{dz^{-1}}. \tag{3.7}$$

Note that d^{-1}/dz^{-1} is an indefinite integral in disguise.

4. Representation $D(u, m_0)$. Consider the representation $D(u, m_0)$ defined for u, $m_0\in C$ such that $0\leqq \operatorname{Re} m_0<1$ and $u\pm m_0$ are not integers [7, Chap. 5]. The representation space W has a basis $\{f_m\}$, $m\in S=\{m_0+n\colon n=0,\pm1,\cdots\}$, such that the action of $sl(2, C)$ on W is given by

$$\begin{gathered} J^+f_m=(m-u)f_{m+1},\quad J^-f_m=-(m+u)f_{m-1},\quad J^3f_m=mf_m,\\ (J^+J^-+J^3J^3-J^3)f_m=u(u+1)f_m. \end{gathered} \tag{4.1}$$

The operators $\{J^+, J^-, J^3\}$ satisfy the commutation relations

$$[J^3, J^\pm]=\pm J^\pm,\qquad [J^+, J^-]=2J^3, \tag{4.2}$$

and as such generate a Lie algebra which is an isomorphic image of the complex Lie algebra $sl(2, C)$ [7, p. 7].

4.1. Type A operators. The type A operators $\{J^+, J^-, J^3\}$ satisfying (4.1), and indeed (4.2), are

$$\begin{aligned} J^+&=t\left(z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-u\right),\\ J^-&=t^{-1}\left[z(1-z)\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}+z(q+u)-u\right],\\ J^3&=t\frac{\partial}{\partial t}, \end{aligned} \tag{4.3}$$

and

$$f_m(z,t)={}_2F_1\left[\begin{matrix} m-u, -q-u;\\ -2u;\end{matrix}\ z\right]t^m. \tag{4.4}$$

The multiplier representation T induced by the operators (4.3) on $\mathscr{F}$, the space of all analytic functions in a neighbourhood of (z_0, t_0), is

$$\begin{aligned}{}[T(g)f](z,t)&=(d+bt)^u\left(a+\frac{c}{t}\right)^{-q}\left(a-\frac{c(z-1)}{t}\right)^{q+u}\\ &\quad\times f\left(\frac{zt}{(d+bt)(at-c(z-1))},\frac{c+at}{d+bt}\right),\end{aligned} \tag{4.5}$$

$|bt/d|<1$, $|c/at|<1$, $|c(z-1)/at|<1$, $-\pi<\arg a$, $\arg d<\pi$, and

$$g=\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad ad\text{-}bc=1,$$

lies in a sufficiently small neighbourhood of the identity element

$$e=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \varepsilon SL(2, C).$$

Now we use the type A operators as building blocks for constructing a new model of irreducible representations of $sl(2, C)$.

To illustrate the method, we rewrite, with the help of (2.3),

$$J^+f_m=t\left(z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-u\right)f_m=(m-u)f_{m+1},$$

$$f_m(z, t)={}_2F_1\left[\begin{matrix} m-u, -q-u; \\ -2u; \end{matrix}\ z\right]t^m,$$

as

$$J^+\left\{z^{1-\lambda}\frac{\partial^{\mu-\lambda}}{\partial z^{\mu-\lambda}}\left(z^{\mu-1}{}_3F_2\left[\begin{matrix} m-u, -q-u, \lambda; \\ -2u, \mu; \end{matrix}\ z\right]t^m\right)\right\}$$

$$=(m-u)z^{1-\lambda}\frac{\partial^{\mu-\lambda}}{\partial z^{\mu-\lambda}}\left(z^{\mu-1}{}_3F_2\left[\begin{matrix} 1+m-u, -q-u, \lambda; \\ -2u, \mu; \end{matrix}\ z\right]t^m\right).$$

It follows that

(4.6) $$K^+h_m=(m-u)h_{m+1},$$

where, by (3.6),

(4.7) $$K^+=\mathscr{D}^{-1}J^+\mathscr{D}=t\left(z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}+1-\mu-u\right),$$

and

(4.8) $$h_m(z, t)=z^{\mu-1}{}_3F_2\left[\begin{matrix} m-u, -q-u, \lambda; \\ -2u, \mu; \end{matrix}\ z\right]t^m,$$

$$m=m_0+n, \quad n=0, \pm1, \pm2, \cdots.$$

Similarly,

(4.9) $$K^-h_m=-(m+u)h_{m-1}, \qquad K^3h_m=mh_m$$

where

(4.10) $$K^-=\mathscr{D}^{-1}J^-\mathscr{D}=t^{-1}\left[z(1-z)\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}+(q+u+2\mu-1-\lambda)z\right.$$
$$\left.+1-\mu-u+(\lambda-\mu)(q+u+\mu)\frac{\partial^{-1}}{\partial z^{-1}}\right],$$

(4.11) $$K^3=\mathscr{D}^{-1}J^3\mathscr{D}=t\frac{\partial}{\partial t}.$$

Indeed,

$$\begin{aligned}&[K^3, K^\pm]=\pm K^\pm, \qquad [K^+, K^-]=2K^3,\\ &(K^+K^-+K^3K^3-K^3)h_m=u(u+1)h_m.\end{aligned}\tag{4.12}$$

We have thus constructed a new model of irreducible representation $D(u, m_0)$ in terms of differintegral operators $\{K^+, K^-, K^3\}$ such that the vectors $h_m(z, t)$ defined by (4.8) form a basis of the representation space. Writing $h_m(z, t)=Z_m(z)t^m$, in terms of the functions $Z_m(z)$ the above relations take the form

$$\begin{aligned}&\left(z\frac{d}{dz}+1+m-\mu-u\right)Z_m(z)=(m-u)Z_{m+1}(z),\\ &\left[z(1-z)\frac{d}{dz}+(\lambda-\mu)(q+u+\mu)\frac{d^{-1}}{dz^{-1}}+(q+u+2\mu-\lambda-1)z+1-u-\mu-m\right]\\ &\quad\cdot Z_m(z)=-(m+u)Z_{m-1}(z),\\ &\left[z^2(1-z)\frac{d^2}{dz^2}+z\{(q+2u+3\mu-3-\lambda-m)z+2(1-\mu-u)\}\frac{d}{dz}\right.\\ &\qquad+(\lambda-\mu)(q+u+\mu)\frac{d^{-1}}{dz^{-1}}+z\\ &\qquad\left.\cdot\{(\lambda-\mu)(q+u+\mu)+(q+u+2\mu-\lambda-1)(1+m-u-\mu)\}\right]Z_m(z)\\ &\quad=(1-\mu)(\mu+2u)Z_m(z).\end{aligned}\tag{4.13}$$

The multiplier representation T' induced by the K-operators on $\mathcal{F}$ is

$$\begin{aligned}[T'(g)h](z, t)&=\mathcal{D}^{-1}\left[\exp\left(-\frac{b}{d}J^+\right)\exp(-cdJ^-)\exp(\tau J^3)(\mathcal{D}h)(z, t)\right.\\ &=\mathcal{D}^{-1}[T(g)(\mathcal{D}h)](z, t), \qquad e^{\tau/2}=d^{-1}.\end{aligned}\tag{4.14}$$

By (4.5) it follows that

$$\begin{aligned}[T'(g)h](z, t)=\mathcal{D}^{-1}\Bigg[&(d+bt)^u\left(a+\frac{c}{t}\right)^{-q}\left(a-\frac{c(z-1)}{t}\right)^{q+u}\\ &\times(\mathcal{D}h)\left(\frac{zt}{(d+bt)(at-c(z-1))}, \frac{c+at}{d+bt}\right)\Bigg],\end{aligned}\tag{4.15}$$

$|bt/d|<1$, $|c/at|<1$, $|c(z-1)/at|<1$ and $g\in SL(2, C)$ lies in small enough neighbourhood of e so that the above expression is uniquely defined.

In [5, § 2] it has been shown that 16 of the Horn functions of two variables [3] can be realized in terms of type A operators. In particular, the function

$$F=F_2[m_0-u; -q-u, \beta'; -2u, \gamma'; z, t]t^{m_0}\tag{4.16}$$

satisfies the Casimir eigenvalue equation

$$(J^+J^-+J^3J^3-J^3)F=u(u+1)F\tag{4.17}$$

as well as

$$[J^3J^3-J^+J^3+(m_0-\beta')J^+-(2m_0-\gamma'+1)J^3]F=m_0(\gamma'-m_0-1)F.\tag{4.18}$$

Likewise, it can be shown that

$$G = t^{m_0}(1-t)^{u-m_0}\,{}_2F_1\left[\begin{matrix} m_0-u, -q-u; & \dfrac{z}{1-t} \\ -2u; & \end{matrix}\right] \tag{4.19}$$

is a simultaneous solution of

$$(J^+J^- + J^3J^3 - J^3)G = u(u+1)G \tag{4.20}$$

and

$$[J^3J^3 - J^+J^3 + m_0J^+ - (2m_0+1)J^3]G = -m_0(1+m_0)G. \tag{4.21}$$

Thus, by (2.5) and (2.8), it follows that

$$\mathscr{H}_1 = t^{m_0}z^{\mu-1}F^{1:2;1}_{0:2;1}\left(\begin{matrix} m_0-u: -q-u, \lambda; \beta'; \\ —: -2u, \mu; \gamma'; \end{matrix}\ z, t\right) \tag{4.22}$$

satisfies the differintegral equation

$$(K^+K^- + K^3K^3 - K^3)\mathscr{H}_1 = u(u+1)\mathscr{H}_1 \tag{4.23}$$

as well as

$$[K^3K^3 - K^+K^3 + (m_0-\beta')K^+ - (2m_0-\gamma'+1)K^3]\mathscr{H}_1 = m_0(\gamma'-m_0-1)\mathscr{H}_1, \tag{4.24}$$

while

$$\mathscr{H}_2 = t^{m_0}z^{\mu-1}(1-t)^{u-m_0}\,{}_3F_2\left[\begin{matrix} m_0-u, -q-u, \lambda; & \dfrac{z}{1-t} \\ -2u, \mu; & \end{matrix}\right] \tag{4.25}$$

is a common solution of both

$$(K^+K^- + K^3K^3 - K^3)\mathscr{H}_2 = u(u+1)\mathscr{H}_2 \tag{4.26}$$

and

$$[K^3K^3 - K^+K^3 + m_0K^+ - (2m_0+1)K^3]\mathscr{H}_0 = -m_0(1+m_0)\mathscr{H}_2. \tag{4.27}$$

Now, we have the expansions [6], [11],

$$T'(g)\mathscr{H}_i = \sum_{n=-\infty}^{\infty} k_{in}(g)h_{m_0+n}, \qquad i = 1, 2, \tag{4.28}$$

each valid in a region determined by the inequalities (4.15).

The identities which follow as special cases from (4.28), are as under:

$$\begin{aligned}(1-t)^{-\alpha}&F^{1:2;1}_{0:2;1}\left(\begin{matrix} \alpha: \beta, \lambda; \beta'; \\ —: \gamma, \mu; \gamma'; \end{matrix}\ \frac{z}{1-t}, \frac{-\omega t}{1-t}\right) \\ &= \sum_{n=0}^{\infty} \frac{(\alpha)_n}{n!}\,{}_3F_2\left[\begin{matrix} \alpha+n, \beta, \lambda; & z \\ \gamma, \mu; & \end{matrix}\right]{}_2F_1\left[\begin{matrix} -n, \beta'; & \omega \\ \gamma'; & \end{matrix}\right]t^n, \\ &\qquad |t|<1, \qquad \left|\frac{z}{1-t}\right| + \left|\frac{\omega t}{1-t}\right| < 1,\end{aligned} \tag{4.29}$$

$$\begin{aligned}\left(1+\frac{c}{t}\right)^{\alpha-\gamma}&(1-c-t)^{-\alpha}F^{2:1;0}_{1:1;0}\left(\begin{matrix} \beta, \lambda: \alpha; —; \\ \mu: \gamma; —; \end{matrix}\ \frac{zt}{(c+t)(1-c-t)}, \frac{cz}{c+t}\right) \\ &= \sum_{n=-\infty}^{\infty} \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\,{}_3F_2\left[\begin{matrix} \alpha+n, \beta, \lambda; & z \\ \gamma, \mu; & \end{matrix}\right]\frac{{}_2F_1(\alpha+n, 1+\alpha-\gamma-n; n+1; c)}{\Gamma(n+1)}t^n, \\ &\qquad \left|\frac{c}{t}\right| < 1, \quad |c+t| < 1, \quad \left|\frac{zt}{(c+t)(1-c-t)}\right| + \left|\frac{cz}{c+t}\right| < 1,\end{aligned} \tag{4.30}$$

$$(4.31)\qquad (1-t)^{-\alpha}{}_3F_2\left[\begin{matrix}\alpha,\beta,\lambda; & \dfrac{z}{1-t}\\ \gamma,\mu; & \end{matrix}\right]=\sum_{n=0}^{\infty}\frac{(\alpha)_n}{n!}{}_3F_2\left[\begin{matrix}\alpha+n,\beta,\lambda; & z\\ \gamma,\mu; & \end{matrix}\right]t^n,\quad |z|+|t|<1.$$

In (4.30), the terms corresponding to $n=-1,-2,-3,\cdots$ are well defined in view of the relation

$$(4.32)\qquad \begin{aligned}&\lim_{s\to-k}\frac{1}{\Gamma(1+s)}{}_2F_1\left[\begin{matrix}\alpha+s,1+\alpha-\gamma+s; & c\\ 1+s; & \end{matrix}\right]\\ &=\frac{(\alpha-k)_k(1+\alpha-\gamma-k)_k}{k!}c^k{}_2F_1\left[\begin{matrix}\alpha,1+\alpha-\gamma; & c\\ 1+k; & \end{matrix}\right].\end{aligned}$$

The above identities are valid for all α, β, β', γ, γ', λ, $\mu\in C$ such that α, $\alpha-\gamma$ are not integers and γ, γ', $\mu\neq 0,-1,-2,\cdots$.

4.2. Type B operators. The type B operators $\{J^+, J^-, J^3\}$ and the corresponding basis vectors $\{f_m\}$ satisfying (4.1) as well as (4.2) are

$$(4.33)\qquad \begin{aligned}J^+&=t\left(z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-z+u+1\right),\\ J^-&=t^{-1}\left(z\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}+u+1\right),\\ J^3&=t\frac{\partial}{\partial t},\end{aligned}$$

and

$$(4.34)\qquad f_m(z,t)=L^{(2u+1)}_{m-u-1}(z)t^m.$$

$L^{(2u+1)}_{m-u-1}(z)$ are generalized Laguerre functions defined as [10]

$$(4.35)\qquad L^{(\alpha)}_n(z)=\frac{\Gamma(1+\alpha+n)}{\Gamma(1+\alpha)\Gamma(1+n)}{}_1F_1\left[\begin{matrix}-n; & z\\ 1+\alpha; & \end{matrix}\right].$$

The multiplier representation induced by the J-operators on the space $\mathcal{F}$ is

$$(4.36)\qquad \begin{aligned}[T(g)f](z,t)&=(d+bt)^{-u-1}\left(a+\frac{c}{t}\right)^{-u-1}\exp\left(\frac{bzt}{d+bt}\right)\\ &\quad\cdot f\left(\frac{zt}{(at+c)(d+bt)},\frac{at+c}{d+bt}\right),\end{aligned}$$

$|c/at|<1$, $|bt/d|<1$, $-\pi<\arg a$, $\arg d<\pi$, $ad-bc=1$, and g lies in a sufficiently small neighbourhood of the identity element $e\in SL(2,C)$.

As in *I*, the operators J^+, J^-, J^3 give rise to the operators K^+, K^-, K^3:

$$(4.37)\qquad \begin{aligned}K^+&=\mathcal{D}^{-1}J^+\mathcal{D}=t\left[z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-z+u-\mu+(\mu-\lambda)\frac{\partial^{-1}}{\partial z^{-1}}+2\right],\\ K^-&=\mathcal{D}^{-1}J^-\mathcal{D}=t^{-1}\left(z\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}+u-\mu+2\right),\\ K^3&=\mathcal{D}^{-1}J^3\mathcal{D}=t\frac{\partial}{\partial t}\end{aligned}$$

such that

$$K^+h_m=(m-u)h_{m+1},\quad K^-h_m=-(m+u)h_{m-1},\quad K^3h_m=mh_m, \tag{4.38}$$

$$h_m(z,t)=\frac{\Gamma(u+m+1)}{\Gamma(2u+2)\Gamma(m-u)}z^{\mu-1}{}_2F_2\left[\begin{matrix}1+u-m,\lambda; \\ 2+2u,\mu;\end{matrix}\ z\right]t^m. \tag{4.39}$$

Again

$$\begin{gathered}[K^3,K^\pm]=\pm K^\pm,\qquad [K^+,K^-]=2K^3,\\ (K^+K^-+K^3K^3-K^3)h_m=u(u+1)h_m.\end{gathered} \tag{4.40}$$

If we write $h_m(z,t)=Z_m(z)t^m$, in terms of the functions $Z_m(z)$ the above relations become

$$\begin{gathered}\left[z\frac{d}{dz}+(\mu-\lambda)\frac{d^{-1}}{dz^{-1}}+m+u-\mu-z+2\right]Z_m(z)=(m-u)Z_{m+1}(z),\\ \left(z\frac{d}{dz}+u-\mu+2-z\right)Z_m(z)=-(m+u)Z_{m-1}(z),\\ \left[z^2\frac{d^2}{dz^2}+(2u-2\mu-z)z\frac{d}{dz}+(\mu-\lambda)(u-\mu-m-1)\frac{d^{-1}}{dz^{-1}}\right.\\ \left.+(2\mu+m-u-\lambda)z+(m-2u)(\mu+1)\right]Z_m(z)=0.\end{gathered} \tag{4.41}$$

The multiplier representation T' induced by the K-operators on $\mathcal{F}$ is

$$\begin{aligned}[T'(g)h](z,t)=\mathcal{D}^{-1}\Bigg[&(b+dt)^{-u-1}\left(a+\frac{c}{t}\right)^{-u-1}\exp\left(\frac{bzt}{d+bt}\right)\\ &\cdot(\mathcal{D}h)\left(\frac{zt}{(at+c)(d+bt)},\frac{at+c}{d+bt}\right)\Bigg],\end{aligned} \tag{4.42}$$

$|c/at|<1$, $|bt/d|<1$ and $g\in SL(2,C)$ lies in a sufficiently small neighbourhood of the identity element so that the above expression is uniquely defined.

It has been shown in [4] that

$$F=t^{m_0}e^z\psi_2[1+u+m_0;\,2+2u,\gamma';\,-z,t] \tag{4.43}$$

satisfies both

$$(J^+J^-+J^3J^3-J^3)F=u(u+1)F \tag{4.44}$$

and

$$[J^3J^3+(\gamma'-2m_0-1)J^3-J^+]F=m_0(\gamma'-m_0)F, \tag{4.45}$$

while it can be shown that

$$G=t^{m_0}(1-t)^{-1-u-m_0}e^z{}_1F_1\left[\begin{matrix}1+u+m_0; \\ 2+2u;\end{matrix}\ -\frac{z}{1-t}\right] \tag{4.46}$$

is a common solution of

$$(J^+J^-+J^3J^3-J^3)G=u(u+1)G \tag{4.47}$$

and

$$[J^3J^3-J^-J^3+m_0J^--(1+2m_0)J^3]G=-m_0(m_0+1)G. \tag{4.48}$$

Therefore, by (2.6) and (2.9), it follows that

$$\mathscr{H}_1 = z^{\mu-1} t^{m_0} G^{1:1;2}_{0:2;1}\left(\begin{matrix} 1+u-m_0\colon \lambda;\ 1+u+m_0,\ m_0-u; \\ \text{—}\colon 2+2u,\ \mu;\ \gamma'; \end{matrix}\quad z, -t\right) \tag{4.49}$$

satisfies

$$(K^+K^- + K^3K^3 - K^3)\mathscr{H}_1 = u(u+1)\mathscr{H}_1 \tag{4.50}$$

as well as

$$[K^3K^3 + (\gamma' - 2m_0 - 1)K^3 - K^+]\mathscr{H}_1 = m_0(\gamma' - m_0)H_1, \tag{4.51}$$

while

$$\mathscr{H}_2 = z^{\mu-1} t^{m_0}(1-t)^{-1-u-m_0} F^{1:1;0}_{1:1;0}\left(\begin{matrix} \lambda\colon 1+u+m_0;\ \text{—}; \\ \mu\colon 2+2u;\ \text{—}; \end{matrix}\quad -\frac{z}{1-t}, z\right) \tag{4.52}$$

satisfies

$$(K^+K^- + K^3K^3 - K^3)\mathscr{H}_2 = u(u+1)\mathscr{H}_2 \tag{4.53}$$

as well as

$$[K^3K^3 - K^-K^3 + m_0K^- - (1+2m_0)K^3]\mathscr{H}_2 = -m_0(m_0+1)\mathscr{H}_2. \tag{4.54}$$

From the expansions

$$T(g)\mathscr{H}_i = \sum_{n=-\infty}^{\infty} j_{in} h_{m_0+n}, \qquad i = 1, 2, \tag{4.55}$$

we obtain the following identities:

$$\begin{aligned} &(1-t)^{-\alpha} F^{1:1;0}_{0:2;1}\left(\begin{matrix} \alpha\colon \lambda;\ \text{—}; \\ \text{—}\colon \mu,\ \gamma;\ \gamma'; \end{matrix}\quad \frac{z}{1-t}, \frac{-\omega t}{1-t}\right) \\ &\quad = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{n!}\, {}_2F_2\left[\begin{matrix} \alpha+n,\ \lambda; \\ \gamma,\ \mu; \end{matrix}\quad z\right] {}_1F_1\left[\begin{matrix} -n; \\ \gamma'; \end{matrix}\quad \omega\right] t^n, \qquad |t| < 1, \end{aligned} \tag{4.56}$$

$$\begin{aligned} &\left(1+\frac{c}{t}\right)^{-\alpha} (1-c-t)^{\alpha-\gamma} F^{1:1;0}_{1:1;0}\left(\begin{matrix} \lambda\colon \gamma-\alpha;\ \text{—}; \\ \mu\colon \gamma;\ \text{—}; \end{matrix}\quad \frac{-zt}{(c+t)(1-c-t)}, \frac{zt}{c+t}\right), \\ &\quad = \sum_{n=-\infty}^{\infty} \frac{\Gamma(\gamma-\alpha+n)}{\Gamma(\gamma-\alpha)}\, {}_2F_2\left[\begin{matrix} \alpha-n,\ \lambda; \\ \gamma,\ \mu; \end{matrix}\quad z\right] \\ &\qquad \cdot \frac{{}_2F_1(\gamma-\alpha+n, 1-\alpha+n; 1+n; c)}{\Gamma(1+n)} t^n, \qquad \left|\frac{c}{t}\right| < 1, \quad |c+t| < 1, \end{aligned} \tag{4.57}$$

where the terms corresponding to $n = -1, -2, -3, \cdots$ are well defined because of the relation of the type (4.32).

The above identities are valid for all α, γ, γ', λ, $\mu \in C$ such that α and $\gamma - \alpha$ are not integers and γ, γ', $\mu \neq 0, -1, -2, \cdots$.

5. Type B representation $\uparrow u$. Consider the representation $\uparrow u$ of $sl(2, C)$, $2u$ not a nonnegative integer. The representation space W has a basis $\{f_m\}$, $m \in S = \{-u+n\colon n \geqq 0\}$, such that the action of $sl(2, C)$ on W is given by

$$\begin{aligned} &J^+ f_m = (m-u) f_{m+1}, \quad J^- f_m = -(m+u) f_{m-1}, \quad J^3 f_m = m f_m, \\ &(J^+J^- + J^3J^3 - J^3) f_m = u(u+1) f_m. \end{aligned} \tag{5.1}$$

The type B operators $\{J^+, J^-, J^3\}$ and the basis functions $\{f_m\}$ satisfying (5.1) are [7, p. 188]

$$(5.2)\qquad \begin{aligned} J^+ &= t\left(z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-z-u\right),\\ J^- &= t^{-1}\left(z\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}-u\right),\\ J^3 &= t\frac{\partial}{\partial t} \end{aligned}$$

and

$$(5.3)\qquad f_m(z,t)=\frac{\Gamma(-2u)n!}{\Gamma(n-2u)}L_n^{(-2u-1)}(z)t^m,\qquad m=-u+n\in S.$$

The representation $\uparrow u$ can be extended to a local multiplier representation T of $SL(2,C)$ on the space $\mathcal{F}$ of all functions analytic in a neighbourhood of the point $(z^0,t^0)=(1,1)$.

Thus

$$(5.4)\qquad [T(g)f](z,t)=(d+bt)^u\left(a+\frac{c}{t}\right)^u e^{bzt/(d+bt)}f\left(\frac{zt}{(at+c)(d+bt)},\frac{at+c}{d+bt}\right),$$

$|c/at|<1, |bt/d|<1, f\in\mathcal{F}$ and g lies in a small enough neighbourhood of $e\in SL(2,C)$ such that the right-hand side of the expression is uniquely defined.

As in § 4, the operators J^+, J^-, J^3 induce operators K^+, K^-, K^3

$$(5.5)\qquad \begin{aligned} K^+ &= \mathcal{D}^{-1}J^+\mathcal{D}=t\left[z\frac{\partial}{\partial z}+t\frac{\partial}{\partial t}-z-u-\mu+1+(\mu-\lambda)\frac{\partial^{-1}}{\partial z^{-1}}\right],\\ K^- &= \mathcal{D}^{-1}J^-\mathcal{D}=t^{-1}\left(z\frac{\partial}{\partial z}-t\frac{\partial}{\partial t}-u-\mu+1\right),\\ K^3 &= \mathcal{D}^{-1}J^3\mathcal{D}=t\frac{\partial}{\partial t}, \end{aligned}$$

$$(5.6)\qquad \begin{aligned} &K^+h_m=(m-u)h_{m+1},\quad K^-h_m=-(m+u)h_{m-1},\quad K^3h_m=mh_m,\\ &(K^+K^-+K^3K^3-K^3)h_m=u(u+1)h_m, \end{aligned}$$

where

$$(5.7)\qquad h_m(z,t)={}_2F_2\left[\begin{matrix}-n,\lambda;\\ -2u,\mu;\end{matrix}\; z\right]t^m,\qquad m=-u+n\in S.$$

In view of the fact that

$$(5.8)\qquad [K^3,K^\pm]=\pm K^\pm,\qquad [K^+,K^-]=2K^3,$$

we have constructed a new model of irreducible representation $\uparrow u$ of $sl(2,C)$ in terms of differintegral operators $\{K^+, K^-, K^3\}$ such that the functions $h_m(z,t)$ defined by (5.7) are basis vectors.

The multiplier representation T' induced by the K-operators is

$$(5.9)\qquad \begin{aligned} [T'(g)h](z,t)=\mathcal{D}^{-1}\bigg[&(d+bt)^u\left(a+\frac{c}{t}\right)^u\exp\left(\frac{bzt}{d+bt}\right)\\ &\cdot(\mathcal{D}h)\left(\frac{zt}{(at+c)(d+bt)},\frac{at+c}{d+bt}\right)\bigg], \end{aligned}$$

$|c/at|<1$, $|bt/d|<1$, $ad-bc=1$, and g lies in a sufficiently small neighbourhood of $e\in SL(2, C)$.

It can be easily shown that

$$F=\psi_2[-2u;\ -2u,\ \gamma';\ -z,\ t]e^z t^{-u} \tag{5.10}$$

satisfies

$$(J^+J^-+J^3J^3-J^3)F=u(u+1)F \tag{5.11}$$

as well as

$$[J^3J^3+(2u+\gamma'-1)J^3-J^+]F=-u(u+\gamma')F. \tag{5.12}$$

It therefore follows that

$$\mathscr{H}=z^{\mu-1}t^{-u}F^{1:1;0}_{1:2;0}\left(\begin{matrix}-2u\colon \lambda;\text{—};\\ \gamma'\colon -2u,\ \mu;\text{—};\end{matrix}\quad -zt,\ t\right) \tag{5.13}$$

satisfies the differintegral equation

$$(K^+K^-+K^3K^3-K^3)\mathscr{H}=u(u+1)\mathscr{H} \tag{5.14}$$

as well as

$$[K^3K^3+(2u+\gamma'-1)K^3-K^+]\mathscr{H}=-u(u+\gamma')\mathscr{H}. \tag{5.15}$$

The expansion

$$T'(g)\mathscr{H}=\sum_{n=0}^{\infty} p_n(g)h_{-u+n} \tag{5.16}$$

leads us to an identity

$$\begin{aligned} &F^{1:1;0}_{1:2;0}\left(\begin{matrix}\gamma\colon \lambda;\text{—};\\ \gamma'\colon \mu,\ \gamma;\text{—};\end{matrix}\quad zt,\ t+c\right)\\ &=\sum_{n=0}^{\infty}\frac{(\gamma)_n}{n!(\gamma')_n}\,{}_2F_2\left[\begin{matrix}-n,\ \lambda;\\ \gamma,\ \mu;\end{matrix}\quad z\right]{}_1F_1\left[\begin{matrix}\gamma+n;\\ \gamma'+n;\end{matrix}\quad c\right]t^n.\end{aligned} \tag{5.17}$$

We are not discussing type A representation $\uparrow u$ since the operators defining it do not fit into the fractional derivative approach.

6. Type A representation $\downarrow u$. consider the representation $\downarrow u$ of $sl(2, C)$, $2u$ not a nonnegative integer [7, p. 205]. The representation space W has a basis $\{f_m\}$, $m\in S=\{u-n\colon n=0, 1, 2, \cdots\}$ such that the action of $sl(2, C)$ on W is given by

$$\begin{gathered} J^+f_m=(m-u)_{m+1},\quad J^-f_m=-(m+u)f_{m-1},\quad J^3f_m=mf_m,\\ (J^+J^-+J^3J^3-J^3)f_m=u(u+1)f_m.\end{gathered} \tag{6.1}$$

The type A operators $\{J^+, J^-, J^3\}$ here are (4.3) while

$$f_m(z, t)={}_2F_1(-n, -q-u;\ -2u;\ z)t^m,\qquad m=u-n,\quad n=0, 1, 2, \cdots. \tag{6.2}$$

Accordingly, K-operators induced by the J-operators are (4.7) and (4.9), while the corresponding basis functions $h_m(z, t)$ are

$$h_m(z, t)=z^{\mu-1}\,{}_3F_2\left[\begin{matrix}-n,\ -q-u,\ \lambda;\\ -2u,\ \mu;\end{matrix}\quad z\right]t^m,\qquad m=u-n,\quad n=0, 1, 2, \cdots, \tag{6.3}$$

where

$$\begin{gathered} K^+h_m=(m-u)h_{m+1}, \quad K^-h_m=-(m+u)h_{m-1}, \quad K^3h_m=mh_m, \\ (K^+K^-+K^3K^3-K^3)h_m=u(u+1)h_m, \\ [K^3, K^\pm]=\pm K^\pm, \quad [K^+, K^-]=2K^3. \end{gathered} \tag{6.4}$$

The multiplier representation T' of $SL(2, C)$ in consequence of the K-operators is (4.15). That is,

$$\begin{aligned} [T'(g)h](z, t)=\mathscr{D}^{-1}\Bigg[&(d+bt)^u\left(a+\frac{c}{t}\right)^{-q}\left(a-\frac{c(z-1)}{t}\right)^{q+u} \\ &\cdot(\mathscr{D}h)\left(\frac{zt}{(d+bt)(at-c(z-1))}, \frac{c+at}{d+bt}\right)\Bigg], \end{aligned} \tag{6.5}$$

$|bt/d|<1$, $|c/at|<1$, $|(c(z-1))/at|<1$, and g lies in a sufficiently small neighbourhood of $e\in SL(2, C)$.

It can be shown that

$$F=F_2\left[-2u; -q-u, \alpha'; -2u, \beta'; \frac{-z}{1-z}, \frac{1}{t}\right](1-z)^{q+u}t^u \tag{6.6}$$

is a solution of both

$$(J^+J^-+J^3J^3-J^3)F=u(u+1)F \tag{6.7}$$

and

$$[J^3J^3+J^-J^3-(u+\alpha')J^-+(1-2u-\beta')J^3]F=u(1-u-\beta')F. \tag{6.8}$$

Therefore, it follows that

$$\mathscr{H}=z^{\mu-1}t^uF^{2:2;0}_{1:2;0}\left(\begin{matrix}-2u, \alpha'\colon -q-u, \lambda; —; \\ \beta'\colon -2u, \mu; —;\end{matrix}\quad -\frac{z}{t}, \frac{1}{t}\right) \tag{6.9}$$

satisfies

$$(K^+K^-+K^3K^3-K^3)\mathscr{H}=u(u+1)\mathscr{H} \tag{6.10}$$

as well as

$$[K^3K^3+K^-K^3-(u+\alpha')K^-+(1-2u-\beta')K^3]\mathscr{H}=u(1-u-\beta')\mathscr{H}. \tag{6.11}$$

From the expansion

$$T'(g)\mathscr{H}=\sum_{n=0}^{\infty} q_n(g)h_{u-n}, \tag{6.12}$$

valid in a region to be determined by the inequalities (6.5), we obtain the following identity as a special case:

$$\begin{aligned} &F^{2:2;0}_{1:2;0}\left(\begin{matrix}\beta, \alpha'\colon \alpha, \lambda; —; \\ \beta'\colon \beta, \mu; —;\end{matrix}\quad \frac{z}{t}, \frac{1+bt}{t}\right) \\ &\quad=\sum_{n=0}^{\infty}\frac{(\alpha')_n(\beta)_n}{n!(\beta')_n}\,{}_3F_2\left[\begin{matrix}-n, \alpha, \lambda; \\ \beta, \mu;\end{matrix}\quad z\right]{}_2F_1\left[\begin{matrix}\alpha'+n, \beta+n; \\ \beta'+n;\end{matrix}\quad b\right]\frac{1}{t^n}, \\ &\qquad\qquad\left|\frac{z}{t}\right|+\left|\frac{1+bt}{t}\right|<1. \end{aligned} \tag{6.13}$$

We are not discussing type B representation $\downarrow u$ since it would lead to the identity (5.17) again and as such give no new results.

7. Conclusion. As we have seen, the operator $\mathscr{D}$ has been highly instrumental in constructing new models of irreducible representation of $sl(2, C)$, induced by type A and type B operators. For $\lambda = \mu$ all the new models reduce to the known ones.

We have not included the discussion on representations of the oscillator algebra or the Lie algebra of the Euclidean group, based on type C′, C″, · · · , operators, since, to deal with these, we need to define the operator $\mathscr{D}$ in a different way. However, we propose to discuss this in a separate article.

Acknowledgment. The authors record their thanks to the referee for suggestions which have led to a better presentation of the paper.

REFERENCES

[1] M. A. Al-Bassam, *Some properties of Holmgren-Riesz tranform,* Ann. Scuola. Norm. Sup. Pisa (3), 15 (1961), pp. 1-24.

[2] A. K. Grunwald, *Uber begrentze Derivationon and deren Anwendung,* Z. Angew. Math. Phys., 12 (1867), pp. 441-480.

[3] J. Horn, *Hypergeometrische funktionon zweier veranderlichen,* Math. Ann., 105 (1931), pp. 381-407.

[4] S. Jain and H. L. Manocha, *Special linear group and generating functions,* Comment. Math. Univ. St. Paul, XXVI (1977), pp. 105-113.

[5] E. G. Kalnins, H. L. Manocha and W. Miller, *Transformations and reduction formulas for two-variable hypergeometric functions on the Sphere S_2,* Stud. Appl. Math., 63 (1980), pp. 155-167.

[6] ———, *Harmonic analysis and expansion formulas for two-variable hypergeometric functions,* Stud. Appl. Math., 66 (1982), pp. 69-89.

[7] W. Miller, *Lie Theory and Special Functions,* Academic Press, New York, 1968.

[8] W. Miller, *Lie theory and specialfunctions satisfying second order nonhomogeneous differential equations,* this Journal, 1 (1970), pp. 246-265.

[9] T. J. Osler, *Fractional derivatives and Leibniz rule,* Amer. Math. Monthly, 78 (1971), pp. 645-649.

[10] H. M. Srivastava and H. L. Manocha, *A Treatise on Generating Functions,* Ellis Horwood Ltd., Halsted Press, John Wiley, Chichester, 1984.

[11] L. Weisner, *Group-theoretic origin of certain generating functions,* Pacific J. Math., 5 (1955), pp. 1083-1089.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

014

A GENERAL FORM FOR SOLVABLE LINEAR TIME VARYING SINGULAR SYSTEMS OF DIFFERENTIAL EQUATIONS*

STEPHEN L. CAMPBELL†

Abstract. A canonical form is derived for all linear solvable systems $E(t)x'(t)+F(t)x(t)=f(t)$ with sufficiently smooth coefficients E, F. Using this form it is shown that for all smooth enough solvable systems a class of recently defined numerical imbedding methods and an algorithm to compute the manifold of consistent initial conditions always work. In addition, necessary and sufficient conditions are given on $E(t)$, $F(t)$ to insure solvability in the case when $E(t)$, $F(t)$ are infinitely differentiable.

Key words. linear time varying system, implicit, descriptor, singular, solvability, numerical imbedding, consistent initial conditions, approximation

AMS(MOS) subject classifications. Primary 34A08; secondary 34A10, 34A30

1. Introduction.

The theory for linear time invariant singular systems (also called differential-algebraic, descriptor, implicit or constrained) has become well developed over the last decade (see [3], [4]). However, progress on the linear time varying problem

$$E(t)x'(t)+F(t)x(t)=f(t) \tag{1.1}$$

is more recent and has been less complete. Work on (1.1) initially took either the form of algorithms utilizing repeated coordinate changes and differentiations to reduce (1.1) to some type of canonical or explicit form [18], [21], or of traditional numerical methods such as backward differences [12].

The canonical form approach was continued in [6], [10], [19]. However, examples in [6], [10] showed that there were solvable systems (to be defined shortly) which could not be put into these canonical forms. It was also observed in [19] that backward difference methods do not converge for all solvable systems although they do work for certain important classes of problems [2], [14], [15], [19]. (Note also [13].) In an attempt to rectify this difficulty, a different numerical approach was introduced in [7] and further developed in [8], [9]. This method was shown to work for many of the standard classes of singular systems (1.1) including some for which backward differences failed and the approach of [18], [21] was numerically impractical. This paper will bring together some of this previous work.

First, we shall give a result which exhibits for the first time the structure of all sufficiently smooth solvable systems. In particular, we will show that the counterexamples in [10] are, in a sense, "generic". Using this structure result, which is similar to, but different from, that of [10], [19], we show that the method [7] works for all sufficiently smooth solvable systems and that [7] also provides a tool for the theoretical analysis of (1.1). Also, under the added assumption that E, F, f are infinitely differentiable we develop a test for solvability that can be verified directly from E, F and their derivatives without any time varying coordinate changes.

2. Terminology and the canonical form.

Let $\mathcal{I}=[t_0, t_1]$ be a finite subinterval of the real line. We assume that E, F are $n\times n$ (possible complex) matrix valued functions, $n\geqq 2$, and E is singular for all $t\in\mathcal{I}$. We do not assume E has constant rank. To avoid

* Received by the editors August 26, 1985; accepted for publication (in revised form) June 3, 1986. This research was supported in part by the National Science Foundation under grant DMS-8318026 and the Air Force Office of Scientific Research under grant 84-0240.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

technical difficulties we assume E, F, are $2n$-times differentiable and f is at least n-times differentiable where differentiable is taken throughout this paper to mean continuously differentiable.

The system (1.1) is *solvable* on $\mathcal{I}$ if for every such f there exists at least one continuously differentiable solution x. (This is a form of target path controllability [24].) Also, all solutions for this f are defined on all of $\mathcal{I}$, are at least differentiable, and are uniquely determined by their value at any $t \in \mathcal{I}$. Finally, all solutions of the homogeneous equation $Ex'+Fx=0$ are at least $(2n+1)$-times differentiable and if f is m-times differentiable with $n \leqq m \leqq 2n$, then the solution x is $(m-n+1)$-times differentiable.

This definition may appear somewhat cumbersome, but when E, F are constant it is equivalent to the usual definition of solvability [3], [4] and allows us to concentrate on structural questions without worrying if we have lost differentiability during coordinate changes.

Those x^o for which (1.1) has a differentiable solution such that $x(t_0)=x^o$ are called *consistent initial conditions* at t_0.

Two singular systems $Ex'+Fx=f$ and $\tilde{E}\tilde{x}'+\tilde{F}\tilde{x}=\tilde{f}$ are said to be *equivalent* if there exists $2n$-times differentiable nonsingular $H(t)$, and $(2n+1)$-times differentiable nonsingular $K(t)$ such that letting $x=K\tilde{x}$ and multiplying by H changes $Ex'+Fx=f$ into $\tilde{E}\tilde{x}'+\tilde{F}\tilde{x}=\tilde{f}$. That is,

$$HEK=\tilde{E},\quad HEK'+HFK=\tilde{F},\quad Hf=\tilde{f}. \tag{2.1}$$

Finally, from [6], a system is in standard canonical form (SCF) if it takes the form

$$y'+C(t)y=g(t), \tag{2.2}$$

$$N(t)z'+z=h(t) \tag{2.3}$$

where N is strictly lower triangular independent of t and (2.2) may be absent. Note that (2.3) has only one solution for each h [6]. By an additional coordinate change on y, one could rewrite (2.2) as $y'(t)=g(t)$. Systems equivalent to one in SCF are clearly solvable.

If E, F are real analytic, then (1.1) is solvable if and only if it can be put into SCF [10]. However, not all solvable systems can be put into SCF. In [19], Petzold and Gear show that if (1.1) is solvable, then there is a collection of open intervals $\mathcal{I}_i$ such that $\bigcup \mathcal{I}_i$ is dense in $\mathcal{I}$ and on each $\mathcal{I}_i$, (1.1) can be put into SCF. However, the coordinate changes and canonical form need not be continuous on $\mathcal{I}$ [10].

Example 2.1. Let N_-, N_+ be any two $n\times n$ nilpotent matrices. Let $\phi(t)$ be a $(2n+1)$-times differentiable function on $[-1,1]$ such that $\phi^{(i)}(0)=0$ for $0\leqq i \leqq 2n+1$. Let $F(t)=I$, $E(t)=\phi(t)N_-$ if $t\leqq 0$ and $E(t)=\phi(t)N_+$ if $t>0$. Then it is not hard to show by the next lemma that

$$Ex'+x=f \tag{2.4}$$

is solvable on $[-1,1]$. If, for example, the nullspaces of N_+, N_- do not intersect, then (2.4) cannot be put into SCF.

LEMMA 2.1. *If N is an $n\times n$ $2n$-times differentiable function such that any n-product of N and its derivatives is zero, then $Nx'+x=f$ is solvable on $\mathcal{I}$ and the unique solution is*

$$x=(I+\bar{N}D)^{-1}f=(I-\bar{N}D+(\bar{N}D)^2+\cdots(-1)^{n-1}(\bar{N}D)^{n-1})f \tag{2.5}$$

where $\bar{N}$ is the operator of multiplication by N and D is the operator of differentiation.

Proof. By assumption, $(\bar{N}D)^n = 0$. □

In Example 2.1, a projection onto the nullspace of $E(t)$ could exhibit a jump discontinuity at zero. An example where the projection is not even piecewise continuous is given in [10]. The next result gives the first general structure theorem for solvable systems and show that the previous examples are typical of how a solvable system can fail to be equivalent to a system in SCF.

THEOREM 2.1. *Suppose that* (1.1) *is solvable on* $\mathcal{I}$. *Then* (1.1) *is equivalent to a system in the form*

$$y' + C(t)z' = g(t), \tag{2.6a}$$

$$N(t)z' + z = h(t) \tag{2.6b}$$

where (2.6a) *may be absent and* (2.6b) *has only one solution for each sufficiently smooth* $h(t)$. *Furthermore, there exists a family of disjoint open intervals* $\mathcal{I}_i$ *such that* $\bigcup \mathcal{I}_i$ *is dense in* $\mathcal{I}$ *and on each* $\mathcal{I}_i$ *the system* (2.6b) *may be transformed by H, K as in* (2.1) *so that* (2.6b) *is in* SCF. *In particular, on* $\mathcal{I}_i$ *the equivalent system will have the form* (2.6) *with* $N(t)$ *at least* $2n$*-times differentiable and strictly lower triangular independent of t.*

Before proceeding to prove Theorem 2.1 we point out that the key differences between Theorem 2.1 and those of [10], [18], [19], [21] are that, while the structure is less detailed and the systems (2.6a), (2.6b) are not completely decoupled, the coordinate changes and matrices in (2.6b) are defined and sufficiently differentiable on *all* of $\mathcal{I}$ and not just on $\bigcup \mathcal{I}_i$. Also, we do not make assumptions of analyticity or constant rank so that Theorem 2.1 applies to all solvable systems which have smooth enough coefficients.

Proof. Suppose that (1.1) is solvable. First we wish to get (1.1) in the form

$$x' + A_1 y' + Cx = f_1, \tag{2.7a}$$

$$A_2 y' + By = f_2 \tag{2.7b}$$

with the bottom equation having only one solution for sufficiently smooth f. If the associated homogeneous equation of (1.1) has only $x = 0$ as a solution, (1.1) is already in the form (2.7b) with (2.7a) absent and this step is complete. So suppose there are nontrivial solutions of (1.1) with $f = 0$. Let $\{\phi_1, \cdots, \phi_r\}$ be a basis for these homogeneous solutions. By the assumption of solvability $\{\phi_1, \cdots, \phi_r\}$ are linearly independent for each t and $(2n+1)$-times differentiable on all of $\mathcal{I}$. Then there exists $\{\phi_{r+1}, \cdots, \phi_n\}$ equally smooth such that $\Phi = [\phi_1, \cdots, \phi_n]$ is invertible on $\mathcal{I}$ (see [22] for example). Let $x = \Phi \tilde{x}$. Then (1.1) becomes

$$E\Phi\tilde{x}' + (E\Phi' + F\Phi)\tilde{x} = f$$

or

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \begin{bmatrix} \tilde{x}_1' \\ \tilde{x}_2' \end{bmatrix} + \begin{bmatrix} 0 & F_{21} \\ 0 & F_{22} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix} \tag{2.8}$$

where $E\Phi = [E_{ij}]$, $E\Phi' + F\Phi = [F_{ij}]$. We omit the tilde and use x_1, x_2 to denote our new variables. We shall now show $[\begin{smallmatrix} E_{11} \\ E_{21} \end{smallmatrix}]$ has full (column) rank. Suppose that $\hat{t} \in \mathcal{I}$ such that $[\begin{smallmatrix} E_{11} \\ E_{21} \end{smallmatrix}]$ is not full rank. Let ϕ be a nonzero constant vector such that

$$\begin{bmatrix} E_{11}(\hat{t}) \\ E_{21}(\hat{t}) \end{bmatrix} \phi = 0.$$

Define

$$f(t)=\begin{cases}\dfrac{1}{t-\hat{t}}\begin{bmatrix}E_{11}(t)\\ E_{21}(t)\end{bmatrix}\phi & \text{if } t\neq\hat{t},\\[2ex] \begin{bmatrix}E_{11}'(\hat{t})\\ E_{21}'(\hat{t})\end{bmatrix}\phi & \text{if } t=\hat{t}.\end{cases}$$

Then f is $2n-1$ times differentiable on $\mathscr{I}$. However $x_1=\ln|t-\hat{t}|\phi$, $x_2=0$ defines an unbounded solution on $\mathscr{I}\backslash\{\hat{t}\}$ wnich contradicts solvability. Hence $[{E_{11} \atop E_{21}}]$ has full (column) rank on $\mathscr{I}$.

Thus there exists a $2n$-times differentiable nonsingular matrix P such that

$$P\begin{bmatrix}E_{11}\\ E_{21}\end{bmatrix}=\begin{bmatrix}I\\ 0\end{bmatrix}.$$

Multiplying (2.8) by P gives that our system is now

$$\begin{bmatrix}I & E_1\\ 0 & E_2\end{bmatrix}\begin{bmatrix}x_1'\\ x_2'\end{bmatrix}+\begin{bmatrix}0 & F_1\\ 0 & F_2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\end{bmatrix}\tag{2.9}$$

and by the construction of Φ, $E_2x_2'+F_2x_2=f_2$ has only one solution for each f_2. Let $x_2=S\bar{x}_2$. Then (2.9) becomes

$$\begin{bmatrix}I & E_1S\\ 0 & E_2S\end{bmatrix}\begin{bmatrix}x_1'\\ \bar{x}_2'\end{bmatrix}+\begin{bmatrix}0 & F_1S+E_1S'\\ 0 & F_2S+E_2S'\end{bmatrix}\begin{bmatrix}x_1\\ \bar{x}_2\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\end{bmatrix}.\tag{2.10}$$

Suppose for the moment that it is possible to choose $2n+1$-times differentiable S such that S and $E_2S'+F_2S$ are both invertible for every $t\in\mathscr{I}$. Then we have (2.9) with F_2 now nonsingular. Multiplying (2.9) by

$$\begin{bmatrix}I & -F_1F_2^{-1}\\ 0 & F_2^{-1}\end{bmatrix}$$

yields the equivalent system

$$\begin{bmatrix}I & C\\ 0 & N\end{bmatrix}\begin{bmatrix}z'\\ w'\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}z\\ w\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\end{bmatrix}\tag{2.11}$$

which is (2.6). To complete the proof, we need to verify that one can define S as indicated. Suppose that $p=$ dimension of the w space in (2.11). Let $\mathscr{C}_m$ be the Banach space of m-times continuously differentiable $p\times p$ complex valued matrix functions defined on $\mathscr{I}$ with $\|X\|_m=\sum_{i=0}^m\|X^{(i)}\|_\infty$, where $\|\cdot\|_\infty$ is the sup norm on $\mathscr{C}_0$. Let $\mathscr{L}(x)=E_2x'+F_2x$ with E_2, F_2 from (2.9). Let Ω_m be those $X\in\mathscr{C}_m$ which are nonsingular for every $t\in\mathscr{I}$. We need the technical fact that

(2.12) For $m\geqq 0$, Ω_m is an open dense subset of $\mathscr{C}_m$.

Lacking a reference for (2.12), we include a proof. Clearly Ω_m is open in $\mathscr{C}_m$ so we need only show Ω_m is dense. Suppose $T\in\mathscr{C}_m\backslash\Omega_m$. Take $\varepsilon>0$. Then there exists $P\in\mathscr{C}_m$ such that $\|P-T\|_m<\varepsilon/2$ and P has polynomial entries. Let Σ be the set of all eigenvalues of $P(t)$ for all $t\in\mathscr{I}$. Then Σ has no interior as a subset of the complex plane. This follows from the fact that $P(t)$ is real analytic in t and hence the eigenvalues are continuous functions of t which are analytic except at a finite number of t values [16]. Hence there exists a number $\varepsilon_1>0$ such that $\varepsilon_1<\varepsilon/2$ and $\varepsilon_1\notin\Sigma$. Let $H=P-\varepsilon_1$. Then $H\in\Omega_m$ and $\|H-T\|_m<\varepsilon$ as desired. This completes the proof of (2.12). Now take $\Psi\in\Omega_n$. By assumption on $\mathscr{L}$ (solvability) there exists a $\Phi\in\mathscr{C}_1$ such that $\mathscr{L}(\Phi)=\Psi$. Let ε be such that the ε neighborhood of $\Psi\in\mathscr{C}_0$ is in Ω_0. By (2.12) there exists $\{\Phi_j\}\subset\Omega_{2n+1}$ such that $\Phi_j\to\Phi$ in $\mathscr{C}_1$. But then $\mathscr{L}(\Phi_j)\to\mathscr{L}(\Phi)=\Psi$ in $\mathscr{C}_0$. Let $\hat{j}$ be such that $\|\mathscr{L}(\Phi_{\hat{j}})-\mathscr{L}(\Phi)\|<\varepsilon$. Then $\Phi_{\hat{j}}\in\Omega_{2n+1}$ and $\mathscr{L}(\Phi_{\hat{j}})\in\mathscr{C}_{2n}\cap\Omega_0=\Omega_{2n}$ so that $\Phi_{\hat{j}}$ is the required S. This completes the proof of (2.11). To get the structure of N we use the fact that if $X(t)$ is any continuous matrix function on $\mathscr{I}$, then there exists a countable collection of disjoint open intervals $\mathscr{I}_i$ such that $\cup\mathscr{I}_i$ is dense in $\mathscr{I}$ and rank X is constant on each $\mathscr{I}_i$. One can then do the reduction process of [19] (or Silverman [21]) where on each step it may be necessary to break each $\mathscr{I}_i$ into another countable family. □

It would be nice if the matrix C could be eliminated in (2.6a) so that (2.6a) and (2.6b) were completely decoupled. We are not sure if this is always possible. However, note that if we let $z=u+Tw$ in (2.11) and then add $-T'$ times the second equation to the first, then (2.11) becomes

$$\begin{bmatrix} I & T+C-T'N \\ 0 & N \end{bmatrix}\begin{bmatrix} u' \\ w' \end{bmatrix}+\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} u \\ w \end{bmatrix}=\begin{bmatrix} \tilde{f}_1 \\ f_2 \end{bmatrix}. \tag{2.13}$$

If $T'N-T=C$ has a smooth solution on $\mathscr{I}$, then (2.13) is in the form (2.11) with $C=0$. There are several sufficient conditions for this including the solvability of $N^*X'-X=C^*$ where $*$ denotes the pointwise conjugate transpose and $X=T^*$.

The next section applies Theorem 2.1 to the numerical methods of [7]-[9].

3. Application to the imbedding method. Theorem 2.1 should not be considered a method of solution but rather a very useful theoretical fact.

In [5] and Lemma 2.1 it was shown that the solutions of (1.1) will generally involve not only f and its derivatives but also derivatives of E, F. For simple small scale problems it may be possible to differentiate and solve for variables [20]. This may not be practical on larger or more complicated problems. One of our goals then is to develop a theory as directly in terms of the derivatives of E, F, f as possible. While the proofs involve time varying coordinate changes the results do not.

Suppose that (1.1) is solvable and that $\hat{t}\in\mathscr{I}$. Let

$$c_i(\hat{t})=\frac{c^{(i)}(\hat{t})}{i!}\quad\text{for } c=E, F, f, x,\quad \hat{t}\in\mathscr{I} \tag{3.1}$$

so that

$$c(t)=\sum_i c_i\delta^i,\qquad \delta=t-\hat{t} \tag{3.2}$$

(where (3.1) is a Taylor approximation with a remainder if c is not analytic). Substituting these expansions into (1.1) gives that for any $j>0$ (j less than the smoothness of E, F, x, f), $t\in\mathscr{I}$,

$$\begin{bmatrix} E_0 & 0 & \cdot & \cdot & \cdot & 0 \\ E_1+F_0 & 2E_0 & \cdot & \cdot & \cdot & \cdot \\ E_2+F_1 & 2E_1+F_0 & 3E_0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ E_{j-1}+F_{j-2} & 2E_{j-2}+F_{j-3} & \cdot & \cdot & \cdot & jE_0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_j \end{bmatrix}=\begin{bmatrix} f_0 \\ f_1 \\ \cdot \\ \cdot \\ \cdot \\ f_{j-1} \end{bmatrix}-\begin{bmatrix} F_0 \\ F_1 \\ \cdot \\ \cdot \\ \cdot \\ F_{j-1} \end{bmatrix}x_0 \tag{3.3}$$

or

$$\mathscr{E}_j\mathbf{x}_j=\not{f}_j-\mathscr{F}_jx_0=\mathbf{f}_j \tag{3.4}$$

where all the terms in (3.3) depend on t. Note that $\mathscr{E}_j$ is singular since E_0 is. The system (3.4) (or (3.3)) is said to be *smoothly* 1-*full* on $\mathscr{I}$ if there exists continuously differentiable nonsingular $R(t)$ on $\mathscr{I}$ such that

$$R(t)\mathscr{E}_j(t)=\begin{bmatrix} I & 0 \\ 0 & L(t) \end{bmatrix}$$

where I is $n\times n$. In this case, if $[R_0,\cdots,R_{j-1}]$ is the top n rows of $R(t)$ we have that

$$x_1=-\sum_{i=0}^{j-1} R_iF_ix_0+\sum_{i=0}^{j-1} R_if_i$$

or

$$x'(t)=Q(t)x(t)+q(t). \tag{3.5}$$

Thus the solutions of the original descriptor system (1.1) corresponding to consistent initial conditions are also solutions of the nonsingular system (3.5) provided (3.4) is smoothly 1-full. But solving (3.4) for x_1, given x_0, is the same as evaluating $Q(t)x(t)+q(t)$ in (3.5) given $x(t)$, t. This makes it possible to numerically solve (1.1) by numerically solving (3.5). This is discussed in more detail in [7], [8]. The key then in utilizing the approach of [7], [8] is that $\mathscr{E}_j$ be smoothly 1-full. Notice that (3.3), (3.5) only require that E, F, f, are $(j-1)$-times differentiable, and as we shall see one may usually assume $j\leqq n+1$. However, forming $\mathbf{x}_j$ requires x to be j-times differentiable which usually requires extra smoothness from E, F, f.

An example in [7] shows that 1-full for each $t\in\mathscr{I}$ does not imply smoothly 1-full.

LEMMA 3.1. *Let*

$$M(t)=\begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \tag{3.6}$$

where M_1 is $n\times n$ and suppose that M is smooth on $\mathscr{I}$ and has constant rank. Then M is smoothly 1-*full on $\mathscr{I}$ if and only if it is* 1-*full for each $t\in\mathscr{I}$.*

Proof. Sufficiency is obvious. Suppose M is 1-full for each $t\in\mathscr{I}$ and has constant rank. From the 1-fullness assumption the submatrix $\left[\begin{smallmatrix} M_1 \\ M_3 \end{smallmatrix}\right]$ has full column rank for all t. Thus there exists a smooth invertible $R_1(t)$ such that

$$R_1(t)M(t)=\begin{bmatrix} I & \tilde{M}_2 \\ 0 & \tilde{M}_4 \end{bmatrix}.$$

But $\tilde{M}_4$ has constant rank by assumption since M does. Thus there exist smooth $R_2(t)$, $R_3(t)$ such that

$$R_2\tilde{M}_4R_3=\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence

$$\begin{bmatrix} I & 0 \\ 0 & R_2 \end{bmatrix} R_1M=\left[\begin{array}{c|c} I & \tilde{M}_2 \\ \hline \begin{matrix} 0 \\ 0 \end{matrix} & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} R_3^{-1} \end{array}\right]=\left[\begin{array}{c|ccc} I & [\hat{M}_{21} & \hat{M}_{22}] & R_3^{-1} \\ \hline 0 & [I & 0] & R_3^{-1} \\ 0 & 0 & 0 & \end{array}\right]. \tag{3.7}$$

Now multiply both sides of (3.7) on the left by

$$\begin{bmatrix} I & -\hat{M}_{21} & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \quad \text{to get} \quad \left[\begin{array}{c|ccc} I & [0 & \hat{M}_{22}] & R_3^{-1} \\ \hline 0 & [I & 0] & R_3^{-1} \\ 0 & 0 & 0 & \end{array}\right]. \tag{3.8}$$

But M is 1-full for each t. Thus $\hat{M}_{22}=0$. Since all the row operations given by R_2, $\hat{M}_{21}$, R_1, were smooth, we have the proof of Lemma 3.1. □

Note that in Lemma 3.1 we may take smooth to mean m-times continuously differentiable or real analytic so that r is as smooth as M. Thus if E, F, f are $2n$-times differentiable, then Q, q, in (3.5) will be at least $2n-j+1$ times differentiable and (3.5) will imply the solutions are $(2n-j+2)$-times differentiable.

In [7] it was shown that for some $j\leqq n+1$ one has $\mathscr{E}_j$ is 1-full and constant rank for many known classes of solvable systems. Using Theorem 2.1, we now show that in fact this holds for all solvable systems with smooth enough coefficients. Let $\mathscr{R}(X)$ denote the range (column space) of a matrix X.

THEOREM 3.1. *Suppose that* (1.1) *is solvable on* $\mathscr{I}$ *and* E, F *are* $2n$*-times differentiable. Then*

(3.9) $\mathscr{E}_j$ *has constant rank on* $\mathscr{I}$ *for* $j=n+1$,

(3.10) $\mathscr{E}_j$ *is smoothly one-full on* $\mathscr{I}$ *for* $j=n+1$,

(3.11) $\mathscr{R}(\mathscr{E}_j)+\mathscr{R}(\mathscr{F}_j)=C^{jn}$ *for every* $t\in\mathscr{I}$ *for* $1\leqq j\leqq n+1$.

Proof. Suppose (1.1) is solvable. As shown in [7], performing smooth coordinate changes amounts to doing block row operations and right to left block column operations on $\mathscr{E}_j$, $\mathscr{F}_j$ where the blocks have size $n\times n$. Thus (3.9)-(3.11) are not altered by such changes of coordinates. By Theorem 2.1 we may assume that (1.1) is in the form (2.6). Thus

(3.12)
$$\mathscr{E}_j=\begin{bmatrix} I & C_0 & & & & & & \\ 0 & N_0 & & & & & & \\ 0 & C_1 & 2I & 2C_0 & & & & \\ 0 & N_1+I & 0 & 2N_0 & & & & \\ 0 & C_2 & 0 & 2C_1 & 3I & 3C_0 & & \\ 0 & N_2 & 0 & 2N_1+1 & 0 & 3N_0 & & \\ 0 & & & & & \ddots & jI & jC_0 \\ 0 & & & & & & 0 & jN_0 \end{bmatrix},\quad F_j=\begin{bmatrix} 0 & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

If we perform column operations to eliminate the C_i we see that (3.9), (3.11) hold for (3.12) if and only if they hold for

$$\hat{\mathscr{E}}_j=\begin{bmatrix} N_0 & & & \\ N_1+I & 2N_0 & & \\ N_2 & 2N_1+I & 3N_0 & \\ \vdots & & \ddots & \\ N_j & 2N_{j-1} & \cdots & jN_0 \end{bmatrix},\quad \hat{\mathscr{F}}_j=\begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.13}$$

Note that because of the block lower triangularity of $\mathscr{E}_j$ that $\mathscr{R}(\mathscr{E}_{j_0})+\mathscr{R}(\mathscr{F}_{j_0})=C^{nj_0}$ implies $\mathscr{R}(\mathscr{E}_j)+\mathscr{R}(\mathscr{F}_j)=C^{nj}$ for $1\leqq j\leqq j_0$. Also if $\hat{\mathscr{E}}_j$ in (3.13) is 1-full, then so is $\mathscr{E}_j$ in (3.12). Thus it suffices to prove Theorem 3.1 for (3.11), that is, the subsystem $Nx'+x=f$. Suppose that N is $n\times n$. Now since (1.1) is solvable, at a given t one may take $\mathscr{f}_j$ arbitrary. Thus (3.11) holds for (3.13). Hence rank $(\hat{\mathscr{E}}_j)\geqq(n-1)j$ for all t, $1\leqq j\leqq 2n$, since rank $(\hat{\mathscr{F}}_j)=n$. Now by Theorem 2.1 there exists disjoint open intervals $\mathscr{I}_i$ such that $\bigcup\mathscr{I}_i$ is dense in $\mathscr{I}$ and $Nx'+x=f$ can be put into SCF on each $\mathscr{I}_i$. But then on $\mathscr{I}_i$ the new coefficient matrices in $\mathscr{E}_j$ are all strictly lower triangular independent of t. Then since any n products of the N_i are zero, we can use a variation of a row operation argument used in [7, Thm. 4.3] to show that $\hat{\mathscr{E}}_j$ is 1-full and has rank $(j-1)n$ on $\mathscr{I}_i$, if $j>n$. But $\hat{\mathscr{E}}_j$ is a continuous function of t and thus the rank can only drop at a discontinuity of the rank. Hence rank $(\hat{\mathscr{E}}_j) \leqq (j-1)n$ on the closure $\overline{\cup \mathscr{I}_i}$ for $j=n+1$. But we already have rank $(\hat{\mathscr{E}}_j) \geqq (j-1)n$ on $\mathscr{I} = \overline{\cup \mathscr{I}_i}$ so that rank $(\hat{\mathscr{E}}_j) = (j-1)n$ on $\mathscr{I}$ and $\hat{\mathscr{E}}_j$ has constant rank if $j=n+1$. Now we can show that $\hat{\mathscr{E}}_j$ is also 1-full. Let $\mathscr{M}$ be the span of the first n standard basis vectors $\{e_1, \cdots, e_n\}$ in C^{jn}. Notice that

(3.14) $\hat{\mathscr{E}}_j$ is 1-full if and only if $\mathscr{M} \perp \mathscr{N}(\hat{\mathscr{E}}_j)$, where $\mathscr{N}$ denotes nullspace

and $\perp$ is determined by the usual inner product on C^{jn}. Suppose $j=n+1$. Since $\hat{\mathscr{E}}_j$ has constant rank there exists a continuous basis $\psi_1, \cdots, \psi_s$ for its nullspace. But then $e_k^* \psi_r = 0$ on $\cup \mathscr{I}_i$ for $1 \leqq k \leqq n$, $1 \leqq r \leqq s$ implies, by continuity, that $e_k^* \psi_j = 0$ on $\overline{\cup \mathscr{I}_i}$ which is $\mathscr{I}$. Thus by Lemma 3.1 and (3.14), we have that (3.10) holds. □

Theorem 3.1 has several important consequences.

COROLLARY 3.1. *If* (1.1) *is a sufficiently smooth solvable system, then the numerical methods of* [7], [8] *can be used, in principle, to integrate it.*

Thus the method of [7], [8] is in fact a general method for (1.1).

COROLLARY 3.2. *Suppose E, F are n-times differentiable where $m \geqq 2n$ and* (1.1) *is a solvable system. If $m+1-n \geqq j \geqq n+1$, then $\eta = \dim (\mathscr{N}(\mathscr{E}_j(t)))$ is independent of t and j. Furthermore, for any given n-times differentiable f, the solutions to* (1.1) *form an $(n-\eta)$-dimensional manifold.*

From [9] we now have the following theorem.

THEOREM 3.2. *Suppose that* (1.1) *is solvable, E, F are $2n$-times differentiable and $n+1 \geqq j \geqq 1+($size of N in* (2.6b)). (*In particular, $j=1+n$ suffices.*) *Then the linear manifold of consistent initial conditions is precisely the set of all x_0 such that $\mathscr{E}_j(t_0)\mathbf{x} = f_j(t_0) - \mathscr{F}_j(t_0)x_0$ is consistent. Equivalently, $f_j(t_0) - \mathscr{F}_j(t_0)x_0$ is in the range of $\mathscr{E}_j(t_0)$. Thus one may determine the manifold of consistent initial conditions by*

> *Computing W such that $\mathscr{E}_j^* W = 0$ and rank $W =$ nullity of $\mathscr{E}_j^*$;*
> *Solving $W^* \mathscr{F}_j x_0 = W^* f_j$ for x_0.*

We have then that if (1.1) is solvable it is straightforward, if time consuming, to compute the initial conditions and integrate the equations.

However, none of the development so far, other than Theorem 2.1, tells us whether (1.1) is solvable on $\mathscr{I}$ and the actual implementation of (2.6) is difficult. We now address this problem by deriving algebraic criteria on $\mathscr{E}_j$, $\mathscr{F}_j$ to ensure solvability. The criteria are to be numerically verifiable. The next result is a partial converse to Theorem 3.2.

THEOREM 3.3. *The system* (1.1) *with E, F, $n \times n$ and real analytic is solvable if and only if there is an integer $j_0 \in [1, n+1]$ such that*

(3.15) Rank $(\mathscr{E}_{j_0})$ *is constant on* $\mathscr{I}$,

(3.16) $\mathscr{E}_{j_0}$ *is* 1-*full at each* $t \in \mathscr{I}$,

(3.17) $\mathscr{R}(\mathscr{E}_{j_0}(t)) + \mathscr{R}(\mathscr{F}_{j_0}(t)) = C^{nj}$ *on* $\mathscr{I}$.

Proof [Only if]. The only if part follows from Theorem 3.1 and the observations that all transformations may be replaced by analytic ones. In particular the Q, R of (3.5) will be analytic so that we get that if f is analytic, then so is the solution x.

Lemma 3.1 goes through for any level of smoothness. The next lemma uses real analyticity in an essential way and will be used to prove the sufficiency of (3.15)–(3.17).

LEMMA 3.2. *Suppose that E, F, f are real analytic and $\mathscr{E}_j$ has constant rank on $\mathscr{I}$. If $\mathscr{E}_j$ is* 1-*full for every $t \in \mathscr{I}$ and if* (1.1) *is solvable on any subinterval $\tilde{\mathscr{I}}$ of $\mathscr{I}$, then* (1.1) *is solvable on* $\mathscr{I}$.

Proof. Assume that $\mathscr{C}_j$ is 1-full for every $t \in \mathscr{I}$ and has constant rank. Then by Lemma 2.1 it is smoothly 1-full. Suppose that (1.1) is solvable on $\tilde{\mathscr{I}} \subset \mathscr{I}$. Let x be a solution of (1.1) on $\tilde{\mathscr{I}}$. Pick $t_0 \in \tilde{\mathscr{I}}$. Let $\bar{x}$ be the solution of $x' = Qx + q$ for $t \in \mathscr{I}$ such that $\bar{x}(t_0) = \tilde{x}(t_0)$. Then $\bar{x}$ is analytic on $\mathscr{I}$ and $\bar{x} = x$ for $t \in \tilde{\mathscr{I}}$ since $x' = Qx + q$ for $t \in \tilde{\mathscr{I}}$. Thus $E\bar{x}' + F\bar{x} = f$ for $t \in \tilde{\mathscr{I}}$. Since E, $\bar{x}$, F, f are all analytic on $\mathscr{I}$, this relationship must hold for all t by analytic continuation and $\bar{x}$ is a solution of (1.1) on $\mathscr{I}$. This shows that solutions of (1.1) exist for any analytic f. The uniqueness follows from the fact they are also solutions of (3.5). Thus (1.1) is solvable on $\mathscr{I}$ for real analytic f. Then using the existence of the SCF on $\mathscr{I}$ gives solvable for $f \in \mathscr{C}_{2n}$. □

We can now complete the proof of Theorem 3.3.

Proof of Theorem 3.3 [If part]. If the system (1.1) is one-dimensional then the theorem is true. Suppose then that (1.1) is a minimal-dimensional counterexample to Theorem 3.3. That is, (1.1) satisfies (3.15)-(3.17), but is not solvable on an interval $\mathscr{I}$. By Lemmas 3.1, 3.2, it is also a counterexample on any nontrivial subinterval $\tilde{\mathscr{I}} \subset \mathscr{I}$. Note that E, F are $n \times n$ matrices with $n \geqq 2$. Since (1.1) is assumed not solvable, E must be singular for some $t \in \mathscr{I}$. But (3.15) then implies E is singular for all $t \in \mathscr{I}$. Let p be a positive integer such that rank $E \leqq p < n$ for all $t \in \mathscr{I}$. Then there exists analytic nonsingular H such that

$$HE = \begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix}$$

where E_{11} is $p \times p$. We now use the fact that (3.15)-(3.17) are unaffected by changes of coordinates. Multiplying (1.1) by H gives the equivalent system

$$\begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} z' \\ w' \end{bmatrix} + \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\begin{bmatrix} z \\ w \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix} \tag{3.18}$$

where (3.15)-(3.17) still hold. But if (3.17) holds, then it must be that $\mathscr{R}(E) + \mathscr{R}(F) = \mathrm{C}^n$ and similarly for (3.18). Thus

$$\mathscr{R}\left(\begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix}\right) + \mathscr{R}\left(\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\right) = \mathrm{C}^n$$

and $[F_{21} \quad F_{22}]$ must be identically full row rank. Hence there exists a nonsingular real analytic $Q(t)$ such that $[F_{21} \quad F_{22}]\ Q = [0 \quad \tilde{F}_{22}]$ with $\tilde{F}_{22}$ invertible. Letting

$$\begin{bmatrix} z \\ w \end{bmatrix} = Q\begin{bmatrix} u \\ v \end{bmatrix}$$

changes (3.18) to the equivalent system

$$\begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} u' \\ v' \end{bmatrix} + \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \\ 0 & \tilde{F}_{22} \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix}. \tag{3.19}$$

But by (3.17) $\tilde{F}_{22}$ must be nonsingular. Multiply (3.19) by

$$\begin{bmatrix} I & -\tilde{F}_{12}\tilde{F}_{22}^{-1} \\ 0 & \tilde{F}_{22}^{-1} \end{bmatrix}$$

to give

$$\begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} u' \\ v' \end{bmatrix} + \begin{bmatrix} \tilde{F}_{11} & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \tilde{g} \\ \tilde{h} \end{bmatrix}. \tag{3.20}$$

Take j to be the j_0 of (3.15)-(3.17). Then for (3.20) the $\mathscr{E}_j$, $\mathscr{F}_j$ take the form

$$(3.21)\qquad \mathscr{E}_j=\begin{bmatrix} \tilde{E}_{110} & \tilde{E}_{120} & & & & & \\ 0 & 0 & & & & & \\ \tilde{E}_{111}+\tilde{F}_{110} & \tilde{E}_{120} & 2\tilde{E}_{110} & 2\tilde{E}_{120} & & & \\ 0 & I & 0 & 0 & & & \\ \vdots & & & & \ddots & & \\ \tilde{E}_{11j-1}+\tilde{F}_{11j-2} & \tilde{E}_{12j-1} & & & & j\tilde{E}_{110} & j\tilde{E}_{120} \\ 0 & 0 & \cdots & & I & 0 & 0 \end{bmatrix}$$

and

$$(3.22)\qquad \mathscr{F}_j=\begin{bmatrix} \tilde{F}_{110} & 0 \\ 0 & I \\ \tilde{F}_{111} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ \tilde{F}_{11j-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

By assumption $\mathscr{R}(\mathscr{E}_j)+\mathscr{R}(\mathscr{F}_j)=C^{jn}$. Let u, v be any two vectors such that $\mathscr{E}_j u+\mathscr{F}_j v\in C^p\oplus 0^{jn-p}\subseteq C^{jn}$ where $C^n\oplus 0^s$ means n arbitrary entries followed by s zero entries. The fourth row of $\mathscr{E}_j$ in (3.21) shows that the $p+1$ through n entries of u are zero and thus $\mathscr{R}(\tilde{\mathscr{E}}_{110})+\mathscr{R}(\tilde{\mathscr{F}}_{110})=C^p$. Pick a point $t_0\in\mathscr{I}$. Then there is an analytic matrix $Q(t)$ such that $\tilde{E}_{110}Q(t)+\tilde{F}_{110}$ is invertible on a neighborhood of t_0. Let U be the solution of $U'=QU$, $U(t_0)=I$, let $u=U\tilde{u}$, and restrict (3.20) to the subinterval $\tilde{\mathscr{I}}$. We still have a counterexample to Theorem 3.3 except now $\tilde{F}_{11}$ in (3.20) is nonsingular. Multiplying by $\tilde{F}_{11}^{-1}$ we finally get that our counterexample is

$$(3.23)\qquad \begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} u' \\ v' \end{bmatrix}+\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix}=\begin{bmatrix} g \\ h \end{bmatrix}$$

(with new E_1, E_2, u, v, g). Also now

$$(3.24)\qquad \mathscr{E}_j=\begin{bmatrix} E_{10} & E_{20} & & & & & \\ 0 & 0 & & & & & \\ E_{11}+I & E_{21} & 2E_{10} & 2E_{20} & & & \\ 0 & I & 0 & 0 & \ddots & & \\ \vdots & & & & & & \\ E_{1j-1} & E_{2j-1} & \cdots & (j-1)E_{11}+I & (j-1)E_{21} & jE_{10} & jE_{20} \\ 0 & 0 & \cdots & 0 & I & 0 & 0 \end{bmatrix},$$

$$\mathscr{F}_j=\begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Notice that if

$$(3.25)\qquad E_1u'+u=g$$

is solvable, then so is (3.18), which is a contradiction. Since E_1 has smaller dimensions than E it suffices to show (3.15)-(3.17) hold for (3.25) to obtain a contradiction. Using the I blocks in (3.24) and doing row operations on $\mathscr{E}_j$ to zero the remaining entries in those columns we see that

$$\tilde{\mathscr{E}}_j = \begin{bmatrix} E_{10} & & & & 0 \\ E_{11}+I & 2E_{10} & & & 0 \\ \vdots & & \ddots & & \vdots \\ E_{1j-1} & & & jE_{10} & jE_{20} \end{bmatrix} = [\Sigma_j|\phi],$$

$$\tilde{\mathscr{F}}_j = \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.26}$$

is 1-full iff (3.24) is. But $\tilde{\mathscr{E}}_j$ 1-full implies Σ_j is 1-full and Σ_j is the $\mathscr{E}_j$ matrix for (3.25). Thus (3.16) holds for (3.25). Since Σ_j is real analytic we may take a subinterval on which it has constant rank so that (3.15) holds for (3.25). There remains only to verify (3.17) for (3.25) to complete the proof. We may assume $j \geqq 2$. Then by assumption on (3.24), (3.17) holds. Observe that then one can find u, v so that $\mathscr{E}_j u + \mathscr{F}_j v$ gives any vector in $\sum_{i=1}^{j} \oplus (\mathrm{C}^p \oplus o^{n-p})$. The last row implies that the $jn-2$ block entry of u is zero. Let $\hat{\mathscr{E}}_j$ be $\tilde{\mathscr{E}}_j$ with the $(j-1)E_{20}$ entry zero. Then this says $\mathscr{R}(\hat{\mathscr{E}}_j) + \mathscr{R}(\tilde{\mathscr{F}}_j) = \mathrm{C}^{jp}$. But $\mathscr{R}(\Sigma_{j-1}) + \mathscr{R}(\mathscr{F}_{j-1}) = \mathrm{C}^{(j-1)p}$ since Σ_{j-1} is the $n(j-1) \times n(j-1)$ principal submatrix of $\hat{E}_j$. As noted Σ_j is the $\mathscr{E}_j$ matrix for (3.24). Thus $\mathscr{R}(\mathscr{E}_i) + \mathscr{R}(\mathscr{F}_i) = \mathrm{C}^{ni}$ for (3.25) for $i = j_0 - 1 = n$. Since E_1 has dimension less than or equal to $n-1$, we have (3.25) satisfies (3.15), (3.16), (3.17) which is a contradiction. □

COROLLARY 3.3. *Suppose that E, F, f are real analytic on $\mathscr{I}$. Then* (1.1) *is solvable on $\mathscr{I}$ if and only if* (3.15) *holds on $\mathscr{I}$,* (3.17) *holds at t_0, and* (3.16) *holds on a dense subset of $\mathscr{I}$.*

Proof. If (3.17) holds at t_0, it will hold in some subinterval $[t_0, t_0+\varepsilon]$. Now use Lemma 3.2, Theorem 3.3, and (3.15). □

We can extend the proof of Theorem 3.2 to cover infinitely differentiable functions. To do this, observe first that if E, F are infinitely differentiable and (1.1) is solvable on $\mathscr{I}$, then by considering only infinitely differentiable solutions one may get (2.11) where $Nw' + w = f_2$ has only one infinitely differentiable solution for each infinitely differentiable f_2. However, if $Nw' + w = 0$ and w is not infinitely differentiable it follows that $w = 0$ on $\bigcup \mathscr{I}_i$ and hence $w = 0$ on $\mathscr{I}$. Also, as noted earlier if E, F, f are infinitely differentiable, then so are Q, q and hence x. Thus for E, F infinitely differentiable, we may consider solvability to mean that x, f are also infinitely differentiable. This version of Theorem 2.1 is actually slightly weaker, since solvability is taken to mean for every infinitely differentiable f, (1.1) has a solution. This does not immediately imply that for every n-times differentiable f, (1.1) will have a smooth solution. If we call this weaker form of solvability $\mathscr{C}^\infty$-solvability, then Theorem 2.1 becomes the following.

THEOREM 3.4. *If* (1.1) *is $\mathscr{C}^\infty$-solvable, then* (1.1) *is equivalent to* (2.6) *where all of the functions including g, h are infinitely differentiable.*

Using Theorem 3.4, we may generalize Theorem 3.3 as follows.

THEOREM 3.5. *The system* (1.1) *with E, F infinitely differentiable is $\mathscr{C}^\infty$-solvable if and only if there exists a $j_0 \geqq n+1$ such that* (3.15)-(3.17) *hold.*

Proof. The "only if" part follows from Theorem 3.1. To prove the "if" part, suppose that E, F, f are infinitely differentiable on $\mathscr{I}$ and that (3.15)-(3.17) hold for

$j_0 \geqq n+1$. There are a countable number of open intervals $\mathcal{I}_i$ such that $\bigcup \mathcal{I}_i$ is dense in $\mathcal{I}$ and rank E is constant on each $\mathcal{I}_i$. Premultiplication (on each $\mathcal{I}_i$) gives

$$\begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix} x' + \begin{bmatrix} F_1 & F_2 \\ F_3 & F_4 \end{bmatrix} x = f.$$

By (3.17), the matrix $[F_3, F_4]$ has full row rank and we may do a change of coordinates $x = Vy$ to give

$$\begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix} y' + \begin{bmatrix} F_1 & F_2 \\ 0 & I \end{bmatrix} y = f$$

with new E_i, F_i. A premultiplication gives finally that

$$\begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix} y' + \begin{bmatrix} F_1 & 0 \\ 0 & I \end{bmatrix} y = \tilde{f}.$$

As shown in [7] these coordinate changes do not affect (3.15)-(3.17). Thus $E_1 y_1' + F_1 y_1 = g$ satisfies (3.17) with $\mathrm{j}_0 \leqq n$. Continuing in this manner we get that there exists a countable family of disjoint open intervals $\mathcal{I}_i$ such that $\bigcup \mathcal{I}_i$ is dense in $\mathcal{I}$ and on each $\mathcal{I}_i$ there exists P_i, V_i so that premultiplication by P_i and letting $x = V_i y$ transfers (1.1) to standard canonical form [6] so that (1.1) is solvable on each $\mathcal{I}_i$. The P_i, V_i are defined only on $\mathcal{I}_i$.

By assumption on $\mathscr{E}_j$, $\mathscr{F}_j$ there exists smooth nonsingular $\mathscr{R}$ on all of $\mathcal{I}$ so that

$$R[\mathscr{E}_j \mid \mathscr{F}_j] = \left[\begin{array}{c|c} X & Y \\ \hline 0 & \theta \end{array}\right]$$

with X of full row rank. Now (1.1) is solvable on each $\mathcal{I}_i$ and from [2], [8] we have the solutions are given by (3.5) and the constraint matrix $\theta(t)$:

$$\dot{x} = Q(t)x(t) + q(t), \tag{3.27}$$

$$0 = \theta(t)x(t) + h(t) \tag{3.28}$$

and g, h are differential operators applied to f,

$$g(t) = \sum_{i=0}^{n} \phi_i(t) f^{(i)}(t), \qquad h(t) = \sum_{i=0}^{n} \psi_i(t) f^{(i)}(t).$$

By assumption on $\mathscr{E}_j$, $\mathscr{F}_j$, the coefficients $Q(t)$, $\theta(t)$, $\phi_i(t)$, $\psi_i(t)$ are in $\mathscr{C}^\infty[\mathcal{I}]$. Also since $\theta(t)$ has full row rank on $\mathcal{I}$ by (3.15), (3.16), (3.17), there exists a change of coordinates W defined on $\mathcal{I}$ so that $\theta(t)W(t) = [0 \ \ I]$. Then (3.27), (3.28) becomes

$$x_1' = Q_{11}x_1 + Q_{12}x_2 + q_1, \tag{3.29}$$

$$x_2' = Q_{21}x_1 + Q_{22}x_2 + q_2, \tag{3.30}$$

$$0 = x_2 + h \tag{3.31}$$

where $[Q_{ij}] = Q^{-1}(QW - W')$, $[q_i] = W^{-1}q$. But (3.29)-(3.31) characterize the solutions of (1.1) on $\mathcal{I}_i$. Letting $f = 0$ in (3.29), (3.30), (3.31) gives, on each $\mathcal{I}_i$,

$$x_1' = Q_{11}x_1, \qquad 0 = Q_{21}x_1.$$

However, $\dim x_1$ is the dimension of the solution manifold on $\mathcal{I}_i$ so $Q_{21} = 0$ on $\bigcup \mathcal{I}_i$. By continuity $Q_{21} \equiv 0$ on $\mathcal{I}$. Returning to (3.29)-(3.31) we now have

$$x_1' = Q_{11}x_1 - Q_{12}h + q_1, \tag{3.32}$$

$$0 = -Q_{22}h + q_2 - h'. \tag{3.33}$$

Given an $f \in \mathscr{C}^\infty[\mathscr{I}]$, by solvability of (3.23) on $\mathscr{I}_i$, we have (3.33) trivially holds on $\bigcup \mathscr{I}_i$. But then (3.33) holds on $\mathscr{I}$ by continuity. Thus (3.32), (3.33), and hence (3.19)-(3.21) have dim x_2 linearly independent solutions defined on all of $\mathscr{I}$. But by construction these solutions satisfy (1.1) on $\bigcup \mathscr{I}_i$. By continuity they satisfy (1.1) on $\mathscr{I}$. □

Observe that if we let

$$\mathscr{L}x = Ex' + Fx, \quad \mathscr{H}x = x' - Qx, \quad \mathscr{R}x = \sum_i R_{1i}x^{(i)}$$

then we can write

$$\mathscr{R}\mathscr{L} = \mathscr{H}. \tag{3.34}$$

Also this proof does not actually require infinite differentiability. Being $3n$-times differentiable would suffice.

4. Comments. This paper provides a fairly complete theory for (1.1) based on the ideas of [7]. In particular, the numerical approach of [7], [8] works on all smooth solvable systems. For infinitely differentiable coefficients, it works if and only if the system is solvable. The conclusion (3.34) implies that the operator L is what Emre calls left-admissable [11]. Examining this connection with [11] would be a major digression and will be done elsewhere.

One difficulty in working with (1.1) is that a system $Ex' + Fx = f$ with E, F close in the $\|\cdot\|_m$ norm to E, F need not be solvable, and if solvable, need not have solutions close to those of (1.1). However, using the differential equation (3.5), Theorem 3.1 and Corollary 3.2, we can now make the following statement (Theorem 4.1). While an immediate consequence of § 3, this is the first result of its kind we are familiar with.

THEOREM 4.1. *Suppose that E^τ, F^τ, E, F are in $\mathscr{C}_{2n}$ where τ is a parameter. Suppose that $E^\tau \to E$, $F^\tau \to F$ in $\mathscr{C}_{2n}$ as $\tau \to \tau_0$. Suppose that $E^\tau x' + F^\tau x = f$ is solvable for each t and that for $j = n+1$*, rank $(\mathscr{C}_j^\tau) =$ rank $(\mathscr{C}_j)$ *for all $t \in \mathscr{I}$. Then $Ex' + Fx = f$ is solvable. Furthermore, the manifolds of initial conditions and solutions are also continuous in τ at τ_0.*

Proof. The only part that needs proof is that $\mathscr{C}_j$ is 1-full. But this follows in the same way we extended 1-fullness to the closure of $\bigcup \mathscr{I}_i$ in Theorem 2.1. □

COROLLARY 4.1. *Suppose* (1.1) *is equivalent to a system in* SCF. *Then there exist polynomial matrices (in t) E^τ, F^τ such that $E^\tau \to E$, $F^\tau \to F$ in $\mathscr{C}_{2n}$ as $\tau \to \tau_0$. $E^\tau x' + F^\tau x = f$ is a solvable system for every τ, and the solution manifold is continuous in τ at τ_0.*

Proof. Take $\mathscr{C}_{2n}$ polynomial approximations of both the SCF and the coordinate changes that put it into SCF. □

This argument does not work if (1.1) is not equivalent to a system in SCF. We do not know if the polynomial solvable systems are dense (in the sense of Corollary 4.1) in all solvable systems.

Theorem 3.3 (or Corollary 3.3), to our knowledge, supplies the first explicit characterization of solvability. However, it appears to involve a great deal of computation. We shall now show how this may usually be done simultaneously with the solution at little additional overhead.

Suppose then that we have (1.1) on $[t_0, t_1]$ with E, F, f real analytic or infinitely differentiable. At t_0, rank $(\mathscr{C}_j(t_0)) = r_0$ is determined, possibly by a singular value decomposition (SVD) and (3.16), (3.17) are also verified. This SVD may also be used to estimate the conditioning of (3.4). Suppose that we now proceed to numerically solve (1.1) by integrating (3.5) as described in [8]. First a QR factorization is performed on the first n columns of $\mathscr{E}_j$ to yield

$$Q^T\mathscr{E}_j=\begin{bmatrix} R_{11} & H_{12} \\ 0 & H_{22}\end{bmatrix}, \qquad Q^T\mathbf{f}_j=\begin{bmatrix} f_1 \\ f_2\end{bmatrix} \tag{4.1}$$

and R_{11} is nonsingular and upper triangular. If this cannot be done, then we could use an SVD. If this still fails to produce a nonsingular R_{11}, then either the problem is not 1-full or is too numerically ill conditioned. Suppose then we have that (3.4) is now (4.1). Now perform a QR on H_{22} using column pivoting so that (3.4) is now

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & \varepsilon_1 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ z_1 \\ z_2 \end{bmatrix}=\begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \varepsilon_2 \end{bmatrix} \tag{4.2}$$

where R_{22} is nonsingular, upper triangular, and $(r_0-n)\times(r_0-n)$. If R_{22} is not that size, then use an SVD instead of a QR. If R_{22} is still not the correct size, either we are near a place where (3.15) is violated or the problem is too ill conditioned to proceed directly (note [23]). Similarly, if ε_1, ε_2 are not negligible, then (3.17) is being violated numerically. It is probably not necessary to check 1-fullness at every time step and one can often use the basic solution in (4.2) by setting $z_2=0$ and backsolving for x_1. However, there are a couple of ways to check 1-fullness. One is to use elementary row operations to use R_{22} to zero R_{12}. Then E_j is 1-full if and only if R_{13} is also zeroed out. An alternative is to perform Householder transformations (which need not be saved) on the right of the coefficient matrix in (4.2) to get

$$\begin{bmatrix} R_{11} & \tilde{R}_{12} & \varepsilon_3 \\ 0 & \tilde{R}_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ z_1 \\ z_2 \end{bmatrix}=\begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ 0 \end{bmatrix} \tag{4.3}$$

where $\tilde{R}_{22}$ is again upper triangular. If ε_3 is not negligible, we consider the system to not be 1-full, (3.16) is violated and we stop. If $\varepsilon_3=0$, then we solve (4.3) for z_1 and then x_1 by back substitution to get

$$\tilde{x}_1=-R_{11}^{-1}\tilde{R}_{12}\tilde{R}_{22}^{-1}\hat{f}_2 \tag{4.4}$$

and proceed with our integration scheme.

The decisions require, as is often the case with numerical procedures, making decisions on numerical rank [15], [23]. However, if (1.1) is solvable on $\mathscr{I}$, with smooth enough coefficients, then there exists a precision and step size (depending on the conditioning of the algebraic equations) and the system (3.5) which will permit us to integrate (1.1) (equivalently (3.5)). Conversely, if (1.1) is not solvable on all of $\mathscr{I}$, then there is a stepsize h such that using any stepsize smaller than this will lead to the determination that (3.15) or (3.16) are violated and the problem is not solvable. In any event, the conditioning of the required algebra in (3.4) over $\mathscr{I}$ is independent of the stepsize, whereas the conditioning of the algebraic problems in a backward difference scheme increases as h decreases.

Finally, note that showing the solution x of (1.1) satisfied (3.5) required f to be in $\mathscr{C}_{2n}$ (or $\mathscr{C}_{2r}$ if there was a SCF) since we need $\mathbf{x}_{n+1}$ to exist. However, by taking limits in $\mathscr{C}_1$ one may show that if $f\in\mathscr{C}_n$, then x satisfies (3.5) for some q with the same Q.

REFERENCES

[1] H. BART, M. A. KAASHOEK AND D. C. LAY, *Relative inverses of meromorphic operator functions and associated holomorphic projection functions*, Math. Ann., 218 (1975), pp. 199-210.

[2] K. BRENAN, *Stability and convergence of difference approximation for higher index differential-algebraic systems with applications in trajectory control*, Ph.D. thesis, University of California, Los Angeles, CA, 1983.

[3] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, New York, 1980.

[4] ———, *Singular Systems of Differential Equations* II, Pitman, New York, 1982.

[5] ———, *Index two linear time varying singular systems of differential equations*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 237-243.

[6] ———, *One canonical form for higher index linear time varying singular systems*, Circuits Systems Signal Process., 2 (1983), pp. 311-326.

[7] ———, *The numerical solution of higher index linear time varying systems of differential equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 334-348.

[8] ———, *Rank deficient least squares and the numerical solution of linear singular implicit systems of differential equations*, in Linear Algebra and Its Role in Systems Theory, AMS Cont. Math. Series, 47, 1985, p. 51-64.

[9] ———, *Consistent initial conditions for linear time varying singular systems*, Proc. Conference on the Mathematical Theory of Networks and Systems, Stockholm, 1985, to apppear.

[10] S. L. CAMPBELL AND L. R. PETZOLD, *Canonical forms and solvable singular systems of differential equations*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 517-521.

[11] E. EMRE, *A polynomial fractional approach to linear time-varying systems*, Proc. Allerton Conference, Urbana-Champaign, IL, 1983, pp. 11-20.

[12] C. W. GEAR, *The simultaneous numerical solution of differential-algebraic equations*, IEEE Trans. Circuit Theory, TC-18 (1971), pp. 89-95.

[13]———, *Maintaining solution invariants in the numerical solution of* ODE*s*, Technical Memorandum No. 40, Argonne National Laboratory, 1984.

[14] C. W. GEAR, G. K. GUPTA AND B. LEIMKUHLER, *Automatic integration of Euler-Lagrange equations with constraints*, J. Comput. Appl. Math, to appear.

[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[16] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1982.

[17] P. LOTSTEDT AND L. R. PETZOLD, *Numerical solution of nonlinear differential equations with algebraic constraints*, Sandia report, 83-8877, 1983.

[18] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 312-321.

[19] L. R. PETZOLD AND C. W. GEAR, ODE *methods for the solution of differential/Algebraic systems*, Sandia Report, 82-8051, 1982.

[20] W. C. RHEINBOLDT, *Differential-algebraic systems as differential equations on manifolds*, Math. Comp., 4 (1984), pp. 473-482.

[21] L. M. SILVERMAN, *Inversion of multivariable systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270-276.

[22] L. M. SILVERMAN AND R. S. BUCY, *Generalizations of a theorem of Dolezal*, Math. Systems Theory, 4 (1970), pp. 334-339.

[23] G. W. STEWART, *Rank degeneracy*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 403-413.

[24] H. W. WOHLTMAN, *Target path controllability of linear time-varying dynamical systems*, IEEE Trans. Automat. Control, AC-20 (1985), pp. 84-87.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

015

A SYNTHETIC STUDY OF ONE-PARAMETER NONLINEAR PROBLEMS*

PATRICK RABIER†

Abstract. We present a study on the local structure of the set of solutions of one-parameter nonlinear problems, based on a recent generalization of the Implicit function theorem. We show how the combination of some nondegeneracy condition together with a priori requirements allows one to treat both cases without or with change of scale. The approach we develop is self-contained and provides most of the classical results on "nondegenerate" bifurcation with various improvements and extensions. Also, the suitable change of scale, if any, is shown to be as predicted by Newton's diagrams. Applications to standard model problems in nonlinear P.D.E.'s are given.

Key words. nondegeneracy condition, one-parameter problems, Newton's diagrams

AMS(MOS) subject classification. 58E07

1. Introduction and preliminaries. Let X and Y be two real Banach spaces and $G(=G(\mu, x))$ a mapping of the variables $\mu \in \mathbb{R}$ and $x \in X$ with values in Y such that $G(0)=0$. We intend to show how a generalization of the Implicit function theorem (Theorem 1.1) can be used for finding the zero set of G around the origin. The first step of our method is classical insofar as it makes use of the Lyapunov-Schmidt reduction so that the problem becomes equivalent to a finite-dimensional one, the mapping G being replaced by the so-called reduced mapping. Unfortunately, the assumptions on the reduced mapping involve higher-order derivatives whose expression in terms of G rapidly becomes very intricate. This fact is even more pronounced when a preliminary change of the parameter μ into $\pm\eta^p$ for some suitable integer $p \geqq 2$ is necessary.

Our first aim has been to give a sharp criterion for the generalization of the Implicit function theorem mentioned above to be available under "simple" assumptions on the mapping G, whether the parameter μ is changed or not. What we mean by "simple" will be made precise later in this section and makes explicit a condition already considered desirable by most authors, who have formulated appropriate particular hypotheses in this aim. Of course, this does not mean that there is no problem of interest which does not fulfill our conditions: for instance, bifurcation near a degenerated eigenray does not fit into our framework in general.

The basic form of the criterion we use is established in § 2, but its most useful versions are derived in § 4. The effect of its combination with Theorem 1.1 are briefly examined in § 3 (no change of parameter) and in more detail in § 4 ($\mu = \pm\eta^p$). In both cases, we obtain results of regularity of the curves that are better than the existing ones (when any exist at all) and our study applies beyond the classical framework.

The analysis of § 4 naturally leads to the observation that the value of p in the change of scale not only cannot be arbitrary but is uniquely determined as a rational number (or $-\infty$) that becomes available for our purposes as soon as its value is an integer. Integer or not, an a posteriori verification shows that it is exactly as provided by using Newton's diagrams. Besides, it is related to another integer κ whose meaning in the problem is evident by showing that p must divide κ or $\kappa-1$. Hence, the two cases $p=\kappa$ and $p=\kappa-1$ only bear a general discussion. Heuristically, this explains

* Received by the editors October 8, 1984; accepted for publication (in revised form) June 17, 1986.

† Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique, 4 Place Jussieu, 75230 Paris Cedex 05, France. Present address, Southern Methodist University, Department of Mathematics, Dallas, Texas 75275.

why they have already attracted attention in general problems (especially $p=\kappa-1$ in the study of problems of bifurcation from the trivial branch). Note however that noninteger values of p have sometimes been successfully used; for instance in problems of bifurcation from the trivial branch when the linearized operator has a nontrivial generalized null-space (see Landman and Rosenblat [10]).

The location of the curves of solutions of the equation $G(\mu, x)=0$ with respect to the hyperplane $\mu=0$ in $\mathbb{R}\times X$ is fully discussed from the parity of p. Our conclusions cover the concepts of "regular," "turning" or "hysteresis" point as well as "supercritical," "subcritical" or "transcritical" bifurcation point.

This paper contains conclusions related to those of Crandall and Rabinowitz [4], McLeod and Sattinger [11], Magnus [12], Szulkin [19] and Buchner, Marsden and Schecter [3], which follow with various complements and extensions. Problems beyond their domain of validity (including standard ones) can also be treated. The results we present are then fairly general though accurate. It is noteworthy that they all derive from the criterion of § 2 in conjunction with the nondegeneracy condition of Theorem 1.1: the whole theory thus reduces to these two basic aspects after technical manipulations which are precisely the purpose of §§ 3 and 4. Related computational algorithms are described in [15], [16].

Due to the abundant literature devoted to the subject, special attention has been given to problems of bifurcation from the trivial branch, although existence of a known branch plays no role in this approach. In particular, nonexistence of a nontrivial generalized null-space (i.e., direct sum of the null-space and the range) for the linearized operator is shown to be a necessity imposed by the standing nondegeneracy condition.

Examples of application to nonlinear partial differential equations are given in § 5. Other examples of interest in which the usual properties of the mapping G may significantly differ are, for instance, concerned with the determination of steady solutions to systems of ordinary differential equations.

We shall now develop a few preliminaries. Throughout this paper, the mapping G is supposed to be of class $\mathscr{C}^m$, $m\geqq 1$, on a neighbourhood of the origin in $\mathbb{R}\times X$. We also assume that the partial derivative $D_xG(0)$ is a Fredholm operator with index 0, namely, that the space

$$X_1=\operatorname{Ker} D_xG(0) \tag{1.1}$$

is finite dimensional, and that the space

$$Z_2=\operatorname{Range} D_xG(0) \tag{1.2}$$

(is closed and) has finite codimension with

$$\dim X_1=\operatorname{codim} Z_2=N\geqq 0. \tag{1.3}$$

Let X_2 and Z_1 be two topological complements of X_1 and Z_2, respectively,

$$X=X_1\oplus X_2, \qquad Y=Z_1\oplus Z_2, \tag{1.4}$$

so that $\dim Z_1=\operatorname{codim} Z_2=N$ and

$$D_xG(0)|_{X_2}\in \operatorname{Isom}(X_2, Z_2). \tag{1.5}$$

Note. A previous version of this work, announced in [14], required $X=Y$ and made use of generalized null-spaces. These restrictions have disappeared in the present exposition.

Here is the basic tool we use in §§ 3 and 4 for determining the local zero set of the mapping G. A slightly more general statement can be found in [14, Thm. 3.2] and, with a loss of one degree of regularity at the origin, is also proved in Buchner et al. [3].

THEOREM 1.1. *Let $n \geqq 0$ be a given integer and $\tilde{X}_1$ and Y_1 two real vector spaces with dimension $n+1$ and n, respectively. On the other hand, let f be a mapping of class $\mathscr{C}^m$, $m \geqq 1$, defined on a neighbourhood of the origin in $\tilde{X}_1$ with values in Y_1. Assume that*

$$(1.6) \qquad D^j f(0) = 0, \qquad 0 \leqq j \leqq k-1,$$

for some $1 \leqq k \leqq m$. Denoting by q the polynomial mapping

$$(1.7) \qquad q : \tilde{x}_1 \in \tilde{X}_1 \to q(\tilde{x}_1) = D^k f(0) \cdot (\tilde{x}_1)^k \in Y_1,$$

assume further that for every nonzero solution $\tilde{x}_1 \in \tilde{X}_1$ of the equation $q(\tilde{x}_1) = 0$ the derivative $Dq(\tilde{x}_1) \in \mathscr{L}(\tilde{X}_1, Y_1)$ is onto.

Then, the zero set of the mapping q in the whole space $\tilde{X}_1$ is the union of a finite number $0^n \leqq j \leqq k^n$[1] of lines through the origin and the zero set of the mapping f around the origin of $\tilde{X}_1$ consists of exactly ν curves of class $\mathscr{C}^m$ away from the origin and $\mathscr{C}^{m-k+1}$ at the origin, where each of them is tangent to a different one of the lines in the zero set of the mapping q (1.7).

Remark 1.1. (i) When $n=0$, Theorem 1.1 applies with any $1 \leqq k \leqq m$ but the best result is obtained with $k=1$: Theorem 1.1 is then nothing but the Implicit function theorem, a fact somewhat hidden by the obviousness of the situation. Conversely, if $k=1$, Theorem 1.1 applies with $n=0$ only since the mapping q (1.7) is the trivial one $q=0$. Thus, assuming $n>1$ implicitly requires $k \geqq 2$.

(ii) When $n=1$ and $k=2$, Theorem 1.1 gives a result which is better than when using the classical Morse lemma (providing $\mathscr{C}^{m-2}$ regularity at the origin instead of $\mathscr{C}^{m-1}$). However, in this case, Theorem 1.1 follows from an improved version of the Morse lemma that can be found in Kuiper [9].

Remark 1.2. The assumption we make on the mapping q (1.7) in Theorem 1.1 will be referred to as "condition of $\mathbb{R}$-nondegeneracy" and abbreviated as ($\mathbb{R}$-N.D.) according to the denomination used in [14]. Some comments on how it can be checked in practice are given in [3], [15].

Remark 1.3. As an odd continuous mapping from $\mathbb{R}^m$ into $\mathbb{R}^n$ always vanishes on the unit sphere of $\mathbb{R}^m$ when $m>n$, we find that $\nu \geqq 1$ when k is odd. When k is even, Buchner et al. have proved that ν is even too [3, Thm. 2.7].

We shall use Theorem 1.1 in the following way: Let $\tilde{X}$ be another real Banach space (in practice, $\tilde{X}$ will be $\mathbb{R} \times X$ but the real variable will not always be μ) and H a mapping of class $\mathscr{C}^m$, $m \geqq 1$, on a neighbourhood of the origin in $\tilde{X}$ with values in Y such that $H(0)=0$. Assume that $DH(0)$ is a Fredholm operator with index 1, namely that

$$(1.8) \qquad \tilde{X}_1 = \operatorname{Ker} DH(0)$$

is finite dimensional, and

$$(1.9) \qquad Y_2 = \operatorname{Range} DH(0),$$

(is closed and) has finite codimension with

$$(1.10) \qquad \operatorname{codim} Y_2 = n \geqq 0, \qquad \dim \tilde{X}_1 = n+1.$$

[1] With $0^0 = 1$ as an understanding.

Let $\tilde{X}_2$ and Y_1 be two topological complements of $\tilde{X}_1$ and Y_2, respectively. Denoting by P_1 and P_2 the (continuous) projection operators onto Y_1 and Y_2, respectively, and after Lyapunov-Schmidt reduction, finding the local zero set of H near the origin amounts to finding the local zero set of the reduced mapping

$$\tilde{x}_1 \in \tilde{X}_1 \to f(\tilde{x}_1) = P_1 H(\tilde{x}_1 + \tilde{\phi}(\tilde{x}_1)) \in Y_1, \tag{1.11}$$

where the function $\tilde{\phi}$ with values in $\tilde{X}_2$ and of class $\mathscr{C}^m$ around the origin is characterized through the Implicit function theorem as solving the equation

$$P_2 H(\tilde{x}_1 + \tilde{\phi}(\tilde{x}_1)) = 0. \tag{1.12}$$

The local zero set of f can be determined by using Theorem 1.1. Nevertheless, in proportion as the index j grows, the derivative $D^j f(0)$ becomes more and more complicated in terms of H because of the increasing number of derivatives of $\tilde{\phi}$ involved in its expression. As a result, writing down the corresponding assumptions on H and checking that they are satisfied seems to be impossible even for relatively small values of k.

However, there is no difficulty if $k = 2$: indeed, implicit differentiation of (1.12) shows that $D\tilde{\phi}(0) = 0$ and it follows that the first two derivatives of the mappings f and $P_1 H(\tilde{x}_1)$ at the origin coincide. This raises the question of whether a simple criterion can be found ensuring that the assumptions of Theorem 1.1 can be checked with the mapping $P_1 H(\tilde{x}_1)$ instead of the reduced mapping. In this case, $D^j f(0) = P_1 D^j H(0)|_{(\tilde{X}_1)^j}$, $0 \leqq j \leqq k$, a particularly simple expression in terms of H. Such a criterion will be established in § 2. In § 3, we take $\tilde{X} = \mathbb{R} \times X$ with generic element (μ, x) and $H = G$. Next, observe that if p is an odd integer, the solutions of the equation $G(\eta^p, x) = 0$ are in one-to-one onto correspondence with the solutions of the equation $G(\mu, x) = 0$ since the mapping $\eta \to \eta^p$ is a $\mathscr{C}^\infty$ homeomorphism. If p is even, the solutions of the equation $G(\eta^p, x) = 0$ provide those of $G(\mu, x) = 0$ with $\mu \geqq 0$ only. In order to get the solutions with $\mu \leqq 0$, we must consider the equation $G(-\eta^p, x) = 0$ as well. That is why, in § 4, setting $\sigma = \pm 1$ we take $\tilde{X} = \mathbb{R} \times X$ with generic element (η, x) and $H(\eta, x) = G(\sigma\eta^p, x)$. It is shown that the integer p must have specific values and that the criterion of § 2 keeps a simple form in terms of G, providing other assumptions made in the literature as particular cases. The usefulness of our generalization is exemplified in § 5 on a model problem. Also, it will be obvious that both choices $\sigma = 1$ and $\sigma = -1$ are equivalent when p is odd ($\geqq 3$). For this reason the case when the parameter p is unchanged will be referred to as the "case $p = 1$" for the sake of convenience.

2. A criterion for the first nonzero derivative. Let the mapping H be as in § 1. Assuming $n = \operatorname{codim} Y_2 = \dim Y_1 \geqq 1$, we shall give a criterion for the first nonzero derivative at the origin of the reduced mapping f (1.11) to coincide with the first nonzero derivative at the origin of the mapping $P_1 H(\tilde{x}_1)$.

Given any integer $0 \leqq l \leqq m$ we define the quantities

$$k_1(l) = \min\{0 \leqq j \leqq l,\ P_1 D^j H(0) \neq 0\}, \tag{2.1}$$

$$k^1(l) = \min\{0 \leqq j \leqq l,\ D^j H(0)|_{(\tilde{X}_1)^j} \neq 0\}. \tag{2.2}$$

In a moment, we shall also need to consider

$$k_2^1(l) = \min\{0 \leqq j \leqq l,\ P_2 D^j H(0)|_{(\tilde{X}_1)^j} \neq 0\}. \tag{2.3}$$

In the definitions (2.1)-(2.3), it is understood that $\min(\varnothing) = +\infty$ and it is then immediate that $k_1(l)$ (resp., $k^1(l)$, $k_2^1(l)$) $= +\infty$ if the derivatives of the mapping $P_1 H$ (resp., $H|_{\tilde{X}_1}$, $P_2 H|_{\tilde{X}_1}$) vanish up to order l, whereas $k_1(l)$ (resp., $k^1(l)$, $k_2^1(l)$) $\leqq l$ otherwise.

As a first step, we establish a simple estimate on the mapping $\tilde{\phi}$.

LEMMA 2.1. *Around the origin in the space $\tilde{X}_1$, one has*

$$\|\tilde{\phi}(\tilde{x}_1)\| = O(\|P_2H(\tilde{x}_1)\|), \tag{2.4}$$

and

$$\|P_2H(\tilde{x}_1)\| = O(\|\tilde{\phi}(\tilde{x}_1)\|), \tag{2.5}$$

so that the derivatives at the origin of order $\leqq m$ of the mappings $\tilde{\phi}$ and $P_2H|_{\tilde{X}_1}$ vanish up to the same order.

Proof. The mapping $\tilde{\phi}$ is characterized by the condition (cf. (1.12))

$$P_2H(\tilde{x}_1 + \tilde{\phi}(\tilde{x}_1)) = 0.$$

Hence, from Taylor's formula

$$0 = P_2H(\tilde{x}_1) + \int_0^1 P_2DH(\tilde{x}_1 + t\tilde{\phi}(\tilde{x}_1)) \cdot \tilde{\phi}(\tilde{x}_1)\,dt. \tag{2.6}$$

Since the mapping H is of class $\mathscr{C}^1$ at least, the mapping

$$\Lambda(\tilde{x}_1) = \int_0^1 P_2DH(\tilde{x}_1 + t\tilde{\phi}(\tilde{x}_1))\,dt \in \mathscr{L}(X, Y_2)$$

is continuous w.r.t. $\tilde{x}_1$ around the origin of $\tilde{X}_1$. As $\Lambda(0) = P_2DH(0)$ is an isomorphism of $\tilde{X}_2$ to Y_2 from the assumptions made in § 1, one has

$$\Lambda(\tilde{x}_1) \in \operatorname{Isom}(\tilde{X}_2, Y_2)$$

for $\tilde{x}_1$ close enough to the origin. Thus, (2.6) becomes

$$P_2H(\tilde{x}_1) = -\Lambda(\tilde{x}_1) \cdot \tilde{\phi}(\tilde{x}_1),$$

or equivalently

$$\tilde{\phi}(\tilde{x}_1) = -\Lambda^{-1}(\tilde{x}_1) \cdot P_2H(\tilde{x}_1).$$

This yields the relations (2.4) and (2.5) since the norms of $\Lambda(\tilde{x}_1)$ and $\Lambda^{-1}(\tilde{x}_1)$ are bounded around the origin. Our last assertion follows from elementary arguments. □

Given an integer $0 \leqq l \leqq m$, it is immediate on (2.1)-(2.3) and the definition of the spaces $\tilde{X}_1$, Y_1 and Y_2 that

$$k_1(l) \geqq 2, \tag{2.7}$$

$$k_2^1(l) \geqq k^1(l) \geqq 2. \tag{2.8}$$

Let us now specify the value of l by setting

$$k = \min\{0 \leqq j \leqq m,\ P_1D^jH(0)|_{(\tilde{X}_1)^j} \neq 0\}. \tag{2.9}$$

Clearly, from (2.7),

$$k \geqq \max(k_1(k), k^1(k)) \geqq 2. \tag{2.10}$$

THEOREM 2.1. *Assume $n \geqq 1$ and k defined by* (2.9) *is finite* (*i.e.* $2 \leqq k \leqq m$). *If*

$$k \leqq k_1(k) + k^1(k) - 2, \tag{2.11}$$

the reduced mapping (1.11) *verifies*

$$D^jf(0) = 0, \qquad 0 \leqq j \leqq k-1, \tag{2.12}$$

$$D^kf(0) = P_1D^kH(0)|_{(\tilde{X}_1)^k} \neq 0. \tag{2.13}$$

Proof. First, observe that

$$k^1(k)=\min(k, k_2^1(k)). \tag{2.14}$$

With (2.10) and since $k \leqq m$ we find $k_1(k) \leqq k \leqq m$. Expanding the mapping P_1H about the origin, we get

$$P_1H(\tilde{x}) = \sum_{j=k_1(k)}^{k} \frac{1}{j!} P_1 D^j H(0) \cdot (\tilde{x})^j + o(\|\tilde{x}\|^k).$$

For $\tilde{x} = \tilde{x}_1 + \tilde{\phi}(\tilde{x}_1)$ we obtain

$$f(\tilde{x}_1) = \sum_{j=k_1(k)}^{k} \frac{1}{j!} P_1 D^j H(0) \cdot (\tilde{x}_1 + \tilde{\phi}(\tilde{x}_1))^j + o(\|\tilde{x}_1 + \tilde{\phi}(\tilde{x}_1)\|^k).$$

As $\tilde{\phi}(0)=0$ and $\tilde{\phi}$ is of class $\mathscr{C}^m$, one has $\tilde{\phi}(\tilde{x}_1) = O(\|\tilde{x}_1\|)$ and the remainder in the above formula can be replaced by $o(\|\tilde{x}_1\|^k)$. Thus

$$f(\tilde{x}_1) = \sum_{j=k_1(k)}^{k} \frac{1}{j!} \sum_{i=0}^{j} \binom{j}{i} P_1 D^j H(0) \cdot ((\tilde{x}_1)^{j-i}, (\tilde{\phi}(\tilde{x}_1))^i) + o(\|\tilde{x}_1\|^k). \tag{2.15}$$

Due to Lemma 2.1 and the definition of $k_2^1(k)$, the derivatives at the origin of the mapping $\tilde{\phi}$ vanish up to order $k_2^1(k)-1$ when $k_2^1(k) < +\infty$ and up to order k when $k_2^1(k) = +\infty$. In any case, (2.14) shows that they vanish up to order $k^1(k)-1$. Hence

$$\|\tilde{\phi}(\tilde{x}_1)\| = O(\|\tilde{x}_1\|^{k^1(k)}).$$

Each term in the sum (2.15) is then of order

$$O(\|\tilde{x}_1\|^{j-i} \|\tilde{\phi}(\tilde{x}_1)\|^i) = O(\|\tilde{x}_1\|^{j-i+(k^1(k)-1)i}).$$

For $1 \leqq i \leqq j$, this term is actually in the remainder of order k. Indeed, $j+(k^1(k)-1)i$ is an increasing function of i so that

$$j+(k^1(k)-1)i \geqq j + k^1(k) - 1 \quad \text{when } 1 \leqq i \leqq j.$$

As j runs over the integers $k_1(k), \cdots, k$

$$j + k^1(k) - 1 \geqq k_1(k) + k^1(k) - 1 > k,$$

from condition (2.11). To sum up, (2.15) simplifies in the form

$$f(\tilde{x}_1) = \sum_{j=k_1(k)}^{k} \frac{1}{j!} P_1 D^j H(0) \cdot (\tilde{x}_1)^j + o(\|\tilde{x}_1\|^k).$$

But, by definition of k and since $\tilde{x}_1 \in \tilde{X}_1$, the expressions $P_1 D^j H(0) \cdot (\tilde{x}_1)^j$ vanish for $k_1(k) \leqq j < k$ and we are left with

$$f(\tilde{x}_1) = \frac{1}{k!} P_1 D^k H(0) \cdot (\tilde{x}_1)^k + o(\|\tilde{x}_1\|^k).$$

Since f is of class $\mathscr{C}^m$, $m \geqq k$, the above relation shows that the first nonzero derivative of f at the origin is of order k with

$$D^k f(0) = P_1 D^k H(0)|_{(\tilde{X}_1)^k},$$

and the proof is complete. □

Note that criterion (2.11) is always satisfied if $k=2$ since $k_1(2) = k^1(2) = 2$ as it follows from (2.7), (2.8) and (2.10). More generally, condition (2.11) is vacuous if

$k=k_1(k)$ or if $k=k^1(k)$ since $k_1(k)\geqq 2$ and $k^1(k)\geqq 2$. The case when $k=k^1(k)<+\infty$ amounts to assuming

$$D^jH(0)|_{(\tilde{X}_1)^j}=0, \qquad 0\leqq j\leqq k-1 \quad \text{(vacuous if } k=2), \tag{2.16}$$

whereas $k=k_1(k)<+\infty$ if and only if

$$P_1D^jH(0)=0, \qquad 0\leqq j\leqq k-1 \quad \text{(vacuous if } k=2). \tag{2.17}$$

A more restrictive version of (2.16) and (2.17) which is also more frequently encountered in the literature (often implicitly) is

$$D^jH(0)=0, \qquad 0\leqq j\leqq k-1. \tag{2.18}$$

Combining Theorems 1.1 and 2.1, we find the following.

THEOREM 2.2. *Let H be a mapping of class $\mathscr{C}^m$, $m\geqq 1$, on a neighbourhood of the origin in $\tilde{X}$ with values in Y such that $H(0)=0$. Assume that $DH(0)$ is a Fredholm operator with index* 1 *and set $\tilde{X}_1=\operatorname{Ker} DH(0)$, $Y_2=\operatorname{Range} DH(0)$. Let $\tilde{X}_2$ and Y_1 be two topological complements of the spaces $\tilde{X}_1$ and Y_2, respectively. Then,* $\dim Y_1=n$ *and* $\dim \tilde{X}_1=n+1$. *Let P_1 denote the projection operator associated with the decomposition $Y=Y_1\oplus Y_2$. Define*

$$k=\begin{cases}1 & \text{if } n=0,\\ \min\{0\leqq j\leqq m, P_1D^jH(0)|_{(\tilde{X}_1)^j}\neq 0\}\geqq 2 & \text{if } n\geqq 1.\end{cases} \tag{2.19}$$

When $n\geqq 1$, and in this case only, set

$$k_1(k)=\min\{0\leqq j\leqq k, P_1D^jH(0)\neq 0\}\geqq 2, \tag{2.20}$$

$$k^1(k)=\min\{0\leqq j\leqq k, D^jH(0)|_{(\tilde{X}_1)^j}\neq 0\}\geqq 2 \tag{2.21}$$

and assume

(i) $k<+\infty$,
(ii) $k\leqq k_1(k)+k^1(k)-2$ (*vacuous for* $k=2$),
(iii) *the mapping*

$$\tilde{x}_1\in\tilde{X}_1\to P_1D^kH(0)\cdot(\tilde{x}_1)^k\in Y_1, \tag{2.22}$$

verifies the condition ($\mathbb{R}$-N.D.).[2]

Then, the zero set of the mapping (2.22) *consists of* $0^n\leqq \nu\leqq k^n$[3] *lines through the origin in the space $\tilde{X}_1$ and the zero set of the mapping H around the origin of $\tilde{X}$ consists of exactly ν curves of class $\mathscr{C}^m$ away from the origin and of class $\mathscr{C}^{m-k+1}$ at the origin where each of them is tangent to a different one of the lines in the zero set of the mapping* (2.22).

Remark 2.1. It is easily seen that k, $k_1(k)$ and $k^1(k)$ are independent of the choice of the spaces $\tilde{X}_2$ and Y_1. The same is true as concerns the fact that the mapping (2.22) verifies the condition ($\mathbb{R}$-N.D.): Indeed, let Y_1' be a second space such that $Y=Y_1'\oplus Y_2$ and call P_1' the associated projection onto Y_1'. As $P_1'P_2=0$ one has $P_1'=P_1'P_1$. Also, $\operatorname{Ker} P_1'=Y_2$ and hence $P'_{1|Y_1}\in\operatorname{Isom}(Y_1, Y_1')$. The assertion is now immediate from this observation. As a result, Theorem 2.2 is independent of the choice of the Lyapunov-Schmidt reduction. Actually, from a result of Beyn [1] showing the equivalence (in some precise sense) of any two Lyapunov-Schmidt reductions, it can be shown that the order of the first nonzero derivative of the reduced mapping (1.11) at the origin

[2] Recall that the condition ($\mathbb{R}$-N.D.) was defined in Remark 1.2.

[3] With $0^0=1$ as an understanding.

and the fact that it verifies the condition ($\mathbb{R}$-N.D.) are *always* independent of the choice of the Lyapunov–Schmidt reduction.

3. Applications to one-parameter nonlinear problems: the case $p=1$. In this section, we come back to the situation described in § 1: given two real Banach spaces X and Y and a mapping $G(=G(\mu, x))$ defined and of class $\mathscr{C}^m$, $m \geqq 1$, on a neighbourhood of the origin in $\mathbb{R}\times X$ with values in Y such that $G(0)=0$, we consider the problem of finding the local zero set of G near the origin. As this paper is intended to emphasize the role of the change of parameter in this problem, we shall especially bring prominence to situations in which a direct approach (i.e. without changing the parameter μ) fails or does not provide all the desirable information.

As in § 1, we shall assume that $D_xG(0)$ is a Fredholm operator with index 0 and set

$$X_1 = \text{Ker } D_xG(0), \tag{3.1}$$

$$Z_2 = \text{Range } D_xG(0). \tag{3.2}$$

Given two topological complements X_2 and Z_1 of X_1 and Z_2, respectively, we shall denote by Q_1 and Q_2 the (continuous) projection operators from Y onto Z_1 and Z_2 and, whenever necessary, write the generic element $x \in X$ in the form $x = x_1 + x_2$, $x_1 \in X_1$, $x_2 \in X_2$. Since $D_xG(0)$ has index 0

$$N = \dim X_1 = \dim Z_1 \geqq 0. \tag{3.3}$$

Setting $\tilde{X} = \mathbb{R}\times X$ with generic element $\tilde{x} = (\mu, x)$ we shall see how the general results of § 2 apply when $H = G$. First, one has

$$DG(0)\cdot(\mu, x) = \mu D_\mu G(0) + D_xG(0)\cdot x, \tag{3.4}$$

so the null-space and the range of $DG(0)$ depend on the alternative $D_\mu G(0) \in Z_2$ or $D_\mu G(0) \notin Z_2$.

First case. $D_\mu G(0) \notin Z_2$. In the notation of § 2 and from (3.4)

$$\tilde{X}_1 = \text{Ker } DG(0) = \{0\}\times X_1 \simeq X_1, \tag{3.5}$$

$$Y_2 = \text{Range } DG(0) = \mathbb{R} D_\mu G(0) \oplus Z_2. \tag{3.6}$$

From the above and our assumptions on G, the space Y_2 is closed. Besides, due to (3.3), $\dim \tilde{X}_1 = \text{codim } Y_2 + 1 = N$. Thus, the codimension n of Y_2 is

$$n = N - 1. \tag{3.7}$$

When $N = 1$ (i.e. $n = 0$), Theorem 2.2 applies without any further assumption and the zero set of G is made of exactly one curve of class $\mathscr{C}^m$. This curve is tangent to the one-dimensional space $\tilde{X}_1 = \{0\}\times X_1$ at the origin ("vertical" tangent). Note that no information is provided as concerns its location in $\mathbb{R}\times X$ relative to the hyperplane $\mu = 0$. This will be complemented in § 4 (case "$p = \kappa$").

When $N \geqq 2$ (i.e. $n \geqq 1$), k is defined by (cf. § 2 and (3.5)–(3.6))

$$k = \min\{0 \leqq j \leqq m, P_1 D_x^j G(0)|_{(X_1)^j}\} \neq 0. \tag{3.8}$$

Assuming $k < +\infty$ (i.e. $2 \leqq k \leqq m$) one has

$$k^1(k) = \min\{0 \leqq j \leqq k, D_x^j G(0)|_{(X_1)^j} \neq 0\}. \tag{3.9}$$

In contrast, $k_1(k)$ involves derivatives of G in all directions

$$k_1(k) = \min\{0 \leqq j \leqq k, P_1 D^j G(0) \neq 0\}. \tag{3.10}$$

Theorem 2.2 will apply if $k<+\infty$ and

$$k \leqq k_1(k)+k^1(k)-2 \quad \text{(vacuous for } k=2) \tag{3.11}$$

and the mapping

$$x_1 \in X_1 \rightarrow P_1 D_x^k G(0) \cdot (x_1)^k \in Y_1, \tag{3.12}$$

verifies the condition ($\mathbb{R}$-N.D.). If so, the largest number of curves in the zero set of G is k^{N-1} and the curves are of class $\mathscr{C}^{m-k+1}$ at the origin and $\mathscr{C}^m$ away from it. Here, the two particular cases when $k=k^1(k)$ or $k=k_1(k)$ (ensuring that (3.11) holds) are, respectively,

$$D_x^j G(0)|_{(X_1)^j}=0, \qquad 0 \leqq j \leqq k-1 \quad \text{(vacuous for } k=2), \tag{3.13}$$

$$P_1 DG(0)=0, \qquad 0 \leqq j \leqq k-1 \quad \text{(vacuous for } k=2). \tag{3.14}$$

In the next section, we shall consider the quantity

$$\kappa=\min\{0 \leqq j \leqq m, \, Q_1 D_x^j G(0)|_{(X_1)^j} \neq 0\} \geqq 2. \tag{3.15}$$

When $N=1$ (i.e., $n=0$), one has $k=1$ in Theorem 2.2 and, in particular, $k<\kappa$. On the contrary, when $N \geqq 2$ (i.e., $n \geqq 1$) the inequality $k \geqq \kappa$ holds. Indeed, since Ker $P_1 = Y_2 \supset Z_2$, one has $P_1 Q_2=0$ and hence $P_1=P_1 Q_1$ so that $k \geqq \kappa$ from the definitions. Anticipating some of the results of the next section, it is not without interest to notice that the equality $k=\kappa$ holds in most cases (when $N \geqq 2$). Indeed, $k=\kappa$ unless $\kappa<+\infty$ and the κ-linear mapping $Q_1 D_x^\kappa G(0)|_{(X_1)^\kappa}$ has one-dimensional range $\mathbb{R} Q_1 D_\mu G(0)$. In the applications, a consequence of this observation is that the necessary assumptions in the study of the same problem, after we change the parameter μ in the next section, will not differ too much from those we make here. The main advantage in considering the approach of § 4 will be that it provides the information on the location of the branches relatively to the hyperplane $\mu=0$ in $\mathbb{R} \times X$, which is not available here.

Remark 3.1. For $N \geqq 2$, a simple case when $k=\kappa$ is when κ is odd and

$$Q_1 D_x^\kappa G(0) \cdot (x_1)^\kappa \neq 0 \quad \text{for every } x_1 \in X_1-\{0\}. \tag{3.16}$$

Indeed, the unit sphere in the space X_1 is connected when $N \geqq 2$ and the polynomial mapping $Q_1 D_x^\kappa G(0) \cdot (x_1)^\kappa$ is odd if κ is odd. Assuming that it takes its values in a one-dimensional subspace of Z_1, we deduce that there is a point on the unit sphere in X_1 at which it vanishes, contradicting (3.16). Condition (3.16) will appear again in § 4 regardless of the parity of κ for other reasons. Note that it is not very restrictive since the spaces X_1 and Y_1 have the same dimension (generically, a homogeneous polynomial mapping between two spaces with the same dimension is known not to vanish away from the origin according to a classical result in elementary algebraic geometry; see e.g. [6]).

Second case. $D_\mu G(0) \in Z_2$. As $D_\mu G(0) \in Z_2$, there exists $\xi \in X$ such that

$$D_x G(0) \cdot \xi = D_\mu G(0). \tag{3.17}$$

From (3.4) one has

$$\tilde{X}_1^\xi \equiv \text{Ker } DG(0)=\{(\mu, x_1-\mu\xi), (\mu, x_1) \in \mathbb{R} \times X_1\} \simeq \mathbb{R} \times X_1, \tag{3.18}$$

$$Y_2=\text{Range } DG(0)=Z_2. \tag{3.19}$$

Thus, Range $DG(0)$ is closed and dim $\tilde{X}_1^\xi = \text{codim } Y_2+1$. The codimension n of Y_2 is

$$n=N. \tag{3.20}$$

When $N=0$ (i.e. $n=0$), Theorem 2.2 applies and the zero set of G is made of exactly one curve of class $\mathscr{C}^m$. This curve is tangent to the one-dimensional space $\tilde{X}_1^\xi \subset \mathbb{R} \times X$ ("oblique" tangent). This is the case when the origin is referred to as a "regular point."

When $N \geqq 1$ (i.e. $n \geqq 1$), a complement of the space Y_2 (3.19) is nothing but a complement of Z_2. In other words, $Y_1 = Z_1$ and k is defined by

$$k = \min\{0 \leqq j \leqq m, \; Q_1 D^j G(0)|_{(\tilde{X}_1^\xi)^j} \neq 0\}. \tag{3.21}$$

Assuming $k < +\infty$ (i.e. $2 \leqq k \leqq m$) one has

$$k^1(k) = \min\{0 \leqq j \leqq m, \; D^j G(0)|_{(\tilde{X}_1^\xi)^j} \neq 0\}, \tag{3.22}$$

$$k_1(k) = \min\{0 \leqq j \leqq m, \; Q_1 D^j G(0) \neq 0\}. \tag{3.23}$$

Theorem 2.2 applies if

$$k \leqq k_1(k) + k^1(k) - 2 \quad \text{(vacuous for } k=2) \tag{3.24}$$

and the mapping

$$(\mu, x) \in \tilde{X}_1^\xi \to Q_1 D^k G(0) \cdot (\mu, x)^k \in Z_1 \quad (= Y_1) \tag{3.25}$$

verifies the condition ($\mathbb{R}$-N.D.). If so, the largest number of curves in the zero set of G is k^N and the curves are of class $\mathscr{C}^{m-k+1}$ at the origin and $\mathscr{C}^m$ away from it. The cases $k = k_1(k)$ or $k = k^1(k)$ (ensuring that (3.24) holds) are, respectively

$$Q_1 D^j G(0) = 0, \qquad 0 \leqq j \leqq k-1 \quad \text{(vacuous for } k=2), \tag{3.26}$$

$$D^j G(0)|_{(\tilde{X}_1^\xi)^j} = 0, \qquad 0 \leqq j \leqq k-1 \quad \text{(vacuous for } k=2). \tag{3.27}$$

As in the previous case when $D_\mu G(0) \notin Z_2$ and assuming $N \geqq 1$, define

$$\kappa = \min\{0 \leqq j \leqq m, \; Q_1 D_x^j G(0)|_{(X_1)^j} \neq 0\} \geqq 2. \tag{3.28}$$

As $\{0\} \times X_1 \subset \tilde{X}_1^\xi$, it follows from (3.21) that

$$k \leqq \kappa. \tag{3.29}$$

Recall when $D_\mu G(0) \notin Z_2$ and $N \geqq 2$ that the opposite inequality $k \geqq \kappa$ holds and we observed that the equality is true in most cases. The latter conclusion is no longer appropriate here. Nevertheless, the condition $k = \kappa$ is *necessary* for the mapping (3.25) to verify the condition ($\mathbb{R}$-N.D.) when $N \geqq 2$. Indeed, if $k < \kappa$, the derivative $Q_1 D_x^k G(0)|_{(X_1)^k}$ vanishes and the N-dimensional space $\{0\} \times X_1$ is readily seen to be in the zero set of the mapping (3.25). Thus, it does not consist of a finite number of lines when $N \geqq 2$ and the mapping (3.25) cannot verify the condition ($\mathbb{R}$-N.D.). As it will be seen on the example of problems of bifurcation from the trivial branch, this may be a serious reason for the failure of the approach developed in this section. Thus, when $D_\mu G(0) \in Z_2$, performing a change of the parameter μ may be a necessity, not merely for locating the curves but for proving their existence (or nonexistence) as well. When $N=1$, k need not equal κ as we shall now see.

Assume then $N=1$. Denoting by $\langle \cdot, \cdot \rangle$ the pairing between the space Y and its dual Y', one has

$$Q_1 y = \langle y^*, y \rangle y_1^0, \tag{3.30}$$

for every $y \in Y$ where y_1^0 is a given nonzero element of the one-dimensional space $Z_1 = Y_1$ and $y^* \in Y'$ is characterized by

$$(3.31) \qquad \langle y^*, y_1^0 \rangle = 1, \qquad \langle y^*, y \rangle = 0 \quad \text{for every } y \in Z_2.$$

As the space X_1 is one-dimensional too, we may set $X_1 = \mathbb{R}x_1^0$, where x_1^0 is a given nonzero element. In the discussion below, we shall limit ourselves to considering the case $D_\mu G(0) = 0$, to which the general situation $D_\mu G(0) \in Z_2$ reduces by changing $G(\mu, x)$ into $\hat{G}(\mu, x) = G(\mu, x - \mu\xi)$. If $D_\mu G(0) = 0$, we may take $\xi = 0$ so $\tilde{X}_1^0 = \mathbb{R} \times X_1$ and the mapping (3.25) is

$$(3.32) \qquad (\mu, x_1) \in \mathbb{R} \times X_1 \rightarrow Q_1 D^k G(0) \cdot (\mu, x_1)^k \in Z_1.$$

From (3.30)-(3.31), the above mapping verifies the condition ($\mathbb{R}$-N.D.) if and only if the mapping

$$(3.33) \qquad (\mu, t) \in \mathbb{R}^2 \rightarrow \langle y^*, D^k G(0) \cdot (\mu, t x_1^0)^k \rangle \in \mathbb{R},$$

does the same. It can be shown (cf. [15] for details) that a sufficient condition for this is that a certain determinant does not vanish. For $k = 2$, this determinant is that of the quadratic form (3.33) (Morse condition) and the condition is also necessary (this remains true for $k = 3$ but not for $k \geqq 4$). When $k = 2$ the result is well known but it does not seem to be widely reported for a general k.

We shall now examine in some detail the example of problems of bifurcation from the trivial branch. With $X = Y$ the mapping G is of the form

$$(3.34) \qquad G(\mu, x) = x - (\lambda_0 + \mu) L x + \Gamma(\mu, x),$$

where $L \in \mathscr{L}(X)$ is a compact operator and $\lambda_0 \in \mathbb{R} - \{0\}$ a given real number. The mapping Γ is of class $\mathscr{C}^m$, $m \geqq 1$ (with values in X) and $\Gamma(\mu, 0) = 0$ for μ around $0 \in \mathbb{R}$. In particular

$$(3.35) \qquad D_\mu^j \Gamma(0) = 0, \qquad 0 \leqq j \leqq m.$$

We shall also assume

$$(3.36) \qquad D_\mu D_x \Gamma(0) = 0,$$

$$(3.37) \qquad D_x \Gamma(0) = 0.$$

The case when λ_0 is not a characteristic value of L (i.e. $N = 0$) is obvious and well known. When λ_0 is a characteristic value of L (i.e. $N \geqq 1$), the integer k is defined by

$$(3.38) \qquad k = \min \{0 \leqq j \leqq m, \; Q_1 D^j G(0)|_{(\mathbb{R} \times X_1)^j} \neq 0\}.$$

PROPOSITION 3.1. *Assume that λ_0 is a characteristic value of L. Then, a necessary condition for our conclusions to hold is*

$$(3.39) \qquad X = \operatorname{Ker}(I - \lambda_0 L) \oplus \operatorname{Range}(I - \lambda_0 L).$$

If so, $k = 2$ and we can take $Z_1 (= Y_1) = X_1 (= \operatorname{Ker}(I - \lambda_0 L))$ without loss of generality. The criterion (3.24) *is then vacuous and it suffices to assume that the mapping*

$$(3.40) \quad (\mu, x_1) \in \mathbb{R} \times \operatorname{Ker}(I - \lambda_0 L) \rightarrow -\frac{2}{\lambda_0} \mu x_1 + Q_1 D_x^2 \Gamma(0) \cdot (x_1)^2 \in \operatorname{Ker}(I - \lambda_0 L),$$

verifies the condition ($\mathbb{R}$-N.D.).

Proof. Due to (3.36) one has $D_\mu D_x G(0) = -L$. Next, by definition of $X_1 = \operatorname{Ker}(I - \lambda_0 L)$,

$$Q_1 D_\mu D_x G(0) \cdot x_1 = -\frac{1}{\lambda_0} Q_1 x_1$$

for every $x_1 \in X_1$. Hence,

$$Q_1 D_\mu D_x G(0)|_{X_1} \neq 0 \Leftrightarrow X_1 \not\subset Z_2 (= Y_2). \tag{3.41}$$

Let us first assume $X_1 \subset Z_2$. From (3.35), (3.36) and (3.41) the value of k is $k = 2$ if and only if $Q_1 D_x^2 \Gamma(0)|_{(X_1)^2} \neq 0$. The criterion (3.24) is satisfied and the mapping (3.25) reduces to

$$(\mu, x_1) \in \mathbb{R} \times X_1 \to Q_1 D_x^2 \Gamma(0) \cdot (x_1)^2 \in Z_1 (= Y_1).$$

But the above mapping does not verify the condition ($\mathbb{R}$-N.D.) because of the elements of the form $(\mu, 0)$ in its zero set at which its derivative vanishes. Therefore, when $X_1 \subset Z_2$, the analysis of this section may be available with $k \geqq 3$ only. In this case, however, the criterion (3.24) is not satisfied. To see this, it suffices to prove that $k^1(k) = k_1(k) = 2$, that is to say that the derivatives $D^2 G(0)|_{(\mathbb{R} \times X_1)^2}$ and $Q_1 D^2 G(0)$ do not vanish. This will follow from the relations

$$D_\mu D_x G(0)|_{X_1} \neq 0, \qquad Q_1 D_\mu D_x G(0) \neq 0.$$

Indeed, assume by contradiction that $D_\mu D_x G(0)|_{X_1} = 0$, namely $Lx_1 = 0$ for every $x_1 \in \operatorname{Ker}(I - \lambda_0 L)$. Thus, $\operatorname{Ker}(I - \lambda_0 L) = \{0\}$ and we reach a contradiction with the fact that λ_0 is a characteristic value of L. Now, assume by contradiction that $Q_1 D_\mu D_x G(0) = 0$. This means $Q_1 L = 0$ and, as $Q_1(I - \lambda_0 L) = 0$ by definition of Q_1, we find $Q_1 = 0$, namely $Z_2 = X$, again a contradiction.

To sum up, we must assume $X_1 \not\subset Z_2$. From (3.41) it follows that $k = 2$ and the criterion (3.24) is satisfied. The mapping (3.25) becomes

$$(\mu, x_1) \in \mathbb{R} \times X_1 \to -\frac{2}{\lambda_0} \mu Q_1 x_1 + Q_1 D_x^2 \Gamma(0) \cdot (x_1)^2 \in Z_1.$$

Its zero set contains the pairs $(\mu, 0)$, $\mu \in \mathbb{R}$ and its derivative at such a point with $\mu \neq 0$ is the linear mapping $-(2/\lambda_0)\mu Q_1|_{X_1} \in \mathcal{L}(X_1, Z_1)$. The nondegeneracy condition requires it to be onto, or, equivalently, one-to-one since $\dim X_1 = \dim Z_1 = N$. This means $Q_1 x_1 \neq 0$ for every $x_1 \in X_1$. In other words, $X_1 \cap Z_2 = \{0\}$ which amounts to saying that $X = X_1 \oplus Z_2$. As $Z_2 = Y_2$ and we observed in § 2 that Theorem 2.2 was independent of the choice of the complement $Y_1 (= Z_1)$ of Y_2, we may choose $Z_1 = X_1$ and the proof is complete. □

When $N = 1$, condition (3.39) is often referred to as "Crandall and Rabinowitz nondegeneracy condition." For $N \geqq 2$ it also appears in McLeod and Sattinger [11]. Assuming that (3.39) holds, it is easily checked that the assumptions we make and the conclusions we draw by using Theorem 2.2 are indeed the same as in Crandall and Rabinowitz [4] when $N = 1$ (including the regularity of the curves at the origin). When $N \geqq 2$, our conclusions are as in McLeod and Sattinger [11] under somewhat weaker assumptions. Besides ours, they also require that

$$Q_1 D_x^2 \Gamma(0) \cdot (x_1)^2 \neq 0 \quad \text{for every } x_1 \in X_1 - \{0\} \tag{3.42}$$

(replacing λL by $L(\lambda)$ as in [11] does not affect this remark). Although very little restrictive in the applications, condition (3.42) merely ensures that bifurcation occurs transcritically, as in the case $N = 1$ for a quadratic nonlinearity. The largest number of bifurcated branches is 2^N-1 and the curves are of class $\mathscr{C}^{m-1}$ at the origin, $\mathscr{C}^m$ away from it. Apparently, the result of regularity at the origin is new when $N\geqq 2$. Bifurcation is ensured since $k=2$ is even (cf. Remark 1.3).

4. Applications to one parameter nonlinear problems: the case $p\geqq 2$. Let the mapping G satisfy the general hypotheses of §§ 1 and 3. Given an integer $p\geqq 2$ and with $\sigma=\pm 1$, let $\tilde{X}$ denote the space $\mathbb{R}\times X$ with generic point (η, x). We intend to use Theorem 2.2 with the mapping $H=H_\sigma$, where

$$H_\sigma(\eta,x)=G(\sigma\eta^p,x). \tag{4.1}$$

As $p\geqq 2$ and by differentiating

$$DH_\sigma(0)\cdot(\eta,x)=D_xG(0)\cdot x \tag{4.2}$$

so that the spaces $\tilde{X}_1=\operatorname{Ker} DH_\sigma(0)$ and $Y_2=\operatorname{Range} DH_\sigma(0)$ are independent of σ and p. More precisely

$$\tilde{X}_1=\operatorname{Ker} DH_\sigma(0)=\mathbb{R}\times\operatorname{Ker} D_xG(0)=\mathbb{R}\times X_1, \tag{4.3}$$

$$Y_2=\operatorname{Range} DH_\sigma(0)=\operatorname{Range} D_xG(0)=Z_2. \tag{4.4}$$

As $\dim X_1=\operatorname{codim} Z_2=N$, the operator $DH_\sigma(0)$ is a Fredholm operator with index 1. Here, the codimension n of $Y_2(=Z_2)$ is

$$n=N. \tag{4.5}$$

A complement Y_1 of Y_2 in Y is of course any complement Z_1 of Z_2 with associated projection operator $P_1=Q_1$. When $N=0$ (i.e. $n=0$) Theorem 2.2 applies. As in the previous section, this situation is essentially obvious and no further information is provided here. We shall henceforth assume $N\geqq 1$. When finite, the value k defined by (2.9) with $H=H_\sigma$ is the order of the first nonzero derivative of the mapping $Q_1H_\sigma|_{\mathbb{R}\times X_1}$ at the origin (that k does not depend on σ will be shown in Proposition 4.1 below). We shall characterize it in terms of G: for every integer $0\leqq l\leqq m$, define the set

$$I_1^1(l)=\{(i,j)\in\mathbb{N}\times\mathbb{N},\ 0\leqq i\leqq j\leqq l,\ Q_1D_\mu^{j-i}D_x^iG(0)|_{(X_1)^i}\neq 0\}. \tag{4.6}$$

Now, for $0\leqq l\leqq m$ and $p\geqq 2$ we introduce the mapping

$$\chi_1^1(l,p)=\min\{pj+(1-p)i,\ (i,j)\in I_1^1(l)\}. \tag{4.7}$$

Clearly,

$$\chi_1^1(l,p)=+\infty\Leftrightarrow I_1^1(l)=\varnothing, \tag{4.8}$$

while there is at least one pair $(i,j)\in I_1^1(l)$ such that $pj+(1-p)i=\chi_1^1(l,p)$ when $I_1^1(l)\neq\varnothing$ since $I_1^1(l)$ is finite in any case.

PROPOSITION 4.1. *Let $0\leqq l\leqq m$ be given. If $\chi_1^1(l,p)>l$, the derivatives of the mapping $Q_1H_\sigma|_{\mathbb{R}\times X_1}$ vanish up to order l at the origin. In contrast, if $\chi_1^1(l,p)\leqq l$, one has*

$$k=\chi_1^1(l,p) \tag{4.9}$$

(*so that k is independent of σ*) *and*

(4.10)$_\sigma$

$$Q_1D^kH_\sigma(0)\cdot(\eta,x_1)^k=\sum_{\substack{0\leqq i\leqq j\leqq l,\\ pj+(1-p)i=k}}\frac{k!}{(j-i)!\,i!}(\sigma\eta^p)^{j-i}Q_1D_\mu^{j-i}D_x^iG(0)\cdot(x_1)^i\in Z_1.$$

Proof. For every pair (h,i) of integers and provided that the derivative of order $h+i$ if H_σ is defined, one has

$$Q_1D_\eta^hD_{x_1}^iH_\sigma(0)=Q_1D_{x_1}^i[D_\eta^hH_\sigma(0,x_1)]|_{x_1=0}.$$

As $H_\sigma(\eta, x_1)$ depends only on η^p (cf. (4.1)) it is easily checked that $D^h_\eta H_\sigma(0, x_1)=0$ unless h is a multiple of p, namely $h=p(j-i)$ for some index $j \geqq i$. Besides, a simple calculation based on Taylor's formula yields

$$D^{p(j-i)}_\eta H_\sigma(0, x_1)=\frac{[p(j-i)]!}{(j-i)!} D^{j-i}_\mu G(0, x_1).$$

Hence,

$$Q_1 D^{p(j-i)}_\eta D^i_{x_1} H_\sigma(0)=\frac{[p(j-i)]!}{(j-i)!} Q_1 D^{j-i}_\mu D^i_{x_1} G(0), \tag{4.11}$$

while all the expressions $Q_1 D^h_\eta D^i_{x_1} H_\sigma(0)$ vanish if h is not of the form $p(j-i)$. Any nonzero derivative of the mapping $Q_1 H_\sigma|_{\mathbb{R}\times X_1}$ at the origin is then of order $p(j-i)+i$ for some pairs (i, j) with $i \leqq j$. This order is less than or equal to a given integer $0 \leqq l \leqq m$ if and only if $p(j-i)+i=pj+(1-p)i \leqq l$ and the derivative involves indices $(i, j) \in I^1_1(l)$ only. Our assertions follow at once from this observation. □

For $0 \leqq l \leqq m$ again, let us now introduce the sets

$$I_1(l)=\{0 \leqq i \leqq j \leqq l,\ Q_1 D^{j-i}_\mu D^i_x G(0) \neq 0\}, \tag{4.12}$$

$$I^1(l)=\{0 \leqq i \leqq j \leqq l,\ D^{j-i}_\mu D^i_x G(0)|_{(X_1)^i} \neq 0\}. \tag{4.13}$$

To these sets, we associate the mappings

$$\chi_1(l, p)=\min\{pj+(1-p)i, (i, j) \in I_1(l)\}, \tag{4.14}$$

$$\chi^1(l, p)=\min\{pj+(1-p)i, (i, j) \in I^1(l)\}. \tag{4.15}$$

Arguing as in Proposition 4.1 we see that the derivatives of the mapping $Q_1 H_\sigma$ (resp., $H_\sigma|_{\mathbb{R}\times X_1}$) vanish up to order l at the origin if $\chi_1(l, p)>l$ (resp., $\chi^1(l, p)>l$) and $\chi_1(l, p)$ (resp., $\chi^1(l, p)$) is the order of the first nonzero derivative of the mapping $Q_1 H_\sigma$ (resp., $H_\sigma|_{\mathbb{R}\times X_1}$) at the origin if $\chi_1(l, p) \leqq l$ (resp., $\chi^1(l, p) \leqq l$). In the notation of § 2 (cf. (2.1) and (2.2)) this means that

$$\chi_1(l, p) \leqq l \Rightarrow \chi_1(l, p)=k_1(l), \qquad \chi^1(l, p) \leqq l \Rightarrow \chi^1(l, p)=k^1(l), \tag{4.16}$$

where $k_1(\cdot)$ and $k^1(\cdot)$ are defined through the mapping $H=H_\sigma$. Since $k_1(l)=k_1(l')$ (resp., $k^1(l)=k^1(l')$) when both sides are finite as it is obvious from the definitions, we obtain

$$\begin{gathered}\{\chi_1(l, p) \leqq l \text{ and } \chi_1(l', p) \leqq l'\} \Rightarrow \chi_1(l, p)=\chi_1(l', p), \\ \{\chi^1(l, p) \leqq l \text{ and } \chi^1(l', p) \leqq l'\} \Rightarrow \chi^1(l, p)=\chi^1(l', p).\end{gathered} \tag{4.17}$$

Finally, as $I^1_1(l)$ is contained in both sets $I_1(l)$ and $I^1(l)$ for every $0 \leqq l \leqq m$, it is clear that

$$\chi^1_1(l, p) \geqq \max(\chi_1(l, p), \chi^1(l, p)). \tag{4.18}$$

When we combine the above observations, and according to the definitions, it is easily seen that

PROPOSITION 4.2. *Let $0 \leqq l \leqq m$ be such that $\chi^1_1(l, p) \leqq l$. Then*

$$k_1(\chi^1_1(l, p))=\chi_1(l, p), \qquad k^1(\chi^1_1(l, p))=\chi^1(l, p). \qquad \square$$

We shall now see that Theorem 2.2 cannot be used with the mapping H_σ unless k and p have specific values. The quantity

$$\kappa=\min\{0 \leqq j \leqq m,\ Q_1 D^j_x G(0)|_{(X_1)^j} \neq 0\} \geqq 2, \tag{4.19}$$

we have already encountered in § 3, will play a key role.

THEOREM 4.1. *Assume that Theorem* 2.2 *applies with* $H = H_\sigma$. *Then*

$$k = \chi_1^1(\kappa, p) = \kappa < +\infty \tag{4.20}$$

and the mapping

(4.21)$_\sigma$

$$Q_1 D^\kappa H_\sigma(0) \cdot (\eta, x_1)^\kappa = \sum_{\substack{0 \leqq i \leqq j \leqq \kappa, \\ pj+(1-p)i=\kappa}} \frac{1}{(j-i)!\,i!} (\sigma\eta^p)^{j-i} Q_1 D_\mu^{j-i} D_x^i G(0) \cdot (x_1)^i \in Z_1$$

verifies the condition ($\mathbb{R}$-N.D.). *In addition*

$$\kappa \leqq \chi_1(\kappa, p) + \chi^1(\kappa, p) - 2 \tag{4.22}$$

and the integer p is characterized by

$$p = \max\left\{\frac{\kappa - i}{j - i}, (i, j) \in I_1^1(\kappa), j < \kappa\right\} \tag{4.23}$$

(*in particular, p is independent of* σ).

Proof. Since k is finite by hypothesis, it coincides with the order of the first nonzero derivative of the mapping $Q_1 H_\sigma|_{\mathbb{R}\times X_1}$ at the origin. From the first part of Proposition 4.1 it follows that there is an integer $0 \leqq l \leqq m$ such that $\chi_1^1(l, p) \leqq l$ and hence $k = \chi_1^1(l, p)$ from the second part of Proposition 4.1. By hypothesis, the mapping (4.10)$_\sigma$ verifies the condition ($\mathbb{R}$-N.D.). If the pair (k, k) is not in the set $I_1^1(l)$, η^p is in factor in (4.10)$_\sigma$ since $pj + (1-p)j = j \neq k$ for each pair $(j, j) \in I_1^1(l)$. But this is impossible. Indeed, as $p \geqq 2$, the mapping (4.10)$_\sigma$ and its derivative vanish on the N-dimensional space $\{0\} \times X_1$, contradicting the condition ($\mathbb{R}$-N.D.) since $N \geqq 1$. Thus

$$Q_1 D_x^k G(0)|_{(X_1)^k} \neq 0. \tag{4.24}$$

On the contrary

$$Q_1 D_x^j G(0)|_{(X_1)^j} = 0, \qquad 0 \leqq j \leqq k-1, \tag{4.25}$$

for there is no pair $(j, j) \in I_1^1(l)$ with $j < k$ since this would contradict the relation $k = \chi_1^1(l, p)$. The relations (4.24) and (4.25) characterize k as being κ (cf. (4.19)). In particular, $\kappa < +\infty$; hence the pair (κ, κ) is in the set $I_1^1(\kappa)$. Thus, $\chi_1^1(\kappa, p) \leqq p\kappa + (1-p)\kappa = \kappa$ and Proposition 4.1 show that $k = \chi_1^1(\kappa, p)$, which proves (4.20). Besides, the mapping (4.10)$_\sigma$ with $k = \kappa$ and $l = \kappa$ is nothing but (4.21)$_\sigma$ (and verifies the condition ($\mathbb{R}$-N.D.)).

Next, from (4.20), (4.16) and (4.18),we find $k_1(k) = \chi_1(\kappa, p)$, $k^1(k) = \chi^1(\kappa, p)$ and the criterion $k \leqq k_1(k) + k^1(k) - 2$ reads as (4.22).

Let us finally show that p is characterized by (4.23): from (4.20) and for every pair $(i, j) \in I_1^1(\kappa)$, one has $pj + (1-p)i \geqq \chi_1^1(\kappa, p) = \kappa$ or, equivalently, $p \geqq (\kappa - i)/(j-i)$. It suffices to prove that there is a pair $(i, j) \in I_1^1(\kappa)$ with $j < \kappa$ for which $pj + (1-p)i = \kappa$. If there is none, the mapping (4.21)$_\sigma$ reduces to

$$(\eta, x_1) \in \mathbb{R} \times X_1 \to Q_1 D_x^\kappa G(0) \cdot (x_1)^\kappa \in Z_1 \tag{4.26}$$

because there is no pair $(i, j) \in I_1^1(\kappa)$ with $j < \kappa$ such that $pj + (1-p)i = \kappa$ by hypothesis and there is no such pair with $j = \kappa > i$ either since $p \geqq 2$ and hence $p\kappa + (1-p)i > \kappa$.

But the mapping (4.26) as well as its derivative vanishes on the line $\mathbb{R} \times \{0\}$ since $\kappa \geqq 2$, so that it does not verify the condition ($\mathbb{R}$-N.D.). □

Remark 4.1. It is easy to check that the value of p given by (4.23) is exactly as provided by using Newton's diagram (see e.g. Sattinger [18]).

The proof of the converse of Theorem 4.1, which is the useful version in practical applications, is based on the same arguments and is omitted. The result is stated in Theorem 4.2 below.

THEOREM 4.2. *Assume* $\kappa<+\infty$ (*i.e.* $2\leqq\kappa\leqq m$) *and suppose that formula* (4.23) *defines p as an integer. Then,*

$$2\leqq p\leqq\kappa \tag{4.27}$$

and the order of the first nonzero derivative of the mapping $Q_1H_\sigma|_{\mathbb{R}\times X_1}$ *at the origin is*

$$k=\chi_1^1(\kappa,p)=\kappa. \tag{4.28}$$

Also, the expression of this derivative is given by $(4.21)_\sigma$ *and*

$$k_1(k)=\chi_1(\kappa,p),\qquad k^1(k)=\chi^1(\kappa,p), \tag{4.29}$$

so that the criterion $k\leqq k_1(k)+k^1(k)-2$ *takes the form*

$$\kappa\leqq\chi_1(\kappa,p)+\chi^1(\kappa,p)-2. \tag{4.30}$$

Finally, if (4.30) *holds and the mapping* $(4.21)_\sigma$ *verifies the condition* ($\mathbb{R}$-N.D.), *Theorem* 2.2 *applies with* $H=H_\sigma$.

Remark 4.2. Observe that κ and the sets $I_1^1(\kappa)$, $I^1(\kappa)$ and $I_1(\kappa)$ are independent of the choice of the space $Z_1(=Y_1)$ and of the complement $\tilde{X}_2$ of $\tilde{X}_1$ $(=\mathbb{R}\times X_1)$. Thus, Theorem 4.2 is independent of the choice of the Lyapunov-Schmidt reduction (in particular, p is independent of it).

According to Proposition 4.2 one has $\chi_1(\kappa,p)\geqq2$, $\chi^1(\kappa,p)\geqq2$. Hence, the criterion (4.30) will be satisfied in particular when $\kappa=\chi_1(\kappa,p)$ or $\kappa=\chi^1(\kappa,p)$, for instance, if the set $I_1^1(\kappa)$ coincides with either set $I_1(\kappa)$ or $I^1(\kappa)$. In such a case, the value of p is without importance, but we shall find that weaker and simpler assumptions can be made to ensure one of the relations $\kappa=\chi_1(\kappa,p)$ or $\kappa=\chi^1(\kappa,p)$ by taking the value of p (given by (4.23)) into account. Besides, we shall see in § 5 that there are standard problems in which the full criterion (4.30) cannot be limited to any of these simple two forms.

We now give prominence to two necessary conditions for Theorem 4.2 to be available. The first one is important because it will allow us to derive precise information on the structure of the zero set of the mapping H_σ and is a common hypothesis in these problems. The second one is not usual and shows that there are only *two* values of p for which a general discussion is possible.

PROPOSITION 4.3. *A first necessary condition for Theorem* 4.2 *to apply in full is*

$$Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa\neq0\quad\textit{for every } x_1\in X_1-\{0\}. \tag{4.31}$$

A second necessary condition is that p is a divisor of κ *and* $(0,\kappa/p)\in I_1^1(\kappa)$ *or p is a divisor of* $\kappa-1$ *and* $(1,[(\kappa-1)/p]+1)\in I_1^1(\kappa)$.

Proof. To show (4.31), we parallel the beginning of the proof of Theorem 4.1: if there is an element $x_1\in X_1-\{0\}$ such that $Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa=0$, η^p factors out in the expression $(4.21)_\sigma$ for this specific value x_1. As $p\geqq2$, the derivative of $(4.21)_\sigma$ at $(0,x_1)$ has rank deficiency $\geqq1$, contradicting the condition ($\mathbb{R}$-N.D.).

The second assertion follows from the observation that the mapping $(4.21)_\sigma$ cannot verify the condition ($\mathbb{R}$-N.D.) if there is no pair $(i,j)\in I_1^1(\kappa)$ such that $pj+(1-p)i=\kappa$ and $i=0$ or $i=1$. Indeed, in this case, the expression $(4.21)_\sigma$ involves pairs with $i\geqq2$ only and vanishes together with its derivative at the points of the line $\mathbb{R}\times\{0\}$. Therefore, there must be a pair $(i_0,j_0)\in I_1^1(\kappa)$ for which $i_0=0$ and $pj_0=\kappa$ so that p divides κ and $(0,\kappa/p)\in I_1^1(\kappa)$ or $i_0=1$ and $pj_0+(1-p)=\kappa$ so that p divides $\kappa-1$ and $(1,[(\kappa-1)/p]+1)\in I_1^1(\kappa)$. □

Remark 4.3. As the sole pair (j, j) such that $pj+(1-p)j(=j)=\kappa$ is obviously the pair (κ, κ), (4.31) means that there is no element of the form $(0, x_1)$, $x_1 \in X_1-\{0\}$ in the zero set of $(4.21)_\sigma$. When $N=1$, (4.31) is not an additional assumption.

A classification of the mappings G to which Theorem 4.2 may apply is then based on the divisors of κ or $\kappa-1$ (observe in passing that κ and $\kappa-1$ are always prime to each other). A general discussion is then impossible unless $p=\kappa$ or $p=\kappa-1$. These cases will be considered later on.

To complete these general considerations, we now show how the structure of the local zero set of G can be derived from the structure of the local zero set of the mappings H_σ. Of course, we suppose that the assumptions of Theorem 4.2 are fulfilled so that the latter consists of a finite number $\nu_\sigma \leqq \kappa^N$ (possibly 0) of curves of class $\mathscr{C}^m$ away from the origin and $\mathscr{C}^{m-\kappa+1}$ at the origin, where each of them is tangent to a different one of the lines in the zero set of the mapping $(4.21)_\sigma$. Let

$$t \to (\eta(t), x(t)) \in \mathbb{R} \times X, \quad \eta(0)=0, \quad x(0)=0 \tag{4.32}$$

be such a curve. From the above, the parameterization can be chosen so that $((d\eta/dt)(0), (dx/dt)(0))$ is a *nonzero* element of the zero set of the mapping $(4.21)_\sigma$ (this can be directly established by looking at the proof of Theorem 1.1 and examining how Theorem 2.2 is derived from it). In particular, $(dx/dt)(0) \in X_1$ and, using Remark 4.3, we deduce

$$(d\eta/dt)(0) \neq 0. \tag{4.33}$$

Also, for each curve (4.32) except that one tangent to the line $\mathbb{R} \times \{0\}$ at the origin (when it is in the zero set of $(4.21)_\sigma$), one has

$$\frac{dx}{dt}(0) \neq 0. \tag{4.34}$$

Remark 4.4. From Proposition 4.3, the line $\mathbb{R} \times \{0\}$ is not in the zero set of $(4.21)_\sigma$ when p is a divisor of κ since the pair $(0, \kappa/p)$ belongs to the set $I_1^1(\kappa)$. On the contrary, it is in the zero set of $(4.21)_\sigma$ when p divides $\kappa-1$ because there is no pair $(0, j) \in I_1^1(\kappa)$ such that $pj=\kappa$ for p would also be a divisor of κ. These remarks are independent of the choice $\sigma=1$ or $\sigma=-1$.

Assume first that $p(\geqq 3)$ is odd. The zero sets of the mappings G and H_σ are homeomorphic and Theorem 2.2 applies with H_1 or H_{-1} equivalently (the verification of this fact is obvious). Thus, the zero set of G consists of the $\nu_1(=\nu_{-1})$ distinct curves

$$t \to \gamma(t)=(\eta^p(t), x(t)) \in \mathbb{R} \times X, \tag{4.35}$$

where (4.32) is one of the ν_1 curves in the local zero set of H_1. As $p \geqq 3$, $(d\gamma/dt)(0)=(0, (dx/dt)(0))$: if p is a divisor of κ, one has $(d\gamma/dt)(0) \neq 0$ (cf. (4.34) and Remark 4.4) and the ν_1 curves (4.35) are of class $\mathscr{C}^{m-\kappa+1}$ at the origin, $\mathscr{C}^m$ away from it. If p is a divisor of $\kappa-1$, $(d\gamma/dt)(0) \neq 0$ except when γ corresponds with that curve (4.32) which is tangent to the line $\mathbb{R} \times \{0\}$ at the origin. As a result, all the ν_1 curves in the zero set of G except possibly one are found to be of class $\mathscr{C}^{m-\kappa+1}$ at the origin and $\mathscr{C}^m$ away from it. The remaining curve is easily seen to be $\mathscr{C}^m$ away from the origin but we can only conclude to its continuity at the origin. Observe finally from (4.33) that $\eta(t)$, and hence $\eta^p(t)$, takes positive and negative values as t is varied around the origin: all the ν_1 curves (4.35) cross the hyperplane $\mu=0$ at the origin.

Next, assume that $p(\geqq 2)$ is even. Taking $\sigma=1$, the solutions of the equation $H_1(\eta, x)=0$ provide the solutions of the equation $G(\mu, x)=0$ with $\mu \geqq 0$. These solutions are then obtained through the ν_1 curves

$$t \to \gamma_+(t)=(\eta^p(t), x(t)) \in \mathbb{R} \times X, \tag{4.36}$$

where (4.32) denotes any of the ν_1 curves in the local zero set of H_1. Contrary to what happens when p is odd, the ν_1 curves (4.36) are not distinct here because $H_1(-\eta, x) = H_1(\eta, x)$, so that the curve $(-\eta(t), x(t))$ is in the zero set of H_1 as soon as $(\eta(t), x(t))$ is in it. Clearly, they provide the same curve γ_+ (4.36) and, except in one case, are distinct. To see this, notice from (4.33) that the vectors $((d\eta/dt)(0), (dx/dt)(0))$ and $(-(d\eta/dt)(0), (dx/dt)(0))$ are not collinear unless $(dx/dt)(0) = 0$, which happens, for one curve only, when p divides $\kappa - 1$. Thus, when p is a divisor of κ, each curve (4.36) is provided by two distinct curves in the local zero set of H_1 and the exact number of distinct curves (4.36) is $\nu_1/2$ (hence, ν_1 is even; this can be found a priori since p is even and divides κ, so that κ is even and it suffices to apply [3, Thm. 2.7]). The regularity of the curves is the same as when p is odd. When p is a divisor of $\kappa - 1$, the two curves $(\eta(t), x(t))$ and $(-\eta(t), x(t))$ with $(dx/dt)(0) = 0$ are both tangent to the line $\mathbb{R} \times \{0\}$ at the origin and hence coincide (which merely means that the curve is symmetric with respect to the hyperplane $\eta = 0$, but of course not that $\eta(t) = 0$). The corresponding curve γ_+ (4.36) is then a half-branch emerging from the origin into the half-space $\mu \geqq 0$ in $\mathbb{R} \times X$. The exact number of curves γ_+, including the half-branch, is $(\nu_1 + 1)/2$ (hence, ν_1 is odd, as κ is since p is even and divides $\kappa - 1$) and, except for the half-branch, their regularity is the same as when p is odd.

Taking $\sigma = -1$ and by the same arguments, we find that the solutions of the equation $G(\mu, x) = 0$ with $\mu \leqq 0$ are provided by the ν_{-1} curves

$$t \to \gamma_-(t) = (-\eta^p(t), x(t)) \in \mathbb{R} \times X, \tag{4.37}$$

where (4.32) is one of the ν_{-1} curves in the local zero set of H_{-1}. There are $\nu_{-1}/2$ distinct curves γ_- when p is a divisor of κ and $(\nu_{-1} + 1)/2$ when p divides $\kappa - 1$ (including a half-branch emerging from the origin into the half-space $\mu \leqq 0$ in $\mathbb{R} \times X$). The regularity results are identical to those when $\sigma = 1$.

Observe that the exact number of curves in the zero set of G is always $(\nu_1 + \nu_{-1})/2$. This is obvious when p is a divisor of κ. When p is a divisor of $\kappa - 1$, a first estimate is $[(\nu_1 + \nu_{-1})/2] + 1$. But if we do so, the two half-branches emerging from the origin into the half-spaces $\mu \geqq 0$ and $\mu \leqq 0$ are counted separately while they extend each other as a single branch.

To complete this section, we shall examine in some detail the two particular cases $p = \kappa$ and $p = \kappa - 1$, as motivated by Proposition 4.3.

The case $p = \kappa$. From Proposition 4.3, the pair $(0, 1)$ is in the set $I_1^1(\kappa)$. Conversely, it is immediately checked that $p = \kappa$ if $(0, 1) \in I_1^1(\kappa)$. By definition of the operator Q_1, we deduce

$$p = \kappa \Leftrightarrow D_\mu G(0) \notin Z_2. \tag{4.38}$$

Surprisingly enough, we find again the first case of the classification of § 3. The criterion (4.30) becomes

$$\kappa \leqq \chi_1(\kappa, \kappa) + \chi^1(\kappa, \kappa) - 2, \tag{4.39}$$

where the quantities $\chi^1(\kappa, \kappa)$ and $\chi_1(\kappa, \kappa)$ are easily found to be

$$\chi^1(\kappa, \kappa) = \min\{0 \leqq j \leqq \kappa, D_x^j G(0)|_{(X_1)^j} \neq 0\}, \tag{4.40}$$

$$\chi_1(\kappa, \kappa) = \min\{0 \leqq j \leqq \kappa, Q_1 D_x^j G(0) \neq 0\}. \tag{4.41}$$

The only pairs (i, j) such that $0 \leqq i \leqq j \leqq \kappa$ and $\kappa j + (1 - \kappa) i = \kappa$ are $(0, 1)$ and (κ, κ). Therefore, the mapping $(4.21)_\sigma$ (with $p = \kappa$) reads

$$(\eta, x_1) \in \mathbb{R} \times X_1 \to \sigma\kappa!\, \eta^\kappa Q_1 D_\mu G(0) + Q_1 D_x^\kappa G(0) \cdot (x_1)^\kappa \in Z_1. \tag{4.42$_\sigma$}$$

In § 3, the analysis of the same problem led us to consider an arbitrary complement Y_1 of $\mathbb{R}D_\mu G(0)\oplus Z_2$ in Y with associated projection operator P_1. Proposition 4.4 applies to this notation.

PROPOSITION 4.4. *Both mappings* $(4.42)_1$ *and* $(4.42)_{-1}$ *verify the condition* ($\mathbb{R}$-N.D.) *if and only if*

$$(4.43)\qquad Q_1 D_x^\kappa G(0)\cdot(x_1)^\kappa \neq 0 \quad \textit{for every } x_1\in X_1-\{0\}$$

and the mapping

$$(4.44)\qquad x_1\in X_1 \to P_1 D_x^\kappa G(0)\cdot (x_1)^\kappa \in Y_1$$

verifies the condition ($\mathbb{R}$-N.D.).

Proof. We already know that whether or not the mapping $(4.42)_\sigma$ verifies the condition ($\mathbb{R}$-N.D.) is independent of the choice of the space Z_1 (cf. Remark 4.2) and that the condition (4.43) is necessary (Proposition 4.3). For the sake of convenience, we may then suppose

$$(4.45)\qquad Z_1=\mathbb{R}D_\mu G(0)\oplus Y_1.$$

If so, $Q_1D_\mu G(0)=D_\mu G(0)$, $P_1Q_1=Q_1P_1=P_1$ and the operator (Q_1-P_1) is the projection onto the space $\mathbb{R}D_\mu G(0)$ associated with the decomposition $Y=\mathbb{R}D_\mu G(0)\oplus(Y_1\oplus Z_2)$. Thus, the mapping $(4.42)_\sigma$ is

$$(4.46)_\sigma\qquad (\eta, x_1)\in\mathbb{R}\times X_1\to \sigma\kappa!\,\eta^\kappa D_\mu G(0)+Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa\in Z_1.$$

Assume first that $(4.46)_1$ and $(4.46)_{-1}$ verify the condition ($\mathbb{R}$-N.D.). Let $x_1\in X_1$ be a nonzero element of the zero set of the mapping (4.44). One has

$$Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa=(Q_1-P_1)D_x^\kappa G(0)\cdot(x_1)^\kappa.$$

From the above, the right-hand side is collinear with $D_\mu G(0)$. As a result, $Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa=\lambda D_\mu G(0)$ for some real number λ. For either $\sigma=1$ or $\sigma=-1$ (and for both if κ is odd) there is $\eta\in\mathbb{R}$ such that $\sigma\kappa!\,\eta^\kappa=-\lambda$ so that the pair (η,x_1) is in the zero set of the mapping $(4.46)_\sigma$. Let $y_1\in Y_1$ be given. In particular, $y_1\in Z_1$ and there is a pair $(\tau,\xi_1)\in\mathbb{R}\times X_1$ such that

$$\sigma\kappa(\kappa!)\eta^{\kappa-1}\tau D_\mu G(0)+\kappa Q_1D_x^\kappa G(0)\cdot((x_1)^{\kappa-1},\xi_1)=y_1.$$

Projecting onto Y_1, we find $\kappa P_1D_x^\kappa G(0)\cdot((x_1)^{\kappa-1},\xi_1)=y_1$, which proves that the mapping (4.44) verifies the condition ($\mathbb{R}$-N.D.).

Conversely, let $(\eta,x_1)\in\mathbb{R}\times X_1$ be a nonzero element of the zero set of the mapping $(4.46)_\sigma$. Observe that $\eta\neq 0$ from (4.43) and it is obvious that $x_1\neq 0$. Projecting onto Y_1, we see that x_1 is a nonzero element of the zero set of the mapping (4.44). Let $z_1\in Z_1$ be given and set $y_1=P_1z_1$. From ($\mathbb{R}$-N.D.), there is $\xi_1\in X_1$ such that $\kappa D_x^\kappa G(0)\cdot((x_1)^{\kappa-1},\xi_1)=y_1$. As (Q_1-P_1) is the projection onto $\mathbb{R}D_\mu G(0)$ and by definition of y_1, one has $\kappa Q_1D_x^\kappa G(0)\cdot((x_1)^{\kappa-1},\xi_1)=z_1+\lambda D_\mu G(0)$ for some real number λ. Setting $\tau=-\lambda/(\sigma\kappa(\kappa!)\eta^{\kappa-1})$ it is clear that

$$\sigma\kappa(\kappa!)\eta^{\kappa-1}\tau D_\mu G(0)+\kappa Q_1D_x^\kappa G(0)\cdot((x_1)^{\kappa-1},\xi_1)=z_1,$$

so that the mapping $(4.46)_\sigma$ verifies the condition ($\mathbb{R}$-N.D.). □

As $p=\kappa$ divides κ, all the curves are of class $\mathscr{C}^{m-\kappa+1}$ at the origin and $\mathscr{C}^m$ away from it. Note when $N=1$ that the assumptions we make in § 3 in the analysis of the same problem are weaker and the results we obtain there are better: for instance, it is known from § 3 that the local zero set of G is made of exactly one curve. Here, the information is only "at most κ" but existence and uniqueness can also be derived from the special form of the mapping $(4.42)_\sigma$. More generally, by arguments similar to those we used in Proposition 4.4, the maximum number of lines in the zero set of either mapping $(4.42)_\sigma$ (hence the maximum number of curves in the zero set of G) can be shown to be κ^{N-1}, an improvement over the general estimate κ^N that is due to the fact that $p=\kappa$. Also, when $N=1$ in § 3, the curve in the zero set of G is found to be of class $\mathscr{C}^m$ around the origin instead of $\mathscr{C}^{m-\kappa+1}$ here. However, its location with respect to the hyperplane $\mu=0$ in $\mathbb{R}\times X$ cannot be specified under the assumptions of § 3. The hypotheses of this section are trivially satisfied as soon as the criterion (4.39) is fulfilled (in particular, note that $Y_1=\{0\}$) and the origin is a turning point if κ is even, a hysteresis point if κ is odd, which agrees with well-known results.

When $N\geqq 2$, the assumptions of § 3 are no longer weaker in general: the analysis involves the order of the first nonzero derivative of the mapping $P_1G(0,\cdot)|_{X_1}$ at the origin, denoted by k, and we observed that $k\geqq\kappa$. Here, the equality $k=\kappa$ is a necessity from Proposition 4.4. However, this is not really restrictive since it was already mentioned in § 3 that the equality $k=\kappa$ holds with most of the values $D_\mu G(0)\notin Z_2$. If so, the result of regularity and the estimate on the number of curves is the same. Condition (4.43) is not required in § 3 but it is also little restrictive in practice (see Remark 3.1). A more significant difference is observed between the criterion (4.39) and its analogue of § 3, which, when $k=\kappa$, is $\kappa\leqq k_1(\kappa)+k^1(\kappa)-2$, where

$$k^1(\kappa)=\min\{0\leqq j\leqq\kappa,\ D_x^jG(0)|_{(X_1)^j}\neq 0\},$$
$$k_1(\kappa)=\min\{0\leqq j\leqq\kappa,\ P_1D^jG(0)\neq 0\}.$$

On comparison with (4.40) and (4.41) we see that $k^1(\kappa)=\chi^1(\kappa,\kappa)$ but there is no relation between $k_1(\kappa)$ and $\chi_1(\kappa,\kappa)$ in general so that either criterion may be satisfied while the other one is not. This is not surprising since a simple examination shows that there is no explicit relation between the two reduced mappings when $D_\mu G(0)\notin Z_2$.

The case $p=\kappa-1$. Of course, this case requires the a priori assumption $\kappa\geqq 3$. From Proposition 4.3, the pair $(1,2)$ belongs to $I_1^1(\kappa)$, namely

$$Q_1D_\mu D_xG(0)\neq 0. \tag{4.47}$$

Also, from the equivalence (4.38) we must assume that

$$D_\mu G(0)\in Z_2. \tag{4.48}$$

Conversely, it is easily checked when (4.47) and (4.48) hold that $p=\kappa-1$. The criterion (4.30) is then

$$\kappa\leqq\chi_1(\kappa,\kappa-1)+\chi^1(\kappa,\kappa-1)-2. \tag{4.49}$$

Elementary considerations lead to

$$\chi^1(\kappa,\kappa-1)=\min\{0\leqq j\leqq\kappa,\ D_x^jG(0)|_{(X_1)^j}\neq 0\}, \tag{4.50}$$

unless $D_\mu G(0)=0$ and $D_x^jG(0)|_{(X_1)^j}=0$, $0\leqq j\leqq\kappa-1$, and, if so

$$\chi^1(\kappa,\kappa-1)=\kappa-1.$$

Next, in any case, $\chi_1(\kappa,\kappa-1)$ is computed to be

$$\chi_1(\kappa,\kappa-1)=\min\{0\leqq j\leqq\kappa,\ Q_1D_x^jG(0)\neq 0\}. \tag{4.51}$$

It is easy to check that the only pairs (i,j) such that $0\leqq i\leqq j\leqq\kappa$ and $(\kappa-1)j+(2-\kappa)i=\kappa$ are $(1,2)$ and (κ,κ). Thus, the mapping $(4.21)_\sigma$ (with $p=\kappa-1$) becomes

$$(\eta,x_1)\in\mathbb{R}\times X_1\to\sigma\kappa!\,\eta^{\kappa-1}Q_1D_\mu D_xG(0)\cdot x_1+Q_1D_x^\kappa G(0)\cdot(x_1)^\kappa\in Z_1. \tag*{$(4.52)_\sigma$}$$

As stated by Proposition 4.3, a necessary condition for this mapping to verify the nondegeneracy condition of this paper is

$$Q_1 D_x^\kappa G(0) \cdot (x_1)^\kappa \neq 0 \quad \text{for every } x_1 \in X_1 - \{0\}. \tag{4.53}$$

As an example, we shall consider the problems of bifurcation from the trivial branch as they were already encountered in § 3: with $X = Y$, the mapping G is of the form

$$G(\mu, x) = x - (\lambda_0 + \mu)Lx + \Gamma(\mu, x), \tag{4.54}$$

where $L \in \mathscr{L}(X)$ is compact, λ_0 is a characteristic value of L and the mapping Γ is of class $\mathscr{C}^m$, $m \geqq 2$, around the origin of $\mathbb{R} \times X$ with values in X and verifies $\Gamma(\mu, 0) = 0$ for μ around $0 \in \mathbb{R}$. In particular

$$D_\mu^j \Gamma(0) = 0, \qquad 0 \leqq j \leqq m, \tag{4.55}$$

and we further require that

$$D_\mu D_x \Gamma(0) = 0, \tag{4.56}$$

$$D_x \Gamma(0) = 0. \tag{4.57}$$

Under these assumptions, the zero set of G was determined in § 3 when the decomposition

$$X = \operatorname{Ker}(I - \lambda_0 L) \oplus \operatorname{Range}(I - \lambda_0 L) \tag{4.58}$$

holds. If so, we can make the choice $Z_1 = X_1$ $(= \operatorname{Ker}(I - \lambda_0 L))$ without loss of generality. We were able to settle the two cases when $\dim \operatorname{Ker}(I - \lambda_0 L) = 1$ or when $\dim \operatorname{Ker}(I - \lambda_0 L) \geqq 2$, $Q_1 D_x^2 \Gamma(0)|_{(X_1)^2} \neq 0$ and the mapping

$$(\mu, x_1) \in \mathbb{R} \times \operatorname{Ker}(I - \lambda_0 L) \to -\frac{2}{\lambda_0} \mu x_1 + Q_1 D_x^2 \Gamma(0) \cdot (x_1)^2 \in \operatorname{Ker}(I - \lambda_0 L),$$

verifies the condition ($\mathbb{R}$-N.D.).

We shall now assume that $Q_1 D_x^2 \Gamma(0)|_{(X_1)^2} = 0$. Clearly, κ is defined by

$$\kappa = \min\{0 \leqq j \leqq m, \; Q_1 D_x^j \Gamma(0)|_{(X_1)^j} \neq 0\}, \tag{4.59}$$

so that $\kappa \geqq 3$. As $D_\mu G(0) = 0$ and $Q_1 D_\mu D_x G(0) = (-1/\lambda_0) Q_1|_{X_1}$, conditions (4.47) and (4.48) are satisfied if and only if $Q_1|_{X_1} \neq 0$ (i.e. $\operatorname{Ker}(I - \lambda_0 L) \not\subset \operatorname{Range}(I - \lambda_0 L)$) so that $p = \kappa - 1$. From (4.50) and (4.51) the quantities $\chi^1(\kappa, \kappa - 1)$ and $\chi_1(\kappa, \kappa - 1)$ are given by

$$\chi^1(\kappa, \kappa - 1) = \min\{0 \leqq j \leqq \kappa, \; D_x^j \Gamma(0)|_{(X_1)^j} \neq 0\}, \tag{4.60}$$

$$\chi_1(\kappa, \kappa - 1) = \min\{0 \leqq j \leqq \kappa, \; Q_1 D_x^j \Gamma(0) \neq 0\}. \tag{4.61}$$

The mapping $(4.52)_\sigma$ is then

$$(\eta, x_1) \in \mathbb{R} \times \operatorname{Ker}(I - \lambda_0 L) \to -\frac{\sigma\kappa!}{\lambda_0} \eta^{\kappa - 1} Q_1 x_1 + Q_1 D_x^\kappa \Gamma(0) \cdot (x_1)^\kappa \in Z_1.$$

Because of the elements of the trivial branch in the zero set, it is easily seen that the condition (4.58) is still necessary for the above mapping to verify the condition ($\mathbb{R}$-N.D.). Again, we may then take $Z_1 = X_1 (= \operatorname{Ker}(I - \lambda_0 L))$ and, with this choice, $(4.52)_\sigma$ reads

$$(\eta, x_1) \in \mathbb{R} \times \operatorname{Ker}(I - \lambda_0 L) \to -\frac{\sigma\kappa!}{\lambda_0} \eta^{\kappa - 1} x_1 + Q_1 D_x^\kappa \Gamma(0) \cdot (x_1)^\kappa \in \operatorname{Ker}(I - \lambda_0 L). \tag*{$(4.62)_\sigma$}$$

The condition (4.53), necessary for it to verify the condition ($\mathbb{R}$-N.D.), is equivalent to

$$Q_1 D_x^\kappa \Gamma(0) \cdot (x_1)^\kappa \neq 0 \quad \text{for every } x_1 \in \operatorname{Ker}(I - \lambda_0 L) - \{0\}. \tag{4.63}$$

As $p = \kappa - 1$ divides $\kappa - 1$, the only curve in the local zero set of G for which continuity at the origin only can be proved in general is the curve tangent to the line $\mathbb{R} \times \{0\}$ at the origin. Here, it is then the trivial branch itself so that the bifurcated branches are of class $\mathscr{C}^{m-\kappa+1}$ at the origin (and $\mathscr{C}^m$ away from it). The location of these curves with respect to the hyperplane $\mu = 0$ is as described before in the general comments and their number cannot exceed $\kappa^N - 1$. When $N = 1$, $\kappa - 1$ is always an overestimate since $\kappa \geqq 3$ and we know that there is exactly one bifurcated branch. When $N = 2$, it is possible to show that a more accurate and optimal estimate is $\kappa + 1$ but we are not aware of a general improvement. When $N = 1$, all the assumptions of this section are satisfied provided the criterion (4.49) is fulfilled and bifurcation is transcritical if κ is even, supercritical if κ is odd. This agrees with well-known results.

When $N \geqq 2$, our assumptions coincide with those of McLeod and Sattinger [11] except for the criterion (4.49) which, in [11], is replaced by

$$D_x^j\Gamma(0) = 0, \qquad 0 \leqq j \leqq \kappa - 1, \tag{4.64}$$

an assumption that makes (4.49) trivially fulfilled. The result on the location of the curves is identical but, beyond continuity, no regularity of the curves at the origin is proved in [11].

Remark 4.5. Assumption (4.64) can be found in a number of publications (see e.g. [5], [11], [17], [18]). Two somewhat more general forms of (4.64), corresponding to the cases $\kappa = \chi^1(\kappa, \kappa - 1)$ and $\kappa = \chi^1(\kappa, \kappa - 1)$ respectively, are

$$Q_1 D_x^j\Gamma(0) = 0, \qquad 0 \leqq j \leqq \kappa - 1, \tag{4.65}$$

or

$$D_x^j\Gamma(0)\big|_{(x_1)^j} = 0, \qquad 0 \leqq j \leqq \kappa - 1. \tag{4.66}$$

Both assumptions are mentioned in [3] and, again, both make the criterion (4.49) trivially fulfilled. Nevertheless, as already mentioned, there are quite standard problems in which (4.49) cannot be reduced to any of these simplified versions (see § 5).

5. Examples. In this section we shall briefly examine the usefulness of some aspects of our approach. We begin with the problem

$$-\Delta u - \lambda u + u^4 + u^5(+\text{higher-order terms}) = 0, \qquad u \in H_0^1(\Omega), \tag{5.1}$$

where $\Omega \subset \mathbb{R}^2$ denotes the square $(0, \pi) \times (0, \pi)$. Problem (5.1) is a typical model of bifurcation from the trivial branch. Denoting by $L \in (H^{-1}(\Omega), H_0^1(\Omega))$ the inverse of $-\Delta$, an equivalent form of (5.1) is

$$u - \lambda L u + L(u^4 + u^5 + \cdots) = 0, \qquad u \in H_0^1(\Omega). \tag{5.2}$$

The characteristic values of L are of the form $\lambda_0 = a^2 + b^2$ where a and b are positive integers and the multiplicity of λ_0 equals the number of such distinct pairs (a, b). Assume first that λ_0 is simple, hence $\lambda_0 = 2a^2$. Then, it is known that there is exactly one bifurcated branch passing through $(\lambda_0, 0)$. If a is odd, it can be shown by classical arguments that bifurcation occurs transcritically. This situation corresponds to the case $\kappa = 4$ in § 4 and the criterion (4.49) holds trivially. On the contrary, if a is even, determining whether bifurcation occurs transcritically or supercritically requires the examination of derivatives of the reduced mapping of order >4. This is precisely when the criterion (4.49) is useful: one can easily check that $\kappa = 5$ and (with the choice $p = \kappa - 1$ as obtained in § 4) $\chi^1(5, 4) = \chi_1(5, 4) = 4$ so that (4.49) is fulfilled and it follows at once that bifurcation occurs supercritically, as if the term u^4 were not present and the problem were

$$-\Delta u - \lambda u + u^5(+\text{higher-order terms}) = 0. \tag{5.3}$$

A similar observation remains true when we study bifurcation at a multiple characteristic value. For instance, bifurcation near the point (10, 0) in the problem (5.1) can be equally studied by the method of McLeod and Sattinger [11] or that of § 4: bifurcation occurs transcritically and there are three bifurcated branches. But bifurcation near the point $(5,0)$ $(\lambda_0=5$ is double) cannot be analyzed under the assumptions of [11] nor those of [3]. Note that the situation is worse than when λ_0 is simple since not merely the location but the very structure of the solutions is unknown. Again, it can easily be shown that $\kappa=5$ while $\chi^1(5,4)=\chi_1(5,4)=4$: due to the criterion (4.49) we see that bifurcation near $(5,0)$ occurs as in the problem (5.3). Further investigation shows that there are exactly four bifurcated branches, located supercritically. This example can be embedded into a general problem, namely

$$(5.4)\qquad -\Delta u-\lambda u+P(u)+\varepsilon u^\kappa(+\text{higher-order terms})=0,\qquad u\in H_0^1(\Omega),$$

where $\Omega=(0,\pi)\times(0,\pi)$, $\kappa\geqq 3$ is odd, $\varepsilon=\pm 1$ and P is an even polynomial with degree $\leqq\kappa-1$ and valuation $k\geqq 2$. Bifurcation is studied near the point $(\lambda_0,0)$ where λ_0 is a double characteristic value of $L=(-\Delta)^{-1}$ of the form $\lambda_0=5a^2(a\in\mathbb{N}-\{0\})$. In any case, κ is then defined as in § 4. If $P=0$, one has $\chi_1(\kappa,\kappa-1)=\chi^1(\kappa,\kappa-1)=\kappa$ and the criterion (4.49) is trivially fulfilled. If $P\neq 0$, it is easy to check that $\chi_1(\kappa,\kappa-1)=\chi^1(\kappa,\kappa-1)=k$ so that (4.49) reduces to the assumption $\kappa\leqq 2k-3$ (which requires $k\geqq 4$ and $\kappa\geqq 5$). After rescaling η, the mapping $(4.62)_\sigma$ identifies with

$$(\eta,\alpha,\beta)\in\mathbb{R}^3\to\begin{pmatrix}\sigma\eta^{\kappa-1}\alpha-\varepsilon\displaystyle\sum_{j=0}^{(\kappa-1)/2}a_{2j+1}\alpha^{2j+1}\beta^{\kappa-2j-1}\\ \sigma\eta^{\kappa-1}\beta-\varepsilon\displaystyle\sum_{j=0}^{(\kappa-1)/2}a_{2j+1}\alpha^{\kappa-2j-1}\beta^{2j+1}\end{pmatrix}\in\mathbb{R}^2,$$

where the coefficients $a_{2j+1}>0$ are explicitly given through the Eulerian function Γ. Arguing as in Bolley [2] it can be shown that the above equation has nontrivial solutions (α,β) for $\sigma=\varepsilon$ only, and in this case its zero set is made of the nine lines generated by $(1,0,0)$, $(1,1,0)$, $(-1,1,0)$, $(1,0,1)$, $(-1,0,1)$, $(\eta_0,1,1)$, $(-\eta_0,1,1)$, $(\eta_0,1,-1)$ and $(-\eta_0,1,-1)$ with

$$\eta_0=\left(\sum_{j=0}^{(\kappa-1)/2}a_{2j+1}\right)^{1/(\kappa-1)}.$$

The nondegeneracy condition is satisfied (such a verification is often straightforward, although sometimes lengthy, when the zero set is known explicitly) and we conclude that problem (5.4) exhibits exactly four bifurcated branches emerging from the point $(\lambda_0,0)$, located supercritically if $\varepsilon=1$ and subcritically for $\varepsilon=-1$. By comparison, the same statement is made in [2] for the problem with $P=0$, $\varepsilon=1$ and no higher-order term

$$-\Delta u-\lambda u+u^\kappa=0$$

only (the equation $-\Delta u-\lambda u-u^\kappa$ cannot be considered in [2] due to growth conditions imposed on the nonlinearity). If $P\neq 0$, our conclusions do not follow from [11] either.

These examples show what criterion (4.49) (or its analogues in other problems) can be good for. They clearly establish that the exponent governing the structure of the bifurcation set in problems like (5.4) depends on the characteristic value λ_0 and, to some extent, allow to determine this "leading" exponent as λ_0 is varied. Other nonlinearities in which the Nemytskii's operator does not only depend on u but also on the generic point $x\in\Omega$ (and also possibly on λ) induce similar phenomena. Problems with nonlocal nonlinearities, such as von Kármán equations, can be considered as well. In this last example, the criterion (4.49) is trivially fulfilled and the nondegeneracy condition alone has to be examined (see Hölder and Schaeffer [7]).

Showing that the condition ($\mathbb{R}$-N.D.) remains appropriate in the study of problems in which no branch of solution is known a priori is our next purpose. For simplicity of calculations and since we do not intend here to discuss the usefulness of some version of the criterion of § 2, we shall consider the problem

$$(5.5)\qquad -\Delta u-\lambda_0 u+u^2=\mu f,\qquad u\in H_0^1(\Omega),$$

where $\Omega\subset\mathbb{R}^2$ is again the square $(0,\pi)\times(0,\pi)$, f is a given element of the space $H^{-1}(\Omega)$ and λ_0 is some eigenvalue of $-\Delta$. With $L=(-\Delta)^{-1}\in\mathscr{L}(H^{-1}(\Omega), H_0^1(\Omega))$ the problem becomes

$$(5.6)\qquad u-\lambda Lu+L(u^2)=\mu F,\qquad u\in H_0^1(\Omega),$$

where we have set $F=Lf$. We shall assume

$$(5.7)\qquad F\notin \text{Range}\,(I-\lambda_0 L),$$

so that the problem falls into the case $D_\mu G(0)\notin Z_2$ of §§ 3 and 4. When λ_0 is simple, hence of the form $\lambda_0=2a^2$, the solutions are given by a single curve. Further, if a is odd, it is easy to check that $\kappa=2$ so that the origin is a turning point (if a is even, higher order derivatives of the reduced mapping are needed to conclude). We shall apply our method to the study of the case when λ_0 is double. More precisely, we shall assume that $\lambda_0=a^2+b^2$ with $a>0$ and $b>0$ odd (and distinct). We leave it to the reader to check that the analysis can be made without changing the parameter μ by the method of § 3 and we shall focus on the method of § 4 since it provides additional information on the location of the curves. The null space Ker $(I-\lambda_0 L)=X_1$ is generated by the two normalized eigenvectors ϕ_{ab} and ϕ_{ba} (the inner product on $H_0^1(\Omega)$ being the usual one $(u,v)=\int_\Omega \nabla u\nabla v$) with

$$\phi_{ab}(t,s)=\frac{2}{\pi\sqrt{\lambda_0}}\sin at\sin bs$$

and $\phi_{ba}(t,s)=\phi_{ab}(s,t)$. With the choice $Z_1=X_1$ one has

$$Q_1F=y_{ab}\phi_{ab}+y_{ba}\phi_{ba},$$

with $y_{ab}, y_{ba}\in\mathbb{R}$ and $y_{ab}^2+y_{ba}^2\neq 0$ from (5.7). After renorming F (i.e. rescaling μ) it is not restrictive to assume

$$y_{ab}^2+y_{ba}^2=1.$$

It is easy to check that $\kappa=2$ and, after rescaling η for ease of calculations, the mapping $(4.42)_\sigma$ identifies with

$$(5.8)_\sigma\qquad (\eta,\alpha,\beta)\in\mathbb{R}^3\to\begin{pmatrix}-\sigma y_{ab}\eta^2+A\alpha^2+2B\alpha\beta+B\beta^2\\ -\sigma y_{ba}\eta^2+B\alpha^2+2B\alpha\beta+A\beta^2\end{pmatrix}\in\mathbb{R}^2,$$

where

$$A=\frac{16}{9ab},\qquad B=\frac{16ab}{(b^2-4a^2)(a^2-4b^2)}.$$

Applying Proposition 4.4 we can see that the condition (4.43) is automatically satisfied from the relations $A\neq B$ and $A+3B\neq 0$ (the former follows from $A/B<1$ and the latter holds because a^2/b^2 could not be rational otherwise). Taking $Y_1=\mathbb{R}(y_{ba}, -y_{ab})$ in Proposition 4.4, it suffices to prove that the mapping

$$(\alpha, \beta)\in\mathbb{R}^2\to(Ay_{ba}-By_{ab})\alpha^2+2B(y_{ba}-y_{ab})\alpha\beta+(By_{ba}-Ay_{ab})\beta^2\in\mathbb{R}$$

verifies the condition ($\mathbb{R}$-N.D.). As the above mapping is quadratic, it is equivalent to check that the discriminant

$$\Delta(F)=y_{ba}^2-\left(\frac{A}{B}+1\right)y_{ab}y_{ba}+y_{ab}^2$$

is nonzero (which will happen for most choices of F). Note that $\Delta(F)>0$ regardless of F when $(A/B)+3>0$ but the sign of $\Delta(F)$ will be negative for some choices of F if $(A/B)+3<0$. If $\Delta(F)<0$, the zero set of $(5.8)_\sigma$ reduces to the origin for both $\sigma=1$ and $\sigma=-1$. In other words, the origin is an *isolated solution* of the problem (5.4) for those F with $\Delta(F)<0$. Note that the condition $(A/B)+3<0$ is satisfied with $\lambda_0=26$ so that this situation does happen. If $\Delta(F)>0$, the zero set of $(5.8)_\sigma$ depends on σ and F. More precisely, define K_i, $i=1, 2$, to be the two distinct roots of the polynomial[4]

$$(By_{ba}-Ay_{ab})K^2+2B(y_{ba}-y_{ab})K+(Ay_{ba}-By_{ab})=0$$

and set

$$\theta_i=(By_{ab}+Ay_{ba})K_i^2+2B(y_{ab}+y_{ba})K_i+(Ay_{ab}+By_{ba}).$$

It can be shown that $\theta_i\neq 0$, $i=1, 2$ (an easy way is to observe that either relation $\theta_1=0$ or $\theta_2=0$ would contradict the fact that condition (4.43) is satisfied) and the discussion is as follows: if θ_1 and θ_2 have the same sign, the solutions of problem (5.4) near the origin consist of two curves located on the same half-space $\mu\geqq 0$ or $\mu\leqq 0$ (depending on the sign of θ_1 and θ_2). If θ_1 and θ_2 have opposite signs, the solutions of problem (5.4) near the origin still consist of two curves, but one of them is located on the half-space $\mu\leqq 0$ and the other is located on the half-space $\mu\geqq 0$. That both situations do occur depending on λ_0 (with the same F) is easily seen: take for instance

$$F=\frac{\sqrt{2}}{2}(\phi_{13}+\phi_{31}+\phi_{15}+\phi_{51}).$$

Then, if $\lambda_0=10$, one has $\theta_1>0$ and $\theta_2<0$ while, if $\lambda_0=26$, one has $\theta_1>0$ and $\theta_2>0$.

Acknowledgments. I wish to thank Prof. F. Weissler for stimulating discussions while I was visiting the University of Texas at Austin. I am also grateful to one of the referees for numerous helpful comments on the presentation of this paper.

REFERENCES

[1] W. J. BEYN, personal communication.

[2] C. BOLLEY, *Multiple solutions of a bifurcation problem*, in Bifurcation and Nonlinear Eigenvalue Problems, Proceedings, Villetaneuse, France (1978), C. Bardos, J. M. Lasry and M. Schatzman, eds., Lecture Notes in Math., 782, Springer, Berlin, New York, 1980, pp. 42-60.

[3] M. BUCHNER, J. MARSDEN AND S. SCHECTER, *Applications of the blowing-up construction and algebraic geometry to bifurcation problems*, J. Differential Equations, 48 (1983), pp. 404-433.

[4] M. CRANDALL AND P. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321-340.

[5] E. DANCER, *Bifurcation theory in real Banach spaces*, Proc. London Math. Soc., 23 (1971), pp. 699-734.

[4] Assuming $By_{ba}-Ay_{ab}\neq 0$. Otherwise, $Ay_{ba}-By_{ab}$ is nonzero and it suffices to exchange the roles of α and β.

[6] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. I, Cambridge Univ. Press, Cambridge, 1968.

[7] E. J. HÖLDER AND D. SCHAEFFER, *Boundary conditions and mode jumping in the von Kármán equations*, this Journal, 15 (1984), pp. 447-458.

[8] M. A. KRASNOSELSKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, MacMillan, Pergamon Press, New York, 1964.

[9] N. H. KUIPER, *$\mathscr{C}^1$ equivalence of functions near isolated critical points*, Symposium on Infinite-Dimensional Topology, R. D. Anderson, ed., Ann. of Math. Stud., 69, Princeton Univ. Press, 1972.

[10] K. A. LANDMAN AND S. ROSENBLAT, *Bifurcation from a multiple eigenvalue and stability of solutions*, SIAM J. Appl. Math., 34 (1978), pp. 743-759.

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

[12] R. J. MAGNUS, *On the local structure of the zero set of a Banach space valued mapping*, J. Funct. Anal., 22 (1976), pp. 58-72.

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

[14] P. RABIER, *A generalization of the Implicit function theorem for mappings from $\mathbb{R}^{n+1}$ into $\mathbb{R}^n$ and its applications*, J. Funct. Anal., 56 (1984), pp. 145-170.

[15] ———, *Topics in one-parameter nonlinear problems*, 76, Tata Institute Lecture Notes, Springer, Berlin, New York, 1985.

[16] P. RABIER AND S. EL HAJJI, *New algorithms for the computation of the branches in one-parameter nonlinear problems*, Comput. Mech., to appear.

[17] D. SATHER, *Branching of solutions of nonlinear equations*, Rocky Mountain J. Math., 3 (1973), pp. 203-250.

[18] D. H. SATTINGER, *Group theoretic methods in bifurcation theory*, Lecture Notes in Math., 762, Springer, Berlin, New York, 1979.

[19] A. SZULKIN, *Local structure of the zero set of differentiable mappings and application to bifurcation theory*, Math. Scand., 45 (1979), pp. 232-242.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

016

THE SADDLE-NODE SEPARATRIX-LOOP BIFURCATION*

STEPHEN SCHECTER†

Abstract. We study vector fieldx $\dot{x}=f(x)$, $x\in\mathbb{R}^2$, having at some point an equilibrium of saddle-node type with a separatrix loop. Such vector fields fill a codimension two submanifold Σ of an appropriate Banach space. We give analytic conditions that determine whether a two-parameter perturbation of $\dot{x}=f(x)$ is transverse to Σ. The new condition is a version of Melnikov's integral around the separatrix loop. If it is nonzero, then as one perturbs away from $\dot{x}=f(x)$ in the direction in which an equilibrium of saddle-node type persists, the separatrix loop breaks in a nondegenerate manner. This integral is shown to be nonzero for the two-parameter pendulum equation $\beta\ddot{\phi}+\dot{\phi}+\sin\phi=\rho$ at its organizing center.

Key words. saddle-node separatrix-loop bifurcation, Melnikov integral, pendulum, Josephson junction

AMS(MOS) subject classification. 58F14

1. Introduction. We shall be concerned with vector fields

$$\dot{x}=f(x),\qquad x\in\mathbb{R}^2 \tag{1}$$

having at some $p\in\mathbb{R}^2$ an equilibrium of saddle-node type with a separatrix loop Γ (see Fig. 1). We assume that the saddle-node has one negative eigenvalue and, of course, one zero eigenvalue. Such vector fields fill a codimension two submanifold Σ of an appropriate Banach space of planar vector fields. Consider a two-parameter unfolding of (1),

$$\dot{x}=\tilde{f}(x,\nu_1,\nu_2),\qquad x\in\mathbb{R}^2,\quad \nu_1,\nu_2\in\mathbb{R} \tag{2}$$

where $\tilde{f}(x,0,0)=f(x)$. We shall give a computable condition that determines whether the family (2) is transverse to Σ at $(\nu_1,\nu_2)=(0,0)$.

If the transversality condition is satisfied, there is a smooth nonsingular change of coordinates in parameter space,

$$(\nu_1,\nu_2)\leftrightarrow(\mu_1,\mu_2),\qquad (0,0)\leftrightarrow(0,0),$$

such that

$$\dot{x}=\tilde{f}(x,\nu_1(\mu_1,\mu_2),\nu_2(\mu_1,\mu_2))\underset{\text{def}}{=}f(x,\mu_1,\mu_2)$$

has the bifurcation diagram of Fig. 2 in a neighborhood of $(\mu_1,\mu_2)=(0,0)$. The curve C lies in $\{(\mu_1,\mu_2)\colon \mu_1\leqq 0,\ \mu_2\geqq 0\}$. It has a quadratic tangency with the μ_2-axis at

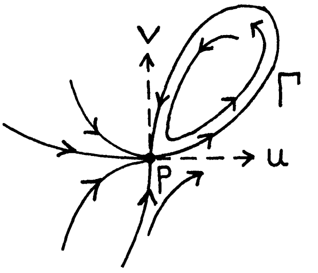


FIG. 1

* Received by the editors September 23, 1985; accepted for publication July 30, 1986.

† Mathematics Department, North Carolina State University, Raleigh, North Carolina 27695-8205.

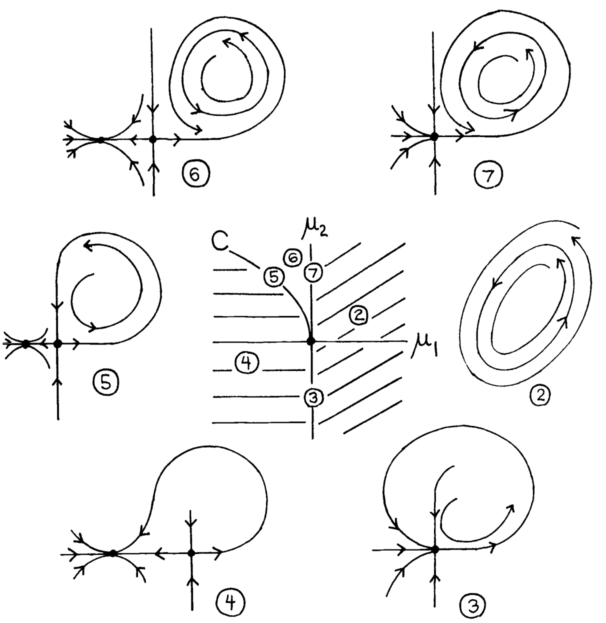


FIG. 2

(0, 0). The phase portrait of $\dot{x}=f(x, \mu_1, \mu_2)$ in a fixed neighborhood of Γ that is positively invariant for each vector field $\dot{x}=f(x, \mu_1, \mu_2)$ is as follows (see Fig. 2):

1) $\mu_1=0$, $\mu_2=0$; a saddle-node and a separatrix loop.

2) $\mu_1>0$; no equilibria, a unique stable closed orbit near Γ.

3) $\mu_1=0$, $\mu_2<0$; a saddle-node.

4) $\mu_1<0$, (μ_1, μ_2) below C; a saddle and a node.

5) $\mu_1<0$, (μ_1, μ_2) on C; a saddle and a node; the saddle has a separatrix loop.

6) $\mu_1<0$, (μ_1, μ_2) above C; a saddle and a node; there is a unique stable closed orbit near Γ.

7) $\mu_1=0$, $\mu_2>0$; a saddle-node and a unique stable closed orbit near Γ.

This bifurcation diagram is developed in [4], except that it is mistakenly stated there that the curve C is transverse to the μ_2-axis.

Perhaps the best known example of this bifurcation diagram occurs in the study of the differential equation for a pendulum with linear damping and constant applied torque, which are the two parameters (see [3], in which the same equation arises in the study of the DC current-driven point Josephson junction). In § 5 we show that the pendulum equation satisfies our transversality condition at its organizing center. Thus the pendulum equation is a generic two-parameter unfolding of the saddle-node separatrix-loop bifurcation.

The heart of this paper is the study, in § 3, of Melnikov's integral (see [2]) around a saddle-node separatrix loop. The same method allows one to study time-periodic perturbations of a saddle-node separatrix loop. This subject is treated in the companion paper [6]. There the motivating example is the pendulum equation with, in addition, sinusoidal applied torque (or, equivalently, the AC–DC current-driven point Josephson junction).

2. Statement of results.

We shall consider vector fields $\dot{x}=f(x)$, $x\in\mathbb{R}^2$, satisfying the following conditions at some $p\in\mathbb{R}^2$:

(i) $f(p)=0$.

(ii) $Df(p)$ has eigenvalues 0 and $-\lambda$, where $\lambda > 0$.

Let u be a right eigenvector and w a left eigenvector of the eigenvalue 0, with w chosen so that $wu > 0$.

(iii) $wD^2f(p)(u, u) > 0$.

(iv) $\dot{x} = f(x)$ has a separatrix loop Γ at p.

Assumptions (i)-(iii) say that $\dot{x} = f(x)$ has a saddle-node at p with one negative eigenvalue (see [7]). Moreover, the assumptions imply that u (not $-u$) is one tangent vector to Γ at p (see Fig. 1). Let v be a right eigenvector of $Df(p)$ for the eigenvalue $-\lambda$, chosen so that v is also tangent to Γ at p as in Fig. 1.

Let $\dot{x} = \tilde{f}(x, \nu_1, \nu_2)$ be a two-parameter family of vector fields on $\mathbb{R}^2$ such that $\dot{x} = \tilde{f}(x, 0, 0)$ satisfies (i)-(iv), and

(v) $\tilde{f}(x, \nu_1, \nu_2)$ is C^{k+1}, $k \geqq 5$.

(vi) $wD_{\nu_1}\tilde{f}(p, 0, 0) > 0$.

Assumptions (iii) and (vi) imply that perturbation in the positive ν_1 direction eliminates the equilibrium p, while perturbation in the negative ν_1 direction splits the equilibrium in two (see [7]).

According to [7] there is a C^k function $\alpha(\nu_2)$, with $\alpha(0) = 0$, such that $\dot{x} = \tilde{f}(x, \nu_1, \nu_2)$ has an equilibrium of saddle-node type near p if and only if $\nu_1 = \alpha(\nu_2)$. Let $f(x, \mu_1, \mu_2) = \tilde{f}(x, \nu_1, \nu_2)$, where (μ_1, μ_2) and (ν_1, ν_2) are related by

$$\mu_1 = \nu_1 - \alpha(\nu_2), \qquad \mu_2 = \nu_2.$$

Then

$$\dot{x} = f(x, \mu_1, \mu_2) \tag{3}$$

is C^k, and has an equilibrium of saddle-node type near p if and only if $\mu_1 = 0$. Let $p(\mu_2)$ denote the saddle-node equilibrium near p of $\dot{x} = f(x, 0, \mu_2)$; $p(\mu_2)$ is C^k. If $\mu_1 < 0$, there are a saddle and a sink of (3) near p; if $\mu_1 > 0$, there are no equilibria of (3) near p.

If w and z are vectors in $\mathbb{R}^2$, let $w \wedge z = w_1 z_2 - w_2 z_1$. Let $q(t)$ be a solution of $\dot{x} = f(x, 0, 0)$ with $q(0) \in \Gamma$. Consider the expression

$$\begin{aligned} I = {} & \frac{dp}{d\mu_2}(0) \wedge \lim_{t_1 \to \infty} f(q(t_1), 0, 0) \exp\left[-\int_0^{t_1} \operatorname{div} f(q(s), 0, 0)\, ds\right] \\ & + \int_{-\infty}^{\infty} \exp\left[-\int_0^{t} \operatorname{div} f(q(s), 0, 0)\, ds\right] f(q(t), 0, 0) \wedge \frac{\partial f}{\partial \mu_2}(q(t), 0, 0)\, dt. \end{aligned} \tag{4}$$

THEOREM 1. *The limit in* (4) *exists and is a negative multiple of v. The improper integral in* (4) *converges. If $I \neq 0$, then there is a C^{k-2} curve $\mu_1 = \psi(\mu_2) = -\ell^2\mu_2^2 + o(\mu_2^2)$, $\ell \neq 0$, such that for (μ_1, μ_2) sufficiently small,* (3) *has a separatrix loop near Γ if and only if $\mu_1 = \psi(\mu_2)$ and $I \cdot (u \wedge v) \cdot \mu_2 \geqq 0$.*

The integral in (4) is just the usual Melnikov integral used to study perturbations of a separatrix loop at a hyperbolic saddle (see [2], [5]). The limit in (4) is zero in the case of a hyperbolic saddle, but must be retained in the case of a saddle-node.

The bifurcation diagram presented in § 1 holds if $I \cdot (u \wedge v) > 0$, in which case separatrix loops occur for $\mu_1 = \psi(\mu_2)$ and $\mu_2 \geqq 0$. If $I \cdot (u \wedge v) < 0$, this bifurcation diagram holds after the further change of parameter $\mu_2 \to -\mu_2$.

We remark that in order to compute I in applications, the only knowledge of the function $\alpha(\nu_2)$ that is needed is $\alpha'(0)$.

We shall now give a precise interpretation of the condition $I \neq 0$ as a transversality condition. Let E denote the space of C^{k+1} vector fields, $k \geqq 5$, on a closed disk $D \subset \mathbb{R}^2$, with the C^{k+1} topology. Let $\Sigma' = \{f \in E : f$ satisfies, at a unique $p \in \operatorname{Int} D$, conditions (i)-(iii); and all other equilibria of f in D are hyperbolic$\}$. Σ' is a C^k codimension one submanifold of E [7]. Let $\Sigma = \{f \in \Sigma' : f$ satisfies condition (iv), and $\Gamma \subset \operatorname{Int} D\}$. We shall show in § 3 that Σ is a C^k codimension two submanifold of E, in fact a codimension one submanifold of Σ'. Let $\dot{x} = \tilde{f}(x, \nu_1, \nu_2)$ be a two-parameter family of vector fields in E with $\dot{x} = \tilde{f}(x, 0, 0) \in \Sigma$. Assume in addition that $\dot{x} = \tilde{f}(x, \nu_1, \nu_2)$ satisfies conditions (v) and (vi). Make the change of parameters $\mu_1 = \nu_1 - \alpha(\nu_2)$, $\mu_2 = \nu_2$ described earlier.

THEOREM 2. *The family $\dot{x} = \tilde{f}(x, \nu_1, \nu_2)$ is transverse to Σ at $(\nu_1, \nu_2) = (0, 0)$ if and only if $I \neq 0$.*

3. Proof of Theorem 1. The equilibrium $(p, 0, 0)$ of

$$\dot{x} = f(x, \mu_1, \mu_2), \quad \dot{\mu}_1 = 0, \quad \dot{\mu}_2 = 0 \tag{5}$$

has a 3-dimensional neutral subspace. The center manifold theorem [1, § 9.2] yields a 3-dimensional C^k local center manifold N_{loc} of (5), tangent at $(p, 0, 0)$ to this subspace. N_{loc} meets each plane $\mathbb{R}^2 \times \{(\mu_1, \mu_2)\}$, (μ_1, μ_2) small, in a curve. $N_{\text{loc}} \cap \mathbb{R}^2 \times \{(0, 0)\}$ contains a portion of $\Gamma \times \{(0, 0)\}$ that is tangent at $(p, 0, 0)$ to $(u, 0, 0)$.

Let N denote the global center manifold that contains N_{loc}, i.e., the union of all integral curves of (5) that meet N_{loc}. N meets each plane $\mathbb{R}^2 \times \{(\mu_1, \mu_2)\}$ in a curve, which we denote $N(\mu_1, \mu_2) \times \{(\mu_1, \mu_2)\}$. Thus $N(\mu_1, \mu_2)$ is a curve in $\mathbb{R}^2$. Let L be a line segment in $\mathbb{R}^2$ perpendicular to Γ at $q(0)$. Then for (μ_1, μ_2) small, $N(\mu_1, \mu_2)$ meets L transversally near $q(0)$. Therefore for (μ_1, μ_2) small there is a C^k function $x(\mu_1, \mu_2)$ such that $x(0, 0) = q(0)$ and $x(\mu_1, \mu_2) \in N(\mu_1, \mu_2) \cap L$. Since a C^k vector field has a C^k flow, there is a C^k family of solutions of (3)

$$q^c(\mu_1, \mu_2, t), \qquad (\mu_1, \mu_2) \text{ small},$$

such that $q^c(\mu_1, \mu_2, 0) = x(\mu_1, \mu_2)$. Then $q^c(0, 0, t) = q(t)$, and each curve $q^c(\mu_1, \mu_2, t)$ lies in $N(\mu_1, \mu_2)$. For $\mu_1 < 0$, $q^c(\mu_1, \mu_2, t)$ is a branch of the unstable manifold of the saddle of (3) near p. Similarly, $q^c(0, \mu_2, t)$ is the unstable separatrix of the saddle-node of $\dot{x} = f(x, 0, \mu_2)$ near p (see Fig. 3).

We shall now define a μ-dependent change of coordinates on $\mathbb{R}^2$ that will make possible our computations. According to [1, § 9.2] there is a C^k change of coordinates

$$y(x, \mu_1, \mu_2) = (y_1(x, \mu_1, \mu_2), y_2(x, \mu_1, \mu_2)), \tag{6}$$

defined for (x, μ_1, μ_2) near $(p, 0, 0)$, such that (1) $y(p, 0, 0) = 0$; (2) $N_{\text{loc}} \cap \mathbb{R}^2 \times \{(\mu_1, \mu_2)\}$ is transformed into the line $y_2 = 0$, which is therefore invariant; (3) the lines $y_1 = \text{constant}$ are mapped into each other by the flow. In other words, in the new coordinates we have a C^k differential equation of the form

$$\dot{y}_1 = a(y_1, \mu_1, \mu_2), \qquad \dot{y}_2 = y_2 b(y_1, y_2, \mu_1, \mu_2).$$

Since $p(\mu_2)$, defined in § 2, is C^k, we may assume that $p(\mu_2)$ is transformed to $(0, 0)$ for all μ_2. In other words, $a(0, 0, \mu_2) \equiv 0$. Since the stable manifold of $\dot{x} = f(x, 0, 0)$ at p is necessarily transformed into the line $y_1 = 0$, it is easy to arrange that

$$D_x y(p, 0, 0)u = (1, 0), \qquad D_x y(p, 0, 0)v = (0, 1). \tag{7}$$

Taking into account assumptions (i)-(iii) and (vi), we have

$$\begin{aligned} \dot{y}_1 &= \eta(\mu_2) y_1^2 (1 + y_1 g(y_1, \mu_2)) + \mu_1 h(y_1, \mu_1, \mu_2), \\ \dot{y}_2 &= -\lambda(y_1, \mu_1, \mu_2) y_2 (1 + y_2 k(y_1, y_2, \mu_1, \mu_2)), \end{aligned} \tag{8}$$

with $\eta > 0$, $h(0, 0, 0) > 0$, $\lambda(0, 0, 0) = \lambda$.

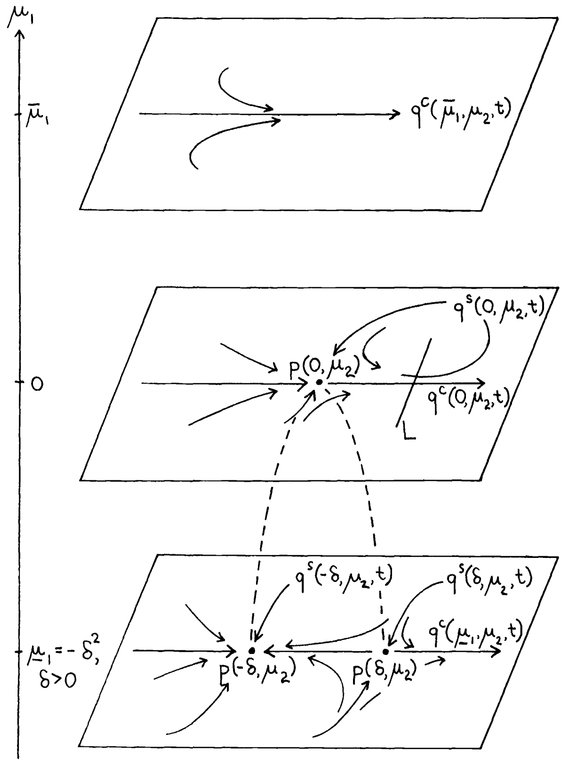


FIG. 3

LEMMA 1. *There is a C^{k-2} mapping $p(\delta, \mu_2)$, defined for (δ, μ_2) near $(0, 0)$, with values in $\mathbb{R}^2$, such that $p(0, 0) = p$, and*

$$p(\delta, \mu_2) \text{ is} \begin{cases} \text{a saddle-node of } \dot{x} = f(x, 0, \mu_2) & \text{if } \delta = 0, \\ \text{a saddle of } \dot{x} = f(x, -\delta^2, \mu_2) & \text{if } \delta > 0, \\ \text{a sink of } \dot{x} = f(x, -\delta^2, \mu_2) & \text{if } \delta < 0. \end{cases}$$

Moreover, the mapping $(\delta, \mu_2) \to (p(\delta, \mu_2), -\delta^2, \mu_2)$ is a C^{k-2} diffeomorphism of a neighborhood of $(0, 0)$ in $\mathbb{R}^2$ onto a neighborhood of $(p, 0, 0)$ in the set of equilibria of (5) *near* $(0, 0, 0)$ (see Fig. 3).

We remark that $p(0, \mu_2)$ equals $p(\mu_2)$ defined in § 2.

Proof. The equilibria of the system

$$\begin{aligned} \dot{y}_1 &= \eta(\mu_2) y_1^2 (1 + y_1 g(y_1, \mu_2)) + \mu_1 h(y_1, \mu_1, \mu_2), \\ \dot{y}_2 &= -\lambda(y_1, \mu_1, \mu_2) y_2 (1 + y_2 k(y_1, y_2, \mu_1, \mu_2)), \\ \dot{\mu}_1 &= 0, \\ \dot{\mu}_2 &= 0 \end{aligned}$$

near $(0, 0, 0, 0)$ comprise a set of the form

$$\{(y_1, 0, \mu_1, \mu_2) \colon \mu_1 = \mu_1(y_1, \mu_2)\},$$

where

$$\mu_1(0, \mu_2) = \frac{\partial \mu_1}{\partial y_1}(0, \mu_2) = 0 \quad \text{and} \quad \frac{\partial^2 \mu_1}{\partial y_1^2}(0, \mu_2) < 0.$$

Therefore,

$$\mu_1 = -y_1^2(A(\mu_2)+y_1B(y_1,\mu_2)),$$

where $A(\mu_2)>0$ and $A(\mu_2)+y_1B(y_1,\mu_2)$ is C^{k-2}. For $\mu_1 \leqq 0$, let $\mu_1 = -\delta^2$. Then

$$\delta = y_1(A(\mu_2)+y_1B(y_1,\mu_2))^{1/2}. \tag{9}$$

(Taking into account the negative square root gives no additional information.) By the implicit function theorem, we can solve (9) for y_1 near $y_1=0$, $\delta=0$, $\mu_2=0$. We obtain

$$y_1=\hat{p}(\delta,\mu_2) \quad \text{with } \hat{p}(0,\mu_2)=0, \qquad \frac{\partial \hat{p}}{\partial \delta}(0,\mu_2)>0. \tag{10}$$

Then $\eta>0$ and (10) imply that the equilibrium $(\hat{p}(\delta,\mu_2),0)$ of (8) with $\mu_1=-\delta^2$ is a saddle-node if $\delta=0$, a saddle if $\delta>0$, and a sink if $\delta<0$.

Let

$$x=x(y,\mu_1,\mu_2) \tag{11}$$

be the change of coordinates inverse to (6). Define

$$p(\delta,\mu_2)=x((\hat{p}(\delta,\mu_2),0),-\delta^2,\mu_2). \tag{12}$$

Then $p(\delta,\mu_2)$ satisfies the assertions of the lemma. □

For future use, we note that

$$\frac{\partial p}{\partial \delta}(0,0) \text{ is a positive multiple of } u. \tag{13}$$

To see this, we compute from (12)

$$\frac{\partial p}{\partial \delta}(0,0)=D_yx((0,0),0,0)\left(\frac{\partial \hat{p}}{\partial \delta}(0,0),0\right). \tag{14}$$

Since $(\partial\hat{p}/\partial\delta)(0,0)>0$ by (10), (14) is a positive multiple of u by (7).

System (8) with $\mu_1=-\delta^2$ has at the equilibrium $(\hat{p}(\delta,\mu_2),0)$ the invariant manifold $\{(y_1,y_2)\colon y_1=\hat{p}(\delta,\mu_2)\}$, a line. For $\delta=0$ this line is the stable manifold of the saddle-node $(0,0)$; for $\delta>0$ it is the stable manifold of the saddle $(\hat{p}(\delta,\mu_2),0)$; and for $\delta<0$ it is the strong stable manifold of the sink $(\hat{p}(\delta,\mu_2),0)$. These lines correspond to invariant manifolds of $\dot{x}=f(x,-\delta^2,\mu_2)$ at $p(\delta,\mu_2)$. Let $v(\delta,\mu_2)=D_yx((\hat{p}(\delta,\mu_2),0),-\delta^2,\mu_2)(0,1)$. Then $\dot{x}=f(x,-\delta^2,\mu_2)$ has at $p(\delta,\mu_2)$ an invariant curve tangent to $v(\delta,\mu_2)$ and these invariant curves vary in a C^{k-2} manner with (δ,μ_2). For $(\delta,\mu_2)=(0,0)$, this invariant curve contains Γ. Now a construction similar to that of $q^c(\mu_1,\mu_2,t)$ yields a C^{k-2} family

$$q^s(\delta,\mu_2,t), \qquad (\delta,\mu_2) \text{ small},$$

each a solution of $\dot{x}=f(x,-\delta^2,\mu_2)$, such that $q^s(\delta,\mu_2,t)\to p(\delta,\mu_2)$ as $t\to\infty$ along the negative $v(\delta,\mu_2)$ direction. Again we require $q^s(\delta,\mu_2,0)\in L$; thus $q^s(0,0,t)=q(t)$. Note that $q^s(0,\mu_2,t)$ is a branch of the stable manifold of the saddle-node $p(0,\mu_2)$ of $\dot{x}=f(x,0,\mu_2)$; and if $\delta>0$, $q^s(\delta,\mu_2,t)$ is a branch of the stable manifold of the saddle $p(\delta,\mu_2)$ of $\dot{x}=f(x,-\delta^2,\mu_2)$ (see Fig. 3).

For any vector $w=(w_1,w_2)\in\mathbb{R}^2$, let $w^\perp=(-w_2,w_1)$. Define $d^c(\mu_1,\mu_2)$ and $d^s(\delta,\mu_2)$ by

$$q^c(\mu_1,\mu_2,0)=q(0)+[d^c(\mu_1,\mu_2)/\|f(q(0),0,0)\|^2]f^\perp(q(0),0,0),$$

$$q^s(\delta,\mu_2,0)=q(0)+[d^s(\delta,\mu_2)/\|f(q(0),0,0)\|^2]f^\perp(q(0),0,0).$$

Then d^c is C^k and d^s is C^{k-2}. The number $d^c(\mu_1, \mu_2)$ (resp. $d^s(\delta, \mu_2)$) determines where on L the curve $q^c(\mu_1, \mu_2, t)$ (resp. $q^s(\delta, \mu_2, t)$) starts. We have

$$\begin{aligned} d^c(\mu_1, \mu_2) &= f^{\perp}(q(0), 0, 0) \cdot [q^c(\mu_1, \mu_2, 0) - q(0)] \\ &= f(q(0), 0, 0) \wedge [q^c(\mu_1, \mu_2, 0) - q(0)]. \end{aligned}$$

Similarly,

$$d^s(\delta, \mu_2) = f(q(0), 0, 0) \wedge [q^s(\delta, \mu_2, 0) - q(0)].$$

There is a separatrix loop of $\dot{x} = f(x, -\delta^2, \mu_2)$ through $p(\delta, \mu_2)$ if and only if $\delta \geqq 0$ and

$$d(\delta, \mu_2) \underset{\text{def}}{=} d^c(-\delta^2, \mu_2) - d^s(\delta, \mu_2) = 0. \tag{15}$$

Here $d(\delta, \mu_2)$ is C^{k-2}.

We shall show that $(\partial d/\partial \delta)(0, 0)$ is a negative multiple of $u \wedge v$ (hence is nonzero), and $(\partial d/\partial \mu_2)(0, 0) = I$. Given these facts, the proof of Theorem 1 is completed as follows: if $I \neq 0$, then $\{(\delta, \mu_2)\colon d(\delta, \mu_2) = 0\}$ is a C^{k-2} curve through $(0, 0)$ of the form

$$\delta = \ell \mu_2 + o(\mu_2), \qquad \ell = -I \Big/ \frac{\partial d}{\partial \delta}(0, 0). \tag{16}$$

Squaring both sides of (16) yields

$$\mu_1 = -\ell^2 \mu_2^2 + o(\mu_2^2).$$

The condition $\delta \geqq 0$, applied to (16), shows that $\mu_2 = 0$ or ℓ and μ_2 have the same sign. But ℓ has the sign of $I \cdot (u \wedge v)$.

We now turn to the computation of $(\partial d/\partial \delta)(0, 0)$ and $(\partial d/\partial \mu_2)(0, 0)$. We shall need the following variational equations for $q^c(-\delta^2, \mu_2, t)$ and $q^s(\delta, \mu_2, t)$:

$$\begin{aligned} &\frac{d}{dt} \frac{\partial q^c}{\partial \mu_2}(0, 0, t) = D_x f(q(t), 0, 0) \frac{\partial q^c}{\partial \mu_2}(0, 0, t) + \frac{\partial f}{\partial \mu_2}(q(t), 0, 0), \\ &\frac{d}{dt} \frac{\partial q^s}{\partial \delta}(0, 0, t) = D_x f(q(t), 0, 0) \frac{\partial q^s}{\partial \delta}(0, 0, t), \\ &\frac{d}{dt} \frac{\partial q^s}{\partial \mu_2}(0, 0, t) = D_x f(q(t), 0, 0) \frac{\partial q^s}{\partial \mu_2}(0, 0, t) + \frac{\partial f}{\partial \mu_2}(q(t), 0, 0). \end{aligned}$$

As in [2], we define

$$\Delta^c_{\mu_2}(t) = f(q(t), 0, 0) \wedge \frac{\partial q^c}{\partial \mu_2}(0, 0, t)$$

and define $\Delta^s_\delta(t)$ and $\Delta^s_{\mu_2}(t)$ analogously.

For $d^c(-\delta^2, \mu_2)$ and $d^s(\delta, \mu_2)$ we have the derivative formulas

$$\begin{aligned} &\frac{\partial d^c}{\partial \delta}(0, 0) = \frac{\partial d^c}{\partial \mu_1}(0, 0) \cdot \frac{d\mu_1}{d\delta}(0) = 0, \qquad \frac{\partial d^c}{\partial \mu_2}(0, 0) = \Delta^c_{\mu_2}(0), \\ &\frac{\partial d^s}{\partial \delta}(0, 0) = \Delta^s_\delta(0), \qquad \frac{\partial d^s}{\partial \mu_2}(0, 0) = \Delta^s_{\mu_2}(0). \end{aligned} \tag{17}$$

Using the variational equations for q^c and q^s, we compute as in [2]:

$$\frac{d}{dt} \Delta^c_{\mu_2}(t) = \operatorname{div} f(q(t), 0, 0) \Delta^c_{\mu_2}(t) + f(q(t), 0, 0) \wedge \frac{\partial f}{\partial \mu_2}(q(t), 0, 0), \tag{18}$$

$$\frac{d}{dt}\Delta^s_\delta(t)=\operatorname{div} f(q(t),0,0)\Delta^s_\delta(t), \tag{19}$$

$$\frac{d}{dt}\Delta^s_{\mu_2}(t)=\operatorname{div} f(q(t),0,0)\Delta^s_{\mu_2}(t)+f(q(t),0,0)\wedge\frac{\partial f}{\partial\mu_2}(q(t),0,0). \tag{20}$$

Solving these linear differential equations, we obtain, for any t_1,

$$\begin{aligned}\Delta^c_{\mu_2}(0)=\Delta^c_{\mu_2}(t_1)\exp\int_{t_1}^0 \operatorname{div} f(q(t),0,0)\,dt\\ +\int_{t_1}^0\exp\left[-\int_0^t \operatorname{div} f(q(s),0,0)\,ds\right]\\ \cdot f(q(t),0,0)\wedge\frac{\partial f}{\partial\mu_2}(q(t),0,0)\,dt,\end{aligned} \tag{21}$$

$$-\Delta^s_\delta(0)=-\Delta^s_\delta(t_1)\exp\left[-\int_0^{t_1}\operatorname{div} f(q(t),0,0)\,dt\right], \tag{22}$$

$$\begin{aligned}-\Delta^s_{\mu_2}(0)=-\Delta^s_{\mu_2}(t_1)\exp\left[-\int_0^{t_1}\operatorname{div} f(q(t),0,0)\,dt\right]\\ +\int_0^{t_1}\exp\left[-\int_0^t \operatorname{div} f(q(s),0,0)\,ds\right]\\ \cdot f(q(t),0,0)\wedge\frac{\partial f}{\partial\mu_2}(q(t),0,0)\,dt.\end{aligned} \tag{23}$$

We shall first evaluate (22) in the limit $t_1\to\infty$. Using the definition of Δ^s_δ, we write

$$-\Delta^s_\delta(0)=\frac{\partial q^s}{\partial\delta}(0,0,t_1)\wedge f(q(t_1),0,0)\exp\left[-\int_0^{t_1}\operatorname{div} f(q(s),0,0)\,ds\right]. \tag{24}$$

LEMMA 2.

$$\lim_{t\to\infty}\frac{\partial q^s}{\partial\delta}(0,0,t)=\frac{\partial p}{\partial\delta}(0,0),\qquad \lim_{t\to\infty}\frac{\partial q^s}{\partial\mu_2}(0,0,t)=\frac{\partial p}{\partial\mu_2}(0,0).$$

Proof. We shall use the coordinates (6). Define

$$\begin{aligned}\tilde q^s(\delta,\mu_2,t)&=y(q^s(\delta,\mu_2,t),-\delta^2,\mu_2)\\ &=(\hat p(\delta,\mu_2),y_2(\delta,\mu_2,t)),\end{aligned} \tag{25}$$

where $\hat p(\delta,\mu_2)$ is given by (10) and $y_2(\delta,\mu_2,t)$ is defined by (25). Since $q^s(\delta,\mu_2,t)\to p(\delta,\mu_2)$ as $t\to\infty$, for each (δ,μ_2) near $(0,0)$, $\tilde q^s(\delta,\mu_2,t)$ and hence $y_2(\delta,\mu_2,t)$ are defined for sufficiently large t. It follows from (7) that $y_2(\delta,\mu_2,t)>0$ for large t. From (8), $y_2(\delta,\mu_2,t)$ satisfies a differential equation of the form

$$\frac{dz}{dt}=-\lambda(\delta,\mu_2)z(1+zG(z,\delta,\mu_2)), \tag{26}$$

where $\lambda(0,0)=\lambda$. Here λ and zG are C^{k-3}.

In order to prove the lemma, we shall study the asymptotic behavior of solutions of (26) as $t\to\infty$ by solving (26) by separation of variables. Let

$$z^{-1}[1+zG(z,\delta,\mu_2)]^{-1}=z^{-1}+H(z,\delta,\mu_2). \tag{27}$$

Then H is C^{k-4}. Fix $z_0>0$. Let

$$J(z, \delta, \mu_2)=\int_{z_0}^{z} H(s, \delta, \mu_2)\, ds.$$

Then J is C^{k-4}. Solving (26) by separation of variables using (27) yields

$$\ln z+J(z, \delta, \mu_2)=-\lambda(\delta, \mu_2)t+A(\delta, \mu_2),$$

or

$$z \exp J(z, \delta, \mu_2)=B(\delta, \mu_2) \exp(-\lambda(\delta, \mu_2)t).$$

Here $B(\delta, \mu_2)$ is determined by the value of $y_2(\delta, \mu_2, t)$ at some $t=t_0$. Hence B is C^{k-4} and $B>0$. Since

$$\frac{\partial}{\partial z}[z \exp J(z, \delta, \mu_2)](0, \delta, \mu_2)\neq 0,$$

by the implicit function theorem we can solve the equation

$$z \exp J(z, \delta, \mu_2)=v$$

for z when z and v are near 0. We obtain

$$z=K(v, \delta, \mu_2),$$

where K is C^{k-4} and

$$K(0, \delta, \mu_2)\equiv 0. \tag{28}$$

Putting $z=y_2$ and $v=B(\delta, \mu_2)\exp(-\lambda(\delta, \mu_2)t)$, we obtain

$$y_2=K(B(\delta, \mu_2) \exp(-\lambda(\delta, \mu_2)t), \delta, \mu_2). \tag{29}$$

From (29) and (28) it follows that $(\partial y_2/\partial\delta)(\delta, \mu_2, t)$ and $(\partial y_2/\partial\mu_2)(\delta, \mu_2, t)$ approach 0 as $t\to\infty$. Therefore (25) implies that as $t\to\infty$,

$$\frac{\partial\tilde{q}^s}{\partial\delta}(\delta, \mu_2, t)\to\left(\frac{\partial\hat{p}}{\partial\delta}(\delta, \mu_2), 0\right), \qquad \frac{\partial\tilde{q}^s}{\partial\mu_2}(\delta, \mu_2, t)\to\left(\frac{\partial\hat{p}}{\partial\mu_2}(\delta, \mu_2), 0\right). \tag{30}$$

Now

$$\frac{\partial q^s}{\partial\delta}(\delta, \mu_2, t)=\frac{\partial}{\partial\delta}x(\tilde{q}^s(\delta, \mu_2, t), -\delta^2, \mu_2),$$

where $x(y, \mu_1, \mu_2)$ is given by (11). Therefore

$$\frac{\partial q^s}{\partial\delta}(0, 0, t)=D_yx(\tilde{q}^s(0, 0, t), 0, 0)\frac{\partial\tilde{q}^s}{\partial\delta}(0, 0, t).$$

By (30) and (14),

$$\lim_{t\to\infty}\frac{\partial q^s}{\partial\delta}(0, 0, t)=D_yx((0, 0), 0, 0)\left(\frac{\partial\hat{p}}{\partial\delta}(0, 0), 0\right)$$

$$=\frac{\partial p}{\partial\delta}(0, 0).$$

Similarly, the second formula of the lemma follows from (30) and the following formula derived from (12):

$$\frac{\partial p}{\partial\mu_2}(0, 0)=D_yx((0, 0), 0, 0)\left(\frac{\partial\hat{p}}{\partial\mu_2}(0, 0), 0\right)+\frac{\partial x}{\partial\mu_2}((0, 0), 0, 0). \qquad \square$$

We shall now use the computations we have done in proving Lemma 2 to study the other terms of (24). By (25),

$$\overset{*}{q}^s(\delta, \mu_2, t) = D_xy(q^s(\delta, \mu_2, t), -\delta^2, \mu_2)\dot{q}^s(\delta, \mu_2, t). \tag{31}$$

Let $\delta = \mu_2 = 0$ in (31). Since $\dot{q}^s(0, 0, t) = \dot{q}(t) = f(q(t), 0, 0)$, we obtain

$$f(q(t), 0, 0) = [D_xy(q(t), 0, 0)]^{-1}\overset{*}{q}^s(0, 0, t).$$

It follows easily from (25), (28) and (29) that

$$\begin{aligned} q(t) &= p + \mathcal{O}(\exp(-\lambda t)), \\ \overset{*}{q}^s(t) &= (0, -C\exp(-\lambda t) + o(\exp(-\lambda t))), \end{aligned} \tag{32}$$

where $C > 0$. Therefore, setting $t = t_1$,

$$\begin{aligned} f(q(t_1), 0, 0) = \{[D_xy(p, 0, 0)]^{-1} + \mathcal{O}(\exp(-\lambda t_1))\} \\ \cdot (0, -C\exp(-\lambda t_1) + o(\exp(-\lambda t_1))). \end{aligned} \tag{33}$$

From (32) we also have

$$\operatorname{div} f(q(t), 0, 0) = -\lambda + \mathcal{O}(\exp(-\lambda t)).$$

Therefore

$$\exp\left[-\int_0^{t_1} \operatorname{div} f(q(s), 0, 0)\, ds\right] = \exp(\lambda t_1) \cdot \exp\int_0^{t_1} \mathcal{O}(\exp(-\lambda s))\, ds. \tag{34}$$

Then (33) and (34) give

$$\begin{aligned} &\lim_{t_1\to\infty} f(q(t_1), 0, 0)\exp\left[-\int_0^{t_1} \operatorname{div} f(q(s), 0, 0)\, ds\right] \\ &\quad = [D_xy(p, 0, 0)]^{-1} \cdot \left(0, -C\exp\int_0^{\infty} \mathcal{O}(\exp(-\lambda s))\, ds\right), \end{aligned} \tag{35}$$

where the integral clearly converges.

Now (24), Lemma 2 and (35) imply that

$$-\Delta_\delta^s(0) = \frac{\partial p}{\partial\delta}(0, 0) \wedge \lim_{t_1\to\infty} f(q(t_1), 0, 0)\exp\left[-\int_0^{t_1} \operatorname{div} f(q(s), 0, 0)\, ds\right], \tag{36}$$

where the limit exists.

Notice that (7) implies that (35) is a negative multiple of v. Then (13) implies that $-\Delta_\delta^s(0)$ is a negative multiple of $u \wedge v$. By (15) and (17), $(\partial d/\partial\delta)(0, 0)$ is also a negative multiple of $u \wedge v$.

We now turn to (23). We claim that

$$\begin{aligned} -\Delta_{\mu_2}^s(0) = \frac{\partial p}{\partial\mu_2}(0, 0) \wedge \lim_{t_1\to\infty} f(q(t_1), 0, 0)\exp\left[-\int_0^{t_1} \operatorname{div} f(q(s), 0, 0)\, ds\right] \\ + \int_0^{\infty} \exp\left[-\int_0^{t} \operatorname{div} f(q(s), 0, 0)\, ds\right] f(q(t), 0, 0) \\ \wedge \frac{\partial f}{\partial\mu_2}(q(t), 0, 0)\, dt. \end{aligned} \tag{37}$$

The proof is modeled on that of (36). Using Lemma 2, we first show that the first summand of (23) approaches, as $t_1 \to \infty$, the first summand of (37), where the limit exists. Then, since $-\Delta^s_{\mu_2}(0)$ is finite, the second summand of (23), the integral, must approach a limit as $t_1 \to \infty$.

Finally we turn to (21).

LEMMA 3.

$$\lim_{t\to-\infty} \frac{\partial q^c}{\partial \mu_2}(0, 0, t) = 0.$$

Proof. Again we shall use the coordinates (6). Define

$$\tilde{q}^c(\mu_2, t) = y(q^c(0, \mu_2, t), 0, \mu_2) = (y_1(\mu_2, t), 0).$$

Since $q^c(0, \mu_2, t) \to p(0, \mu_2)$ as $t \to -\infty$, for each μ_2 near 0, $\tilde{q}^c(\mu_2, t)$, and hence $y_1(\mu_2, t)$, is defined for sufficiently negative t. From (7), $y_1(\mu_2, t) > 0$ for sufficiently negative t. From its definition, $y_1(\mu_2, t)$ satisfies a differential equation of the form

$$\frac{dz}{dt} = \eta(\mu_2) z^2(1 + zG(z, \mu_2)). \tag{38}$$

Here η and zG are C^{k-2}.

Let

$$z^{-2}(1 + zG(z, \mu_2))^{-1} = z^{-2} + A(\mu_2) z^{-1} + H(z, \mu_2). \tag{39}$$

Here A is C^{k-3} and H is C^{k-4}. Fix $z_0 > 0$. Let

$$J(z, \mu_2) = \int_{z_0}^{z} H(s, \mu_2)\, ds.$$

J is C^{k-4}. Then solving (38) by separation of variables using (39) yields

$$-z^{-1} + A(\mu_2) \ln z + J(z, \mu_2) = \eta(\mu_2) t + B(\mu_2).$$

Here $B(\mu_2)$ is determined by the value of $y_1(\mu_2, t)$ at some $t = t_0$. Hence B is C^{k-4}. Rearranging yields

$$z[1 - A(\mu_2) z \ln z - zJ(z, \mu_2)]^{-1} = -[\eta(\mu_2) t + B(\mu_2)]^{-1}. \tag{40}$$

Let $\Phi(z, \mu_2)$ equal the left-hand side of (40). $\Phi(z, \mu_2)$ is a C^1 function of z and μ_2 on a neighborhood of $(0, 0)$ in $\{(z, \mu_2): z \geqq 0\}$; $\Phi(0, \mu_2) \equiv 0$, and $(\partial/\partial z)\Phi(0, \mu_2) \equiv 1$. By the implicit function theorem, we can solve the equation $\Phi(z, \mu_2) = v$ for z when z and v are near 0, $v \geqq 0$ (in which case $z \geqq 0$). We obtain

$$z = v + R(v, \mu_2),$$

where R is C^1, $R(0, \mu_2) \equiv 0$ and R is $o(v)$. Putting $z = y_1$ and $v = -[\eta(\mu_2) t + B(\mu_2)]^{-1}$, t large negative, we obtain

$$y_1 = -[\eta(\mu_2) t + B(\mu_2)]^{-1} + R(-[\eta(\mu_2) t + B(\mu_2)]^{-1}, \mu_2).$$

It follows that $(\partial y_1/\partial \mu_2)(\mu_2, t) \to 0$ as $t \to -\infty$, so

$$\frac{\partial \tilde{q}^c}{\partial \mu_2}(\mu_2, t) \to 0 \quad \text{as } t \to -\infty. \tag{41}$$

Lemma 3 follows from (41) the way Lemma 2 follows from (30). □

To evaluate (21) in the limit $t_1 \to -\infty$, we use the definition of $\Delta^c_{\mu_2}$ to write (21) as

$$\begin{aligned}\Delta^c_{\mu_2}(0) &= -\frac{\partial q^c}{\partial \mu_2}(0,0,t_1) \wedge f(q(t),0,0) \exp \int_{t_1}^0 \operatorname{div} f(q(s),0,0)\,ds \\ &\quad + \int_{t_1}^0 \exp\left[-\int_0^t \operatorname{div} f(q(s),0,0)\,ds\right] f(q(t),0,0) \wedge \frac{\partial f}{\partial \mu_2}(q(t),0,0)\,dt.\end{aligned} \tag{42}$$

Of course

$$f(q(t),0,0) \to 0 \quad \text{as } t_1 \to -\infty. \tag{43}$$

Moreover,

$$\operatorname{div} f(q(t),0,0) = -\lambda + \mathcal{O}([\eta(0)t + c(0)]^{-1}) \quad \text{as } t \to -\infty. \tag{44}$$

By Lemma 3, (43) and (44), we have

$$\lim_{t_1 \to -\infty} -\frac{\partial q^c}{\partial \mu_2}(0,0,t_1) \wedge f(q(t_1),0,0) \exp \int_{t_1}^0 \operatorname{div} f(q(s),0,0)\,ds = 0.$$

Therefore the second summand of (42), the integral, approaches a limit as $t_1 \to -\infty$, so

$$\Delta^c_{\mu_2}(0) = \int_{-\infty}^0 \exp\left[-\int_0^t \operatorname{div} f(q(s),0,0)\,ds\right] f(q(t),0,0) \wedge \frac{\partial f}{\partial \mu_2}(q(t),0,0)\,dt. \tag{45}$$

Finally, we complete the proof of Theorem 1 by calculating

$$\frac{\partial d}{\partial \mu_2}(0,0) = \frac{\partial d^c}{\partial \mu_2}(0,0) - \frac{\partial d^s}{\partial \mu_2}(0,0) = \Delta^c_{\mu_2}(0) - \Delta^s_{\mu_2}(0) = (45) + (37) = I.$$

4. Proof of Theorem 2. First we show that Σ is a C^k codimension one submanifold of Σ'. Let $f \in \Sigma$ with saddle-node p and let L be a line segment perpendicular to the separatrix loop Γ as in § 3. For g near f in Σ', there is a unique saddle-node p_g near p; p_g is a C^k function of $g \in \Sigma'$. The stable and center manifolds of p_g also depend C^k on g [1, § 9.2]. Thus their intersections with L are C^k functions of g. Therefore the function $d(0,\mu_2)$ from § 2 extends to a C^k function $d(g)$ defined for $g \in \Sigma'$ near f; $d(g)$ measures the separation of these points of intersection. (We remark that $d(0,\mu_2)$ is C^k although $d(\delta,\mu_2)$ is only C^{k-2}.) Then $d(g)=0$ if and only if $g \in \Sigma$. Since it is easy to find a perturbation $f+\varepsilon h$ in Σ' such that $d/d\varepsilon|_{\varepsilon=0} d(f+\varepsilon h) \neq 0$, Σ is a C^k codimension one submanifold of Σ'.

To prove Theorem 2, it suffices to show that $f(\cdot,\mu_1,\mu_2)$ is transverse to Σ at $(\mu_1,\mu_2)=(0,0)$ if and only if $I \neq 0$. Since $f(\cdot,0,\mu_2) \in \Sigma'$ for all small μ_2, $(\partial f/\partial \mu_2)(\cdot,0,0)$ is tangent to Σ' (see Fig. 4). But $(\partial \tilde{f}/\partial \nu_1)(\cdot,0,0) = (\partial f/\partial \mu_1)(\cdot,0,0) - (\partial f/\partial \mu_2)(\cdot,0,0)\alpha'(0)$. Since $(\partial \tilde{f}/\partial \nu_1)(\cdot,0,0)$ is transverse to Σ' by

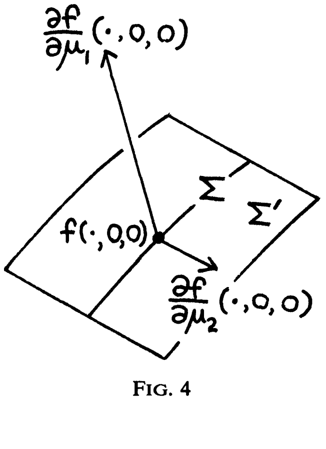

$\frac{\partial f}{\partial \mu_1}(\cdot,0,0)$
Σ
Σ'
$f(\cdot,0,0)$
$\frac{\partial f}{\partial \mu_2}(\cdot,0,0)$

FIG. 4

assumption (vi) (see [7]), and $(\partial f/\partial\mu_2)(\cdot,0,0)$ is tangent to Σ', $(\partial f/\partial\mu_1)(\cdot,0,0)$ is transverse to Σ'. Thus we need to show that $(\partial f/\partial\mu_2)(\cdot,0,0)$ is transverse to Σ if and only if $I\neq 0$. But this follows from the formula $(\partial d/\partial\mu_2)(0,0)=I$.

5. The pendulum equation. We consider the differential equation for the damped pendulum with constant applied torque, in the dimensionless form studied in [5]:

$$\beta\ddot{\phi}+\dot{\phi}+\sin\phi=\rho.$$

Putting $y=\dot{\phi}$, we have

$$\dot{\phi}=y,\qquad \dot{y}=\frac{1}{\beta}(-y-\sin\phi+\rho). \tag{46}$$

We identitify ϕ and $\phi+2\pi$, so that (46) defines a vector field on a cylinder; ρ and β are parameters, which we shall assume positive. We remark that Theorem 1 applies equally well to vector fields on the compact set $\{(\phi,y)\colon |y|\leqq d\}$.

Let $x=(\phi,y)$,

$$f(x,\beta,\rho)=(f_1((\phi,y),\beta,\rho),f_2((\phi,y),\beta,\rho))=\left(y,\frac{1}{\beta}(-y-\sin\phi+\rho)\right).$$

If $\rho<1$, $\dot{x}=f(x,\rho,\beta)$ has two equilibria, one a saddle and one a sink; $\dot{x}=f(x,1,\beta)$ has one equilibrium, at $(\pi/2,0)$ independent of β; if $\rho>1$, $\dot{x}=f(x,\rho,\beta)$ has no equilibria. We note that

$$D_xf\left(\left(\frac{\pi}{2},0\right),1,\beta\right)=\begin{bmatrix}0 & 1\\ 0 & -1/\beta\end{bmatrix},$$

which has eigenvalues $0,-1/\beta$. Corresponding right eigenvectors are $u=(1,0)$ and $v=(-\beta,1)$; a left eigenvector for the eigenvalue 0 is $(1,\beta)$.

To show that $\dot{x}=f(x,1,\beta)$ has a saddle-node at $(\pi/2,0)$, we compute (assumption (iii)):

$$\begin{aligned}(1,\beta)\cdot D_x^2f\left(\left(\frac{\pi}{2},0\right),1,\beta\right)\cdot((1,0),(1,0))&=(1,\beta)\cdot\left(0,\frac{\partial^2 f}{\partial\phi^2}\left(\left(\frac{\pi}{2},0\right),1,\beta\right)\right)\\&=(1,\beta)\cdot\left(0,\frac{1}{\beta}\right)=1.\end{aligned}$$

We also note (assumption (iv)):

$$(1,\beta)\cdot D_\rho f\left(\left(\frac{\pi}{2},0\right),1,\beta\right)=(1,\beta)\cdot\left(0,\frac{1}{\beta}\right)=1.$$

It is shown in [3] that there is a unique positive β_0 such that $\dot{x}=f(x,1,\beta_0)$ has a separatrix loop at the saddle-node $(\pi/2,0)$. The separatrix loop, considered as a curve in ϕy-space with ϕ and $\phi+2\pi$ not yet identified, can be expressed as $y=y(\phi)$, $\pi/2<\phi<5\pi/2$; $y(\phi)>0$ for all ϕ. As $\phi\to\pi/2$, $y\to 0$ and $y/(\phi-\pi/2)\to 0$; as $\phi\to 5\pi/2$, $y\to 0$ and $y/(\phi-5\pi/2)\to -1/\beta)$ (see Fig. 5).

Thus assumptions (i)-(vi) are verified if we put $\tilde{f}=f$, $\nu_1=\rho-1$, $\nu_2=\beta-\beta_0$. The change of variables from ν to μ is not necessary here, i.e., we may put $\mu_1=\nu_1$, $\mu_2=\nu_2$. For simplicity we shall continue to use the parameters ρ and β.

To compute I, we first note that the saddle-node $p(\beta)$ of $\dot{x}=f(x,1,\beta)$ is identically $(\pi/2,0)$, so that the first summand of I is 0. Now $\operatorname{div} f(x,\rho,\beta)=-1/\beta$, and

$$f((\phi,y),\rho,\beta)\wedge\frac{\partial f}{\partial\beta}((\phi,y),\rho,\beta)=-\frac{1}{\beta^2}y(-y-\sin\phi+\beta)=-\frac{1}{\beta}y\dot{y}.$$

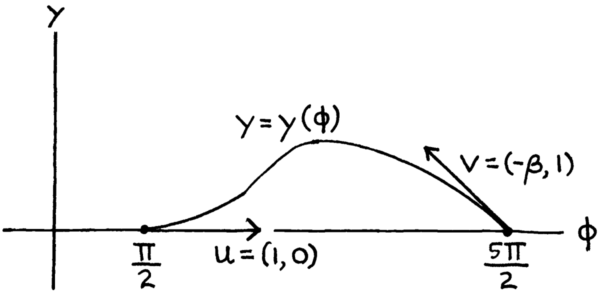


FIG. 5

Therefore,

$$I=-\frac{1}{\beta_0}\int_{-\infty}^{\infty} e^{t/\beta_0}y\dot{y}\,dt$$

$$=\lim_{S,T\to\infty}\left\{-\frac{1}{\beta_0}e^{t/\beta_0}\cdot\frac{y^2}{2}\bigg|_{-S}^{T}+\frac{1}{\beta_0^2}\int_{-S}^{T}e^{t/\beta_0}\cdot\frac{y^2}{2}\,dt\right\}.$$

From (29), as $t\to\infty$, $(\phi(t)-(5\pi/2), y(t))=c\,e^{-t/\beta_0}(-\beta_0, 1)+o(e^{-t/\beta_0})$ for some positive constant c. Therefore $[y(t)]^2=c^2\,e^{-2t/\beta_0}+o(e^{-2t/\beta_0})$ as $t\to\infty$. Hence

$$\lim_{T\to\infty}-\frac{1}{\beta_0}e^{T/\beta_0}\cdot\frac{[y(T)]^2}{2}=0.$$

Of course,

$$\lim_{S\to\infty}-\frac{1}{\beta_0}e^{-S/\beta_0}\cdot\frac{[y(-S)]^2}{2}=0,$$

since $y(t)\to 0$ as $t\to-\infty$. Therefore

$$I=\frac{1}{\beta_0^2}\int_{-\infty}^{\infty}e^{t/\beta_0}\cdot\frac{y^2}{2}\,dt>0.$$

Since $u\wedge v>0$, Theorem 1 implies that for (ρ,β) near $(1,\beta_0)$, $\dot{x}=f(x,\rho,\beta)$ has a separatrix loop if and only if $\rho=1-\ell^2(\beta-\beta_0)^2+\cdots$, with $\ell\neq 0$, and $\beta-\beta_0\geqq 0$ (see Fig. 6). This result is in agreement with statements in [5].

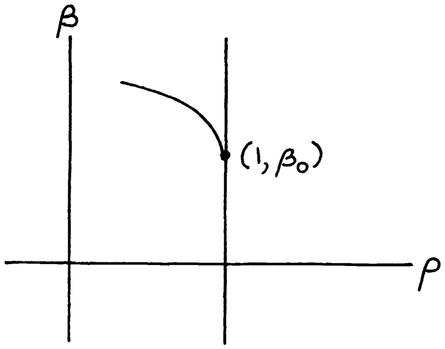


FIG. 6

REFERENCES

[1] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[2] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[3] M. Levi, F. C. Hoppensteadt and W. L. Miranker, *Dynamics of the Josephson junction*, Quart. Appl. Math., 36 (1978), pp. 167-198.

[4] V. I. Luk'yanov, *Bifurcations of dynamical systems with a saddle-point-separatrix loop*, Differentsial'nye Uravneniya, 18 (1982), pp. 1493-1506 (J. Differential Equations, 18 (1982), pp. 1049-1059).

[5] F. M. A. Salam and S. S. Sastry, *Dynamics of the forced Josephson junction: the regions of chaos*, IEEE Trans. Circuits and Systems, CAS-32 (1985), pp. 784-796.

[6] S. Schecter, *Melnikov's method at a saddle-node and the dynamics of the forced Josephson junction*, this Journal, to appear.

[7] J. Sotomayor, *Generic one-parameter families of vector fields*, Inst. Hautes Etudes Sci. Publ. Math., 43 (1974), pp. 5-46.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

017

NONEXISTENCE FOR THE KASSOY PROBLEM*

J. BEBERNES† AND W. TROY‡

Abstract. We prove that the Kassoy problem

$$y''-\frac{x}{2}y'+e^y-1=0, \quad y(0)=\alpha, \quad y'(0)=0, \quad \alpha\in\mathbb{R} \tag{1}$$

has no solution which has the asymptotic property

$$y(x)\sim -2\ln x+K \quad \text{as } x\to\infty. \tag{2}$$

Key words. combustion, supercritical, final time asymptotic analysis, nonexistence

AMS(MOS) subject classification. 34A34

1. Introduction. The purpose of this paper is to rigorously prove that a solution to the initial value problem

$$\begin{gathered} y''-\frac{x}{2}y'+e^y-1=0, \\ y(0)=\alpha, \quad y'(0)=0, \quad \alpha\in\mathbb{R} \end{gathered} \tag{1}$$

cannot have the asymptotic property

$$y(x)\sim -2\ln x+K \quad \text{as } x\to\infty. \tag{2}$$

This problem, hereafter referred to as the Kassoy problem, has an interesting history.

The nondimensional induction period equation for a high activation energy thermal explosion in a bounded container Ω can be written in the form

$$\theta_t-\Delta\theta=\delta\, e^{\theta} \tag{3}$$

where $\theta(x, t)$ must satisfy the initial-boundary conditions

$$\begin{gathered} \theta(x,0)=\psi(x), \qquad x\in\Omega, \\ \theta(x,t)=0, \qquad x\in\partial\Omega, \quad t>0. \end{gathered} \tag{4}$$

The dependent variable $\theta(x, t)$ can be interpreted as a perturbation temperature, describing deviations from a prescribed initial state. The temperature variation is driven by the heat release term $\delta\, e^{\theta}$, where $\delta>0$ is the Frank-Kamenetski parameter.

For Ω a radially symmetric container centered at $x=0$ and $\psi(x)\equiv 0$, Bebernes and Kassoy [1] and Lacey [8] proved that there exists a critical value $\delta_{\mathrm{CR}}>0$, which depends on the geometry of the system such that for $\delta>\delta_{\mathrm{CR}}$ the solution $\theta(x, t)$ becomes unbounded at some $x\in\Omega$ as t approaches a finite time $t_e(\delta)$, and blowup occurs. Recently, Friedman and McLeod [4] proved that this blowup occurs at a single point $x_0=0\in\Omega$ as $t\to t_e$.

* Received by the editors July 1, 1985; accepted for publication (in revised form) July 30, 1986.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309. The work of this author was supported by U.S. Army Research Office contract DAAG 29-82-K-0069.

‡ Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260. The work of this author was supported by National Science Foundation grant MCS 8301085.

These supercritical processes, which are characterized by the appearance of a singularity somewhere in the spatial domain Ω at a finite time t_e, were considered earlier by Kassoy and Poland [6], [7] and Kapila [5]. Using computational methods for symmetric slab, cylindrical, and spherical geometries, they predicted that $\theta(x, t)$ becomes unbounded at the symmetry point $x = 0$ at a time t_e. Elsewhere $\theta(x, t) \to \theta_e(x)$, $x \neq 0$, $x \in \Omega$ as $t \to t_e$. By applying a final-value asymptotic analysis at t_e, they also studied the character of the singularity function $\theta_e(x)$. In their analysis of this singularity function $\theta_e(x)$ in the case of a slab, they numerically predicted the existence of a solution $y(x)$ of the Kassoy problem.

Lacey [9] and Dold [3] have questioned the existence of a solution to the Kassoy problem. In fact, Dold rejects the formulation given in [6] and [5] on numerical grounds and generates a new final value theory which leads to a different description of the singular solution. In this paper we resolve the issue by proving the following theorem.

THEOREM. *There is no solution of problem* (1) *which satisfies condition* (2).

The proof of nonexistence is given in the next three sections. In § 2, for $\alpha < 0$ we prove the existence of a special solution to the IVP(1) assuming the Kassoy problem has a solution. This special solution has several precisely describable properties. In § 3, we prove that our special solution cannot have one of these properties. This contradiction leads us to the conclusion that no solution to the Kassoy problem can exist for $\alpha > 0$. In § 4, we prove that the Kassoy problem has no solution for $\alpha < 0$.

2. A special solution for $\alpha > 0$.

We begin this section by showing some needed properties of solutions of IVP(1). In particular, we first show that for α sufficiently large, the solution $y(x, \alpha)$ of IVP(1) is strictly concave down for $x \geqq 0$.

LEMMA 1. *If* $\alpha \geqq 1$, *then* $y''(x, \alpha) < 0$ *for all* $x \geqq 0$.

Proof. A solution $y(x) = y(x, \alpha)$ of (1) satisfies $y(0) = \alpha$, $y'(0) = 0$, $y''(0) < 0$. From this and (1), it follows that $y''(x) < 0$ as long as $y(x) \geqq 0$. Hence $y(x)$ has a first zero $a > 0$ with $y'(a) < 0$ and $y''(a) < 0$.

If $y''(x_1) = 0$ for some first $x_1 > a$, then $y(x_1) < 0$ and $y'''(x_1) \geqq 0$. Let $g(x) = xy'(x) + 1$. Then $g(0) = 1$ and $g' < y'$ on $(0, x_1)$. By integrating, we have $g(x) < 1 + y(x) - \alpha$ on $(0, x_1)$. Thus $g(x_1) \leqq y(x_1) < 0$; hence $y'(x_1) < -1/x_1$. From this and (1) we conclude that

$$0 = y''(x_1) = \frac{x_1}{2} y'(x_1) + 1 - e^{y(x_1)}$$

$$< \frac{1}{2} - e^{y(x_1)}.$$

Finally, a differentiation of (1) leads to $y'''(x_1) = (\frac{1}{2} - e^{y(x_1)}) y'(x_1) < 0$, a contradiction. Hence, $y''(x) < 0$ for all $x > 0$.

COROLLARY. *If a solution* $y(x, \alpha)$ *of* (1), (2) *exists for* $\alpha > 0$, *then* $0 < \alpha < 1$.

LEMMA 2. *If a solution* $y(x, \alpha)$ *of* (1), (2) *exists, then there exists a first* x_1 *such that*

$$y''(x, \alpha) < 0 \quad \text{on } (0, x_1), \quad y''(x_1, \alpha) = 0, \quad y'''(x_1, \alpha) \geqq 0, \quad y(x_1, \alpha) \geqq -\ln 2.$$

Proof. The solution $y(x) = y(x, \alpha)$ of (1), (2) satisfies $y(0) = \alpha$, $y'(0) = 0$, and $y''(0) = 1 - e^{\alpha} < 0$. This implies $y'(x)$ is negative in a right neighborhood of 0, and, because $y(x)$ satisfies (2), $y'(x) \to 0$ as $x \to \infty$. Hence there exists a first x_1 such that $y''(x) < 0$ on $(0, x_1)$, $y'(x_1) < 0$, $y''(x_1) = 0$, $y'''(x_1) \geqq 0$. Since $y'''(x_1) = (\frac{1}{2} - e^{y(x_1)}) y'(x_1) \geqq 0$ and $y'(x_1) < 0$, we have $y(x_1) \geqq -\ln 2$.

LEMMA 3. *If $y(x, \alpha)$ is a solution of* (1) *with $y''(x, \alpha) < 0$ on $[0, x_2]$ and $y(x_2, \alpha) \leqq -\ln 2$. Then $y''(x) < 0$ for all $x \geqq 0$.*

Proof. At x_2, $y'(x_2, \alpha) < 0$ and $y''(x_2, \alpha) < 0$. If there is a first $x_3 > x_2$ for which $y''(x_3, \alpha) = 0$, then $y'''(x, \alpha) \geqq 0$. However, differentiating (1), we find that $y'''(x_3) = y'(x_3)(\frac{1}{2} - e^{y(x_3)}) < 0$, a contradiction. From this we conclude that $y'' < 0$ for all $x \geqq 0$.

We are now in a position to define our special solution under the following assumption.

Assumption. There exists $\beta \in (0, 1)$ such that $y(x, \beta)$ is a solution of (1), (2).

In order to make use of this assumption we define the set

$$A = \{\hat{\alpha} > 0 \,|\, \text{if } \alpha > \hat{\alpha}, \text{then } y''(x, \alpha) < 0 \text{ for all } x \geqq 0\}.$$

From Lemma 1 we observe that $[1, \infty) \subseteq A$. Furthermore, continuity of solutions with respect to initial conditions, together with Lemma 3, shows that A is open. Finally, it is clear that A is bounded below by β. Thus the value

$$\bar{y} \equiv \inf A$$

is well defined.

Hereafter we let $y(x) = y(x, \bar{y})$. In the next lemma we list four properties which $y(x)$ must satisfy. Our goal in the remainder of the paper is to show that $y(x)$ cannot satisfy property (iv) of Lemma 4. From this contradiction we conclude that $y(x)$ cannot exist, which in turn implies that the original problem (1), (2) has no solution.

LEMMA 4. *There is a first $\bar{x} > 0$ such that*

(i) $y(\bar{x}) = -\ln(2)$,

(ii) $y''(x) < 0$ *for* $0 < x < \bar{x}$,

(iii) $y''(\bar{x}) = y'''(\bar{x}) = 0$,

(iv) $y'(\bar{x}) = -\bar{x}^{-1}$.

Proof. It follows from (1) that $y'' < 0$ for $x > 0$ as long as $y \geqq 0$. As shown in the proof of Lemma 1 there is a first value, $x = a$, for which $y(a) = 0$ with $y'(a) < 0$ and $y''(a) < 0$. Suppose that $y'' < 0$ for all $x > a$. Then $\bar{y} \in A$. It follows from continuity and Lemma 3 that $\alpha \in A$ if $\bar{y} - \alpha > 0$ is sufficiently small, contradicting the definition of $\bar{y}$. Thus, there exists a first $\bar{x} > a$ such that $y''(\bar{x}) = 0$, and $y'''(\bar{x}) \geqq 0$. If $y(\bar{x}) < -\ln(2)$, then $y'''(\bar{x}) = (\frac{1}{2} - e^{y(\bar{x})})y'(\bar{x}) < 0$, a contradiction. If $y(\bar{x}) > -\ln 2$, then $y'''(\bar{x}) > 0$.

It then follows from continuity that $\alpha \notin A$ if $\alpha - \bar{y} > 0$ is sufficiently small, again contradicting the definition of $\bar{y}$. Therefore, it must be the case that $y(\bar{x}) = -\ln(2)$ and $y'''(\bar{x}) = 0$. Finally, from these observations and (1) we conclude that $y'(\bar{x}) = -\bar{x}^{-1}$, and the lemma is proved.

Further technical properties of the solution $y(x, \bar{y}) = y(x)$ are given in the following lemma.

LEMMA 5. *The solution $y(x)$ has initial value $\bar{y} \leqq 1 - \ln 2$. Furthermore, $y''' \geqq 0$ for all $x \in [0, \bar{x}]$.*

Proof. Suppose that $\bar{y} > 1 - \ln(2)$ and recall the function $g(x) = xy'(x) + 1$. Then $g(0) = 1$ and $g'(x) < y'(x)$ on $(0, \bar{x})$, where $\bar{x}$ is the $\bar{x}$ of Lemma 4. An integration leads to $g(\bar{x}) < \ln(2) + y(\bar{x}) < 0$. Thus $\bar{x}y'(\bar{x}) + 1 < 0$, contradicting (iv) of Lemma 4. We conclude that $\bar{y} \leqq 1 - \ln(2)$.

Next, recall that $y(a) = 0$. If $a \geqq \sqrt{2}$, then $y'''(a) \leqq 0$. Also, from (1) we obtain

$$y^{(4)} = \frac{x}{2} y''' + (1 - e^y)y'' - e^y(y')^2 \tag{5}$$

and it follows that $y''' < 0$ for all $x > a$, contradicting Lemma 4. Thus $a \in (0, \sqrt{2})$ and $y'''(a) = ((a^2/4) - \frac{1}{2})y'(a) > 0$. Furthermore, Lemma 4 shows that $y'''(\bar{x}) = 0$. We need to show that y''' is nonnegative on each of the intervals $(0, a)$ and $(a, \bar{x})$. Suppose, first of all, that $y''' < 0$ at some point in $(a, \bar{x})$. Thus, since $y'''(\bar{x}) = 0$ there must be a value $x^* \in (a, \bar{x}]$ such that $y'''(x^*) = 0$ and $y^{(4)}(x^*) \geqq 0$. However, from (5) it follows that $y^{(4)}(x^*) < 0$, a contradiction. Therefore $y''' \geqq 0$ for all $x \in (a, \bar{x}]$. Next, suppose that $y''' < 0$ at some point in $(0, a)$. From (1) and (5) it follows that $y'''(0) = 0$ and $y^{(4)}(0) > 0$. Thus y''' must reach a negative minimum at some value $\hat{x} \in (0, a)$. That is, $y'''(\hat{x}) < 0$, $y^{(4)}(\hat{x}) = 0$ and $y^{(5)}(\hat{x}) \geqq 0$. A differentiation of (5) leads to

$$y^{(5)} = \frac{x}{2} y^{(4)} + y'''\left(\frac{3}{2} - e^y\right) - 3y'y''\,e^y - (y')^3\,e^y. \tag{6}$$

At $\hat{x}$, since $y^{(4)}(\hat{x}) = 0$, we have

$$0 = \frac{\hat{x}}{2} y''' + y''(1 - e^y) - e^y (y')^2.$$

Thus $-e^y(y')^3 = y'y''(e^y - 1) - (\hat{x}/2)y'y'''$. Substituting this into (6), and noting that $3/2 - e^y > 0$ for $y(\hat{x}) < \bar{y} \leqq 1 - \ln(2)$, we obtain, at $x = \hat{x}$, $y^{(5)}(\hat{x}) = y'''(3/2 - e^y) - 2y'y''e^y - y'y'' - (\hat{x}/2)y'y''' < 0$, a contradiction. We conclude that $y''' \geqq 0$ on $(0, a]$ which completes the proof of the lemma.

COROLLARY. *On* $(0, \bar{x})$, $y(x) \geqq p(x) \equiv y''(0)(x^2/2) + \bar{y}$.

Define $\tilde{x}(\bar{y})$ by $p(\tilde{x}(\bar{y})) = -\ln 2$; then $\tilde{x}(\bar{y}) = [2(\ln 2 + \bar{y})/(e^{\bar{y}} - 1)]^{1/2} < \bar{x}$.

3. Nonexistence for $\alpha > 0$. To prove our theorem for the case $\alpha > 0$ it suffices to eliminate the existence of the solution satisfying Lemma 4 (recall that such a solution must exist if the Kassoy problem (1), (2) has a solution). In particular we show that the solution described in Lemma 4 satisfies

$$y'(\tilde{x}) < -\tilde{x}^{-1} \tag{7}$$

where $\tilde{x} = \tilde{x}(\bar{y}) \equiv [2(\ln(2) + \bar{y})/(e^{\bar{y}} - 1)]^{1/2}$. Since $\tilde{x} < \bar{x}$ and $y'' < 0$ on $[0, \bar{x}]$, then (7) implies that $y'(\bar{x}) < y'(\tilde{x}) < -1/\tilde{x} < -1/\bar{x}$. Hence, property (iv) of Lemma 4 is violated and a contradiction is reached. The next three technical lemmas are necessary to show that (7) holds. We introduce the comparison function

$$G(x, \bar{y}) \equiv x\,e^{x^2/4} \int_0^x \left(1 - e^{\bar{y}}\,e^{(1 - e^{\bar{y}})s^2/2}\right) e^{-s^2/4}\,ds + 1. \tag{8}$$

In the next lemma we show that $g(x) \leqq G(x, \bar{y})$ for $0 \leqq x \leqq \bar{x}$. Subsequently, we prove that $G(\tilde{x}(\bar{y}), \bar{y}) < 0$ for all $\bar{y} \in (0, 1 - \ln(2)]$. Thus it follows that $g(\tilde{x}(\bar{y})) < 0$ for all $\bar{y} \in (0, 1 - \ln(2)]$ and (7) is proved.

LEMMA 6. $g(x) \leqq G(x, \bar{y})$ *for* $0 \leqq x \leqq \bar{x}$.

Proof. From Lemma 5, $y''' \geqq 0$ on $[0, \bar{x}]$, which in turn implies that $y(x) \geqq y''(0)(x^2/2) + \bar{y} = \bar{y} + ((1 - e^{\bar{y}})/2)x^2$. Thus $e^{y(x)} \geqq e^{\bar{y}}\,e^{(1 - e^{\bar{y}})x^2/2}$ on $[0, \bar{x}]$. Since $g(x) = xy'(x) + 1 = x\,e^{x^2/4} \int_0^x (1 - e^{y(s)})\,e^{-s^2/4}\,ds + 1$ it follows immediately that $g(x) \leqq G(x, \bar{y})$ on $[0, \bar{x}]$.

LEMMA 7. $\lim_{\bar{y} \to 0^+} G(\tilde{x}(\bar{y}), \bar{y}) = -\infty$.

Proof. We observe that

$$\begin{aligned} G(\tilde{x}(\bar{y}), \bar{y}) &= \tilde{x}\,e^{\tilde{x}^2/4} \int_0^{\tilde{x}} e^{-s^2/4}\,ds - \tilde{x}\,e^{\tilde{x}^2/4} \int_0^{\tilde{x}} e^{\bar{y}}\,e^{(1 - 2e^{\bar{y}})s^2/4}\,ds + 1 \\ &= \tilde{x}\,e^{\tilde{x}^2/4} \int_0^{\tilde{x}} e^{-s^2/4}\,ds - \frac{\tilde{x}\,e^{\tilde{x}^2/4}\,e^{\bar{y}}}{(2\,e^{\bar{y}} - 1)^{1/2}} \int_0^{\tilde{x}(2e^{\bar{y}} - 1)^{1/2}} e^{-s^2/4}\,ds + 1 \\ &= \tilde{x}\,e^{\tilde{x}^2/4}\left(1 - \frac{e^{\bar{y}}}{(2\,e^{\bar{y}} - 1)^{1/2}}\right) \int_0^{\tilde{x}} e^{-s^2/4}\,ds + 1. \end{aligned}$$

Since $\lim_{\bar{y}\to 0}\int_0^{\tilde{x}(y)} e^{-s^2/4}\,ds = \pi^{1/2}$, it suffices to consider

$$H(\bar{y}) \equiv \frac{((2e^{\bar{y}}-1)^{1/2}-e^{\bar{y}})}{\tilde{x}^{-1}\,e^{-\tilde{x}^2/4}}\cdot\frac{1}{(2\,e^{\bar{y}}-1)^{1/2}}$$

$$=\frac{-(e^{\bar{y}}-1)^2}{\tilde{x}^{-1}\,e^{-\tilde{x}^2/4}}\cdot\frac{1}{((2\,e^{\bar{y}}-1)^{1/2}+e^{\bar{y}})}\cdot\frac{1}{(2\,e^{\bar{y}}-1)^{1/2}}.$$

We set $h(\bar{y}) = -(e^{\bar{y}} = 1)^2/(\tilde{x}^{-1}\,e^{-\tilde{x}^2/4})$ and use the definition of $\tilde{x}(\bar{y})$ to obtain

$$h(\bar{y}) = -4\,e^{\tilde{x}^2/4}(\ln(2)+\bar{y})^2/\tilde{x}^3.$$

Since $\tilde{x}(\bar{y})\to\infty$ as $\bar{y}\to 0^+$ it follows that $\lim_{\bar{y}\to 0^+} h(\bar{y}) = -\infty$. Therefore we conclude that

$$\lim_{\bar{y}\to 0^+} H(\bar{y}) = \lim_{\bar{y}\to 0^+} G(\tilde{x}(\bar{y}),\bar{y}) = -\infty$$

and the lemma is proved.

LEMMA 8. *$G(\tilde{x}(\bar{y}),\bar{y})<0$ for all $\bar{y}\in(0, 1-\ln(2)]$.*

Proof. An integration by parts shows that

$$x = e^{x^2/4}\int_0^x e^{-s^2/4}(1-s^2/2)\,ds \quad \text{for all } x>0. \tag{9}$$

If there is a first $\hat{y}\in(0, 1-\ln(2))$ for which $G(\tilde{x}(\hat{y}),\hat{y}) = 0$, then

$$\frac{dG}{d\bar{y}}(\tilde{x}(\bar{y}),\bar{y})|_{\bar{y}=\hat{y}} \geqq 0. \tag{10}$$

Our goal is to obtain a contradiction of (10). First, from the definition of $\tilde{x}$ and the assumption that $G(\tilde{x}(\hat{y}),\hat{y}) = 0$ we obtain

$$\frac{dG}{d\bar{y}}(\tilde{x}(\bar{y}),\bar{y})|_{\bar{y}=\hat{y}} = \left(-\frac{\tilde{x}'}{\tilde{x}} - \tilde{x}\,e^{\tilde{x}^2/4}\int_0^{\tilde{x}} e^{\bar{y}}\,e^{(1-2e^{\bar{y}})s^2/4}\left(1-\frac{e^{\bar{y}}s^2}{2}\right)^2 ds\right)\Bigg|_{\bar{y}=\hat{y}}. \tag{11}$$

Next, substitute $u = s(2\,e^{\bar{y}}-1)^{1/2}$ into (11) and use (9) to conclude that

$$\frac{dG}{d\bar{y}}(\tilde{x}(\bar{y}),\bar{y})\Bigg|_{\bar{y}=\hat{y}} = -\frac{\tilde{x}'}{\tilde{x}} - \frac{\tilde{x}^2}{2} - \frac{\tilde{x}\,e^{\tilde{x}^2/4}(e^{\bar{y}}-1)}{2(2\,e^{\bar{y}}-1)^{3/2}}\int_0^{\tilde{x}(2e^{\bar{y}}-1)^{1/2}/2} u^2\,e^{-u^2/4}\,du. \tag{12}$$

A straightforward calculation shows that

$$-\frac{\tilde{x}'}{\tilde{x}} - \frac{\tilde{x}^2}{2} = \frac{\tilde{x}^2(e^{\bar{y}}-2\ln(2)-2\bar{y})-2}{2\tilde{x}^2(e^{\bar{y}}-1)} < 0 \tag{13}$$

for $0<\bar{y}\leqq 1-\ln(2)$. Thus, from (12), (13) and the assumption that $\hat{y}\in(0, 1-\ln(2)]$, it follows that $dG/d\bar{y}(\tilde{x}(\bar{y}),\bar{y})|_{\bar{y}=\hat{y}}<0$, contradicting (10).

This completes the proof of the first part of our theorem, that (1), (2) has no solution for $\alpha>0$.

4. Nonexistence for $\alpha<0$.

We now show that the Kassoy problem has no solution for $\alpha<0$. This completes the proof of the theorem.

LEMMA 9. *For $\alpha<0$, the solution $y(x,\alpha)$ of (1) cannot satisfy (2).*

Proof. If $y(x) = y(x,\alpha)$ satisfies (2), then y attains a relative maximum at some $\hat{x}>0$. It follows from (1) and uniqueness of solutions that $y(\hat{x})>0$ and $y''(\hat{x})<0$. Thus there exist $x_1\in(0,\hat{x})$ and $x_2>\hat{x}$ such that $y(x_1) = 0 = y(x_2)$ and $y(x)>0$ on (x_1, x_2).

Since $e^z - 1 > z$ for $z \neq 0$, we have that $y(x)$ satisfies $y'' - (x/2)y' + y < 0$ on $[0, x_1]$. Thus, $y(x)$ is an upper solution of

$$z'' - \frac{x}{2}z' + z = 0 \tag{14}$$

and if $u(x)$ is the solution of (14) with $u(0) = \alpha$, $u'(0) = 0$, then $y(x) < u(x)$ on $(0, \sqrt{2}]$, where $\sqrt{2}$ is the only zero of $u(x)$. We conclude that $x_2 > x_1 > \sqrt{2}$.

At x_2, $y'(x_2) < 0$, $y''(x_2) < 0$, $y'''(x_2) \leqq 0$ and $y^{(4)}(x_2) < 0$. It follows from (5) that $y'''(x) < 0$ for all $x > x_2$ and hence $y''(x) < 0$ for all $x > x_2$. This contradicts our assumption that $y(x)$ satisfies (2).

REFERENCES

[1] J. Bebernes and D. Kassoy, *A mathematical analysis of blowup for thermal reactions—the spatially nonhomogeneous case*, SIAM J. Appl. Math., 40 (1981), pp. 476-484.

[2] J. Bebernes and W. Fulks, *The small heat-loss problem*, J. Differential Equations, 57 (1985), pp. 324-332.

[3] J. W. Dold, *Analysis of the early stage of thermal runaway*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 361-387.

[4] A. Friedman and B. McLeod, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425-447.

[5] A. K. Kapila, *Reactive-diffusive system with Arrhenius kinetics: Dynamics of ignition*, SIAM J. Appl. Math., 39 (1980), pp. 21-36.

[6] D. R. Kassoy and J. Poland, *The thermal explosion confined by a constant temperature boundary: I. The induction period solution*, SIAM J. Appl. Math., 39 (1980), pp. 412-430.

[7] ———, *The thermal explosion confined by a constant temperature boundary: II. The extremely rapid transient*, SIAM J. Appl. Math., 41 (1981), pp. 231-246.

[8] A. A. Lacey, *Mathematical analysis of thermal runaway for spatially inhomogeneous reactions*, SIAM J. Appl. Math., 43 (1983), pp. 1350-1366.

[9] ———, *The form of blowup for nonlinear parabolic equations*, Proceedings Royal Soc. Edinburgh Sect. A, 98 (1984), pp. 183-202.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

018

ORTHOGONAL POLYNOMIALS AND THEIR DERIVATIVES, II*

S. BONAN†, D. S. LUBINSKY‡ AND P. NEVAI§

Abstract. Let $d\alpha$ and $d\beta$ be nonnegative mass distributions on the real line, with all moments finite, and with infinitely many points of increase. Let $\{p_n\}$ and $\{q_n\}$ be the orthonormal polynomials associated with $d\alpha$ and $d\beta$ respectively. We characterize $d\alpha$ and $d\beta$ in the case when there exists a fixed rational function R, a positive integer j and nonnegative integers s and t such that, for $n=0, 1, 2, 3, \cdots$, $Rp_n^{(j)}$ may be expressed as a linear combination of $q_{n-j-t}, q_{n-j-t+1}, \cdots, q_{n-j+s}$.

Key words. orthogonal polynomials, derivatives of orthogonal polynomials, generalised Jacobi weights, exponential weights

AMS(MOS) subject classification. 42C05

1. Introduction. Let $d\alpha$ and $d\beta$ be finite positive measures on the real line with infinitely many points of increase and all moments finite. We shall call $d\alpha$ and $d\beta$ distributions, and shall denote the corresponding sequences of orthonormal polynomials by $\{p_n\}$ and $\{q_n\}$.

The purpose of this paper is to solve the following problem: Characterize $d\alpha$ and $d\beta$ for which there exist a rational function $R=S/T$, a positive integer j, and nonnegative integers s and t such that

$$Rp_n^{(j)} = \sum_{k=n-j-t}^{n-j+s} c_{nk}q_k, \qquad n=0,1,2,\cdots, \tag{1.1}$$

where the coefficients $\{c_{nk}\}$ are real numbers and $c_{nk}=0$ if $k<0$. It turns out that, at least in a description of $d\alpha$ and $d\beta$, the integer j and the denominator polynomial T are unimportant, while S plays an important role.

In solving (1.1), our results largely resolve a problem raised by Askey (see Al-Salam and Chihara [1, p. 69]) to characterize sequences of polynomials $\{p_n\}$ satisfying (1.1) with $R=S, j=1$ and $\{p_n\}=\{q_n\}$. An algebraic solution to Askey's problem has recently been provided by Maroni [13], [14], involving quasi-orthogonality and linear forms defined on the space of all polynomials. Earlier work along this line is due to Ronveaux [20] and Hendriksen and van Rossum [8]. By contrast, our characterization of polynomials satisfying (1.1) is analytic, and provides a complete description of the distributions $d\alpha$ and $d\beta$ associated with $\{p_n\}$ and $\{q_n\}$.

Relationships such as (1.1) are useful in studying analytic aspects of orthonormal polynomials. For example, Bonan [3] recently established a relationship similar to (1.1) with $R=1, j=1$ and $d\alpha=d\beta$, for the weights $d\alpha(x)=\exp(-x^m)\,dx$, where m is a positive even integer. He used this identity to prove part of a conjecture of Nevai [16] concerning bounds on orthonormal polynomials. (See also [10], [18].)

Further, (1.1) is a unifying feature of the most important families of orthogonal polynomials. It is well known that the classical orthogonal polynomials are the only orthogonal polynomials that satisfy (1.1) with $R=1, j=1$ and $s=t=0$ (Freud [5, p.

* Received by the editors July 15, 1985; accepted for publication (in revised form) April 22, 1986.

† Department of Mathematics, University of Rhode Island, Kingston, Rhode Island 02881.

‡ National Research Institute for Mathematical Sciences, CSIR, P.O. box 395, Pretoria 0001, Republic of South Africa. The author began working on this material during a visit to Ohio State University.

§ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The work of this author was supported by the National Science Foundation under grant MDS-83-00882.

52])—this special case has been called the problem of W. Hahn [22]. It turns out that the generalised Jacobi weights (GJ) studied by Badkov [2], Geronimus [7], Nevai [15], [17], [19] satisfy several relationships such as (1.1), as do exponential weights on $[0, \infty)$ of the form $\exp(-P(x))$, $P(x)$ a polynomial of positive degree with positive leading coefficient. In this context, the latter have been studied by Ronveaux [20]. For further historical remarks and references, see Bonan and Nevai [4], Hendriksen and van Rossum [8], and Maroni [13], [14]. We note that our main result, Theorem 1.1, essentially contains [8, Thms. 3.1 and 3.2]. The latter two theorems in [8] largely characterize "semiclassical orthogonal polynomial systems" in terms of a differentiation property that is a special case of (1.1).

Given an interval I, we let χ_I denote the characteristic function of I, while δ_x denotes a (unit) point mass at a real number x.

THEOREM 1.1. *Let $d\alpha$ and $d\beta$ be distributions, with corresponding sequences of orthonormal polynomials $\{p_n\}$ and $\{q_n\}$. The following are equivalent:*

I. *There exist a real rational function $R = S/T$ (not identically zero), a positive integer j and nonnegative integers s' and t' such that*

$$Rp_n^{(j)} = \sum_{k=n-j-t'}^{n-j+s'} c'_{nk} q_k, \qquad n = 0, 1, 2, \cdots, \tag{1.2}$$

where the $\{c'_{nk}\}$ are real numbers with $c'_{nk} = 0$, $k < 0$.

II. *There exist a real polynomial S (not identically zero) and nonnegative integers s and t such that*

$$Sp'_n = \sum_{k=n-1-t}^{n-1+s} c_{nk} q_k, \qquad n = 0, 1, 2, \cdots, \tag{1.3}$$

where the $\{c_{nk}\}$ are real numbers with $c_{nk} = 0$, $k < 0$.

III. *There exist nonnegative integers s and t, and real polynomials S, U and V, not identically zero, and with degrees at most s, $t+1$ and $t+2$ respectively. Further, there exist nonnegative integers N and N', and nonnegative numbers A_0, $A_1 \cdots A_N$ (not all zero), $\lambda_1, \lambda_2 \cdots \lambda_N$, and $\mu_1, \mu_2 \cdots \mu_{N'}$, with the following properties: V has N real zeros*

$$-\infty < v_1 < v_2 < \cdots < v_N < \infty$$

and S has N' real zeros

$$-\infty < s_1 < s_2 \cdots < s_{N'} < \infty.$$

Let $v_0 = -\infty$, $v_{N+1} = \infty$ and

$$I_K = (v_k, v_{k+1}), \qquad k = 0, 1, 2, \cdots, N. \tag{1.4}$$

Then

$$d\alpha(x) = \frac{H(x)}{|V(x)|} \exp\left(-\int^x \frac{U(u)}{V(u)}\,du\right) dx + \sum_{k=1}^{N} \lambda_k \delta_{v_k}(x) \tag{1.5}$$

and

$$d\beta(x) = \frac{H(x)}{|S(x)|} \exp\left(-\int^x \frac{U(u)}{V(u)}\,du\right) dx + \sum_{k=1}^{N'} \mu_k \delta_{s_k}(x), \tag{1.6}$$

where

$$H(x) = \sum_{k=0}^{N} A_k \chi_{I_k}(x), \qquad x \in \mathbb{R}, \tag{1.7}$$

$$\lambda_k > 0 \Rightarrow U(v_k) = 0, \tag{1.8}$$

and S and V have the same (nonzero) sign in any interval I_k in which $\alpha' \neq 0$.

If any of the assertions of Theorem 1.1, I, II, III, hold, we may choose the polynomial S to be the same in all of I, II and III. Further, we may relate s, s', t and t' by

$$s-s'=\text{degree}\,(T) \quad\text{and}\quad t-t'=\text{degree}\,(T)+2j-1. \tag{1.9}$$

The degree of S is s, while U and V admit the representation

$$U(x)=\int_{-\infty}^{\infty}\sum_{k=0}^{t+1} p_k(x)p_k'(u)S(u)\,d\beta(u), \tag{1.10}$$

$$V(x)=\int_{-\infty}^{\infty}\sum_{k=0}^{t+2} p_k(x)p_k(u)S(u)\,d\beta(u). \tag{1.11}$$

The following result gives more information on $d\alpha$ and $d\beta$, but is an immediate consequence of Theorem 1.1 and the partial fraction decomposition of U/V.

THEOREM 1.2. *Let $d\alpha$ and $d\beta$ be distributions, with corresponding sequences of orthonormal polynomials $\{p_n\}$ and $\{q_n\}$. Then assertions* I, II *and* III *of Theorem* 1.1 *hold if and only if $d\alpha$ and $d\beta$ are given by* (1.4)-(1.8). *Further,*

$$\phi(x)=\exp\left(-\int^{x} U(u)/V(u)\,du\right) \tag{1.12}$$

satisfies

$$\begin{aligned}\phi(x)&=\exp(-P(x))\prod_{k=1}^{N}|x-v_k|^{\Gamma_k}\\&\quad\cdot\exp\left(-\sum a_{ik}(x-v_i)^{-k}-\int^{x}U^*(u)/V^*(u)\,du\right),\end{aligned} \tag{1.13}$$

where $P(x)$ is a polynomial and $\Gamma_1, \Gamma_2\cdots\Gamma_N$ are positive. In addition, U^* and V^* are polynomials with degree $(U^*)<$ degree (V^*) *and V^* is a product of positive quadratic factors of V so that $V^*(x)>0$, $x\in(-\infty,\infty)$. Consider a term $a_{ik}(x-v_i)^{-k}$ in the finite sum $\sum a_{ik}(x-v_i)^{-k}$.*

If $A_{i-1}=A_i=0$, so that $\alpha'\equiv 0$ in $I_{i-1}\cup I_i$, then the sign of a_{ik} is arbitrary, and k may be any positive integer.

If $A_{i-1}=0$ but $A_i\neq 0$, so that $\alpha'\equiv 0$ in I_{i-1}, but $\alpha'\neq 0$ in I_i, then $a_{ik}\geqq 0$, while k may be any positive integer.

If $A_{i-1}\neq 0$ but $A_i=0$, so that $\alpha'\neq 0$ in I_{i-1}, but $\alpha'\equiv 0$ in I_i then $a_{ik}\geqq 0$ if k is even, while $a_{ik}\leqq 0$ if k is odd.

If $A_{i-1}\neq 0$ and $A_i\neq 0$, so that $\alpha'\neq 0$ in $I_{i-1}\cup I_i$, then $a_{ik}=0$ if k is odd, while $a_{ik}\geqq 0$ if k is even.

Finally, if $A_0\neq 0$ and $A_N\neq 0$ so that $\alpha'(x)\neq 0$ for large enough $|x|$, then $P(x)$ must be a polynomial of positive even degree with positive leading coefficient. If $A_0=0$ but $A_N\neq 0$, so that $\alpha'(x)\neq 0$ for large positive x, but $\alpha'(x)=0$ for large negative x, then $P(x)$ must be a polynomial of positive degree with positive leading coefficient.

Using the partial fraction decomposition of U^*/V^*, we may express $\exp(-\int^x U^*(u)/V^*(u)\,du)$ as a product of powers $(x^2+Bx+C)^D$ of positive quadratic factors of V, and of terms of the form $\exp(-E\arctan(Fx+g))$ and $\exp(-Y(x)/(x^2+Bx+C)^J)$, where J is a positive integer and Y is a polynomial of degree less than $2J$. At this stage, it is pertinent to illustrate the above theorems with some examples.

Example 1.3: GJ. The generalized Jacobi weights GJ [15] take the form

$$w(x)=\begin{cases}\prod_{j=1}^{N}|x-v_j|^{\Gamma_j}, & x\in[-1,1],\\ 0\end{cases}\tag{1.14}$$

where $N\geqq 2$, $\Gamma_j>-1$, $j=1,2,\cdots,N$, and

$$-1=v_1<v_2<\cdots<v_N=1.$$

Let us set

$$d\alpha(x)=w(x)\,dx.$$

The choice of V, U and $d\beta$ in (1.5) to (1.11) depend on S. In particular V and S must have the same sign in $(-1,1)$. We consider two choices for S.

CASE 1: $S=1$. Since V must be nonnegative in $(-1,1)$ and vanish at $v_1, v_2\cdots v_N$, we set

$$V(x)=(1-x^2)\prod_{j=2}^{N-1}(x-v_j)^2$$

and

$$U(x)=-V(x)\left\{\frac{\Gamma_1+1}{x+1}+\frac{\Gamma_{N+1}+1}{x-1}+\sum_{j=2}^{N-1}\frac{\Gamma_j+2}{x-v_j}\right\},$$

so that V has degree $2N-2$, and U has degree $2N-3$, as each $\Gamma_j+1>0$. It is clear that (1.5) holds with the obvious choice for H, and from (1.6), we see that

$$d\beta(x)=w_1(x)\,dx$$

where

$$w_1(x)=\begin{cases}(1+x)^{\Gamma_1+1}(1-x)^{\Gamma_N+1}\prod_{j=2}^{N-1}|x-v_j|^{\Gamma_j+2}, & x\in[-1,1],\\ 0 & \text{otherwise.}\end{cases}$$

From (1.10) and (1.11), we see that we may choose $t=2N-4$. Thus, by (1.3), the orthonormal polynomials $\{p_n\}$ and $\{q_n\}$ associated with w and w_1 satisfy

$$p_n'(x)=\sum_{k=n-2N+3}^{n-1}c_{nk}q_k(x),\qquad n=0,1,2,\cdots.$$

In the case $N=2$, this is precisely the classical result that the derivative of an orthonormal polynomial associated with a Jacobi weight $w(x)$ is an orthogonal polynomial associated with the Jacobi weight $w_1(x)=(1-x^2)w(x)$.

CASE 2: $S(x)=\prod_{j=1}^{N}(x-v_j)$. In this case we may choose $V(x)=S(x)$ and

$$U(x)=-V(x)\sum_{j=1}^{N}(\Gamma_j+1)/(x-v_j),$$

so that U and V have degree $N-1$ and N, respectively. It is clear that (1.5) is valid, while from (1.6), we may set $d\beta(x)=d\alpha(x)$. Finally from (1.10) and (1.11), we may choose $t=N-2$. Thus, by (1.3),

$$S(x)p_n'(x)=\sum_{k=n+1-N}^{n-1+N}c_{nk}p_k(x),\qquad n=0,1,2,\cdots.$$

Example 1.4: GJ and point masses. Let

$$d\alpha(x)=w(x)\,dx+\sum_{j=1}^{N}\lambda_j\delta_{v_j}(x),$$

where $w(x)$ is the generalised Jacobi weight (1.14), and λ_1 and λ_N are positive, while $\lambda_2, \lambda_3\cdots\lambda_N$ are nonnegative. In the case $N=2$, this weight has been considered by Koornwinder [9].

Let $S=1$, and let us determine what U, V and $d\beta$ should be. From (1.8), $U(\pm1)=0$, while V must be nonnegative throughout $(-1,1)$. We may set

$$V(x)=\prod_{j=1}^{N}(x-v_j)^2$$

and

$$U(x)=-V(x)\sum_{j=1}^{N}(\Gamma_j+2)/(x-v_j),$$

so that V and U have degree $2N$ and $2N-1$, respectively. From (1.6), we see that we must choose

$$d\beta(x)=w_1(x)\,dx,$$

where

$$w_1(x)=\begin{cases}\prod_{j=1}^{N}|x-v_j|^{\Gamma_j+2}, & x\in[-1,1],\\ 0 & \text{otherwise.}\end{cases}$$

From (1.10) and (1.11), we may choose $t=2N-2$. Thus

$$p_n'(x)=\sum_{k=n-2N+1}^{n-1}c_{nk}q_k(x),\qquad n=0,1,2,\cdots.$$

Example 1.5: exponential weights. Let

$$d\alpha(x)=\exp\left(-P(x)\right)dx,\qquad x\in\mathbb{R},$$

where $P(x)$ is a polynomial of even positive degree m with positive leading coefficient. In this case, we may set $V=1$ and $U=P'$, so that V and U have degree 0 and $m-1$, respectively. If we set $S=1$, then $d\beta(x)=d\alpha(x)$, and from (1.10) and (1.11), we may choose $t=m-2$. Thus

$$p_n'(x)=\sum_{k=n-m+1}^{n-1}c_{nk}p_k(x),\qquad n=0,1,2,\cdots.$$

Using the results and methods of [3], [6], [11] this may be used to establish bounds on $p_n(x)$. Magnus [12] recently proved Freud's Conjecture for the above weights.

2. Proof of Theorems 1.1 and 1.2. To prove Theorem 1.1, we shall show I⇔II⇔III. First however, we need to introduce notation: The orthonormal polynomial sequences $\{p_n\}$ and $\{q_n\}$ satisfy three term recurrence relations

$$xp_n=a_{n+1}p_{n+1}+a_np_{n-1}+b_np_n,\qquad n=1,2,3,\cdots,\tag{2.1}$$

$$xq_n=e_{n+1}q_{n+1}+e_nq_{n-1}+f_nq_n,\qquad n=1,2,3,\cdots,\tag{2.2}$$

where

$$a_0=e_0=0,\quad a_n>0,\quad e_n>0,\quad n=1,2,3,\cdots,$$

while b_n and f_n are real, $n=1, 2, 3, \cdots$. Further, if $\gamma_n(d\alpha)$ and $\gamma_n(d\beta)$ denote the leading coefficients of p_n, $n=0, 1, 2, \cdots$, it is well known that

$$a_n=\gamma_{n-1}(d\alpha)/\gamma_n(d\alpha), \quad e_n=\gamma_{n-1}(d\beta)/\gamma_n(d\beta), \quad n=1,2,3,\cdots.$$

Proof of I⇔II. The implication II⇒I is trivial, so we must show I⇒II. By hypothesis, we have $R=S/T$ satisfying

$$Sp_n^{(j)}=T\sum_{k=n-j-t'}^{n-j+s'} c'_{nk}q_k, \qquad n=0,1,2,\cdots. \tag{2.3}$$

We shall first show that for some real $\{c''_{nk}\}$,

$$Sp_n^{(j)}=\sum_{k=n-j-t''}^{n-j+s''} c''_{nk}q_k, \qquad n=0,1,2,\cdots \tag{2.4}$$

where $t''-t'=s''-s'=\text{degree}\,(T)$. It clearly suffices to show that there exist real numbers a_{jkl} with $a_{jkl}=0$ for $l<0$, such that

$$x^jq_k(x)=\sum_{l=k-j}^{k+j} a_{jkl}q_l(x), \qquad j, k=0,1,2,\cdots. \tag{2.5}$$

But this is trivial for $j=0$ and $k=0, 1, 2, \cdots$, while for $j=1$ and $k=0, 1, 2, \cdots$, it follows from the recurrence relation. An easy induction on j establishes (2.5) for all $j, k=0, 1, 2, \cdots$. Thus, (2.4) follows from (2.3) and (2.5).

Next, applying Leibniz's formula for the jth derivative of a product of two functions to (2.1), we see that

$$xp_n^{(j)}+jp_n^{(j-1)}=a_{n+1}p_{n+1}^{(j)}+a_np_{n-1}^{(j)}+b_np_n^{(j)}, \qquad n=0,1,2,\cdots,$$

so that

$$\begin{aligned} Sp_n^{(j-1)}&=j^{-1}\{a_{n+1}Sp_{n+1}^{(j)}+a_nSp_{n-1}^{(j)}+b_nSp_n^{(j)}-xSp_n^{(j)}\}\\ &=\sum_{k=n-j-1-t''}^{n-j+1+s''} c_{nk}^{\#}q_k, \end{aligned}$$

for some real numbers $\{c_{nk}^{\#}\}$. Here we have used (2.4) and the recurrence relation (2.2) for xq_k. Proceeding in the same way, we obtain after altogether $j-1$ such steps that (1.3) holds, where s, s', t and t' are related by (1.9). □

The proof of II⇒III will be split into several steps.

LEMMA 2.1. *Assume that assertion* II *of Theorem* 1.1 *holds.*

(i) *Then there exist real numbers* $\{c_{nk}^*\}$ *such that* $c_{nk}^*=0$ *for* $k<0$, *and*

$$Sp_n=\sum_{k=n-2-t}^{n+s} c_{nk}^*q_k, \qquad n=0,1,2,\cdots. \tag{2.6}$$

(ii) *Define polymials* U *and* V *of degree at most* $t+1$ *and* $t+2$ *respectively by* (1.10) *and* (1.11). *Then for every polynomial* P,

$$\int_{-\infty}^{\infty} P(x)V(x)\,d\alpha(x)=\int_{-\infty}^{\infty} P(x)S(x)\,d\beta(x) \tag{2.7}$$

and

$$\int_{-\infty}^{\infty} P(x)U(x)\,d\alpha(x)=\int_{-\infty}^{\infty} P'(x)S(x)\,d\beta(x). \tag{2.8}$$

Proof. (i) From the recurrence relation (2.1)

$$xp_n'(x)+p_n(x)=a_{n+1}p_{n+1}'(x)+a_np_{n-1}'(x)+b_np_n'(x),$$

so that

$$S(x)p_n(x)=a_{n+1}S(x)p_{n+1}'(x)+a_nS(x)p_{n-1}'(x)+b_nS(x)p_n'(x)-xS(x)p_n'(x).$$

Then (2.6) follows easily from (1.3) and the recurrence relation (2.2) for xq_k.

(ii) It suffices to show that (2.7) and (2.8) hold for $P=p_n$, $n=0, 1, 2, \cdots$. Now if V is given by (1.11), so that V has degree at most $t+2$,

$$\int_{-\infty}^{\infty} p_n(x)V(x)\,d\alpha(x)=\begin{cases}0, & n>t+2,\\ \displaystyle\int_{-\infty}^{\infty} p_n(u)S(u)\,d\beta(u), & n\leqq t+2.\end{cases}$$

But (2.6) and orthonormality of $\{q_k\}$ with respect to $d\beta$, show that

$$\int_{-\infty}^{\infty} p_n(u)S(u)\,d\beta(u)=0, \qquad n>t+2.$$

Thus (2.7) holds for $P=p_n$, $n=0, 1, 2, \cdots$. Next, if U is given by (1.10)

$$\int_{-\infty}^{\infty} p_n(x)U(x)\,d\alpha(x)=\begin{cases}0, & n>t+1,\\ \displaystyle\int_{-\infty}^{\infty} p_n'(u)S(u)\,d\beta(u), & n\leqq t+1.\end{cases}$$

But (1.3) and orthonormality of $\{q_k\}$ with respect to $d\beta$ show that

$$\int_{-\infty}^{\infty} p_n'(u)S(u)\,d\beta(u)=0, \qquad n>t+1.$$

Thus (2.8) holds for $P=p_n$, $n=0, 1, 2, \cdots$. □

Next, we estimate some modified moments.

LEMMA 2.2. *Assume that assertion* II *of Theorem* 1.1 *holds. Let*

$$L_n=\int_{-\infty}^{\infty}|x|^n\,d\alpha(x), \qquad n=0,1,2,\cdots, \tag{2.9}$$

and

$$M_n=\int_{-\infty}^{\infty}|x|^n\,d\beta(x), \qquad n=0,1,2,\cdots. \tag{2.10}$$

Then

$$\limsup_{n\to\infty} L_n^{1/n}/n<\infty \tag{2.11}$$

and

$$\limsup_{n\to\infty} M_n^{1/n}/n<\infty. \tag{2.12}$$

Proof. First note that as $S\equiv 0$, (2.7) and (2.8) show that $U\equiv 0$ and $V\equiv 0$. Now from (2.8) with $P(x)=x^{n+1}$,

$$\begin{aligned}\int_{-\infty}^{\infty} x^{n+1}U(x)\,d\alpha(x)&=(n+1)\int_{-\infty}^{\infty} x^nS(x)\,d\beta(x)\\ &=(n+1)\int_{-\infty}^{\infty} x^nV(x)\,d\alpha(x),\end{aligned}$$

by (2.7). Thus

$$\int_{-\infty}^{\infty} x^n[xU(x)-(n+1)V(x)]\,d\alpha(x)=0, \qquad n=0,1,2,\cdots. \tag{2.13}$$

Next, there exist nonzero constants C_1 and C_2 and nonnegative integers p and q such that

$$xU(x)=C_1x^p(1+o(1)), \qquad |x|\to\infty$$

and

$$V(x)=C_2x^q(1+o(1)), \qquad |x|\to\infty.$$

We distinguish between the cases where $xU(x)$ or $(n+1)V(x)$ dominates in (2.13).

CASE 1: $p\leqq q$. As usual, for any real x, we let sign (x) denote the sign of x if $x\neq 0$, while sign $(0)=0$. Now there exist $C>0$ and n_0 such that

$$\operatorname{sign}(-C_2x^q)\{xU(x)-(n+1)V(x)\}\geqq 1, \qquad |x|\geqq C, \quad n\geqq n_0.$$

Then for $n\geqq n_0$ and $n+q$ even, so that sign $(x^n)=$ sign (x^q),

$$\int_{|x|\geqq C}|x|^n\,d\alpha(x)\leqq \operatorname{sign}(-C_2)\int_{|x|\geqq C}x^n\{xU(x)-(n+1)V(x)\}\,d\alpha(x)$$
$$=-\operatorname{sign}(-C_2)\int_{|x|\leqq C}x^n\{xU(x)-(n+1)V(x)\}\,d\alpha(x)=O(nC^n),$$

by (2.13). Since

$$\int_{|x|\leqq C}|x|^n\,d\alpha(x)=O(C^n),$$

we obtain, for $n\geqq n_0$ and $n+q$ even,

$$L_n=O(nC^n).$$

However for $n+q$ odd,

$$L_n\leqq\int_{-1}^{1}d\alpha(x)+\int_{|x|\geqq 1}|x|^{n+1}\,d\alpha(x)\leqq L_0+L_{n+1} \tag{2.14}$$

and then (2.11) follows.

CASE 2: $p>q$. Let $X(n)=16|C_2/C_1|n$, $n=1,2,3,\cdots$. Then for $|x|\geqq X(n)$, and n large enough,

$$\operatorname{sign}(C_1x^p)\{xU(x)-(n+1)V(x)\}\geqq\{|C_1x^p|/2-2(n+1)|C_2x^q|\}$$
$$\geqq|C_1x^p|/2\{1-8n|C_2/C_1|\,|x|^{q-p}\}\geqq 1$$

as $q-p\leqq -1$. Hence if $n+p$ is even and n is large enough, so that sign $(x^n)=$ sign (x^p),

$$\int_{|x|\geqq X(n)}|x|^n\,d\alpha(x)\leqq\operatorname{sign}(C_1)\int_{|x|\geqq X(n)}x^n\{xU(x)-(n+1)V(x)\}\,d\alpha(x)$$
$$=-\operatorname{sign}(C_1)\int_{|x|\leqq X(n)}x^n\{xU(x)-(n+1)V(x)\}\,d\alpha(x)\quad \text{by (2.13))}$$
$$=O(nX(n)^{n+p}).$$

Further,

$$\int_{|x|\le X(n)} |x|^n \, d\alpha(x) = O(X(n)^n),$$

and thus if $n+p$ is even,

$$L_n = O(nX(n)^{n+p}) = O((C_3 n)^{n+p}),$$

for some $C_3 > 0$. This establishes (2.11) if we restrict n to those integers for which $n+p$ is even. As before, (2.14) yields (2.11) in the general case. Finally, (2.12) follows easily from (2.7) and (2.11). □

Finally, we need two Fourier transform identities.

LEMMA 2.3. *Assume that assertion* II *of Theorem* 1.1 *holds. Then for all real u,*

$$\int_{-\infty}^{\infty} e^{iux} V(x) \, d\alpha(x) = \int_{-\infty}^{\infty} e^{iux} S(x) \, d\beta(x) \tag{2.15}$$

and

$$\int_{-\infty}^{\infty} e^{iux} U(x) \, d\alpha(x) = iu \int_{-\infty}^{\infty} e^{iux} S(x) \, d\beta(x). \tag{2.16}$$

Proof. Our proof will be similar to that of the uniqueness criterion of M. Riesz for the moment problem [5, pp. 79-80]. Let

$$\varphi_1(u) = \int_{-\infty}^{\infty} e^{iux} V(x) \, d\alpha(x)$$

and

$$\varphi_2(u) = \int_{-\infty}^{\infty} e^{iux} S(x) \, d\beta(x) \tag{2.17}$$

for all real u. We shall show that there exists $r > 0$ such that both $\varphi_1(u)$ and $\varphi_2(u)$ are analytic in the strip $\{u \in \mathbb{C}: |\text{Im } u| < r\}$. To this end, we first note that by Lemma 2.2, there exists $r > 0$ such that the series $\sum_{k=0}^{\infty} L_{k+t+2} z^k / k!$ and $\sum_{k=0}^{\infty} M_{k+t+2} z^k / k!$ converge absolutely and uniformly in compact subsets of $|z| < r$. Then given real u and $|z| < r$, we have formally

$$\begin{aligned}\varphi_1(u+z) &= \int_{-\infty}^{\infty} e^{ix(u+z)} V(x) \, d\alpha(x) \\ &= \int_{-\infty}^{\infty} e^{ixu} \sum_{k=0}^{\infty} (ixz)^k / k! \, V(x) \, d\alpha(x) \\ &= \sum_{k=0}^{\infty} (iz)^k \int_{-\infty}^{\infty} e^{ixu} x^k V(x) \, d\alpha(x) / k!.\end{aligned}$$

As V has degree at most $t+2$, it is not difficult to see that the coefficient of z^k in this last series is bounded by $C \ (L_{k+t+2} + L_k)/k!$, where C is a positive constant independent of k. Thus the power series converges uniformly for z in compact subsets of $|z| < r$, where $r > 0$ is independent of u. We deduce that the formal interchanges above are analytically valid and that $\varphi_1(u)$ is analytic in the strip $\{u : |\text{Im } u| < r\}$. Similarly φ_2 is analytic in this strip.

If we can show that for some $\varepsilon > 0$,

$$\varphi_1(u) = \varphi_2(u), \qquad u \in (-\varepsilon, \varepsilon),$$

so that (2.15) is valid for $u \in (-\varepsilon, \varepsilon)$, then analytic continuation will show that (2.15) is valid for all real u. Now from (2.7), we have that (2.15) is valid for real u, provided we replace e^{iux} by its $(n+1)$th partial sum,

$$P_n(xu) = \sum_{k=0}^{n} (iux)^k/k!,$$

that is,

$$\int_{-\infty}^{\infty} P_n(xu) V(x)\, d\alpha(x) = \int_{-\infty}^{\infty} P_n(xu) S(x)\, d\beta(x). \tag{2.18}$$

Next, applying Taylor's formula to the real and imaginary parts of e^{iy}, y real, we see (cf. [5, p. 79]) that

$$|e^{iy} - P_n(y)| \leqq 2|y|^n/n!, \qquad y \text{ real}.$$

Then, using (2.18), we see that

$$\begin{aligned}\left|\int_{-\infty}^{\infty} e^{iux} V(x)\, d\alpha(x) - \int_{-\infty}^{\infty} e^{iux} S(x)\, d\beta(x)\right| \\ &= \left|\int_{-\infty}^{\infty} \{e^{iux} - P_n(xu)\}(V(x)\, d\alpha(x) - S(x)\, d\beta(x))\right| \\ &\leqq 2|u|^n \int_{-\infty}^{\infty} |x|^n \{|V(x)|\, d\alpha(x) + |S(x)|\, d\beta(x)\}/n! \\ &\to 0 \quad \text{as } n \to \infty,\end{aligned}$$

if $|u|$ is small enough, by Lemma 2.2, and as S and V are polynomials. Thus (2.15) is valid.

The proof of (2.16) is rather similar. One shows that the left and right members of (2.16) are analytic functions of u in a strip containing the real line. Further, one uses (2.8) to show that

$$\int_{-\infty}^{\infty} P_n(ux) U(x)\, d\alpha(x) = iu \int_{-\infty}^{\infty} P_{n-1}(ux) S(x)\, d\beta(x)$$

and proceeds much as above. □

Recall that we can write

$$d\alpha\ (x) = \alpha'(x)\, dx + d\alpha_j(x) + d\alpha_s(x),$$

where $\alpha'(x)\, dx$ is the absolutely continuous part of $d\alpha$, while $d\alpha_j$ is a series of point masses, and $d\alpha_s$ is the singularly continuous part of $d\alpha$. Similar remarks apply to $d\beta$.

Proof of II⇒III. Let U and V be given by (1.10) and (1.11), and let ϕ_2 be given by (2.17). From (2.16), we see that

$$\varphi_2(u) = O(u^{-1}), \qquad |u| \to \infty,$$

and hence

$$\int_{-x}^{x} |\varphi_2(u)|\, du = o(x), \qquad x \to \infty.$$

It then follows from Wiener's Theorem (Zygmund [21, p. 261]) that

$$F(u) = \int_{-\infty}^{u} S(x)\, d\beta(x) \tag{2.19}$$

is continuous in $(-\infty, \infty)$. Next, integrating the left member of (2.16) by parts, and using

$$\int_{-\infty}^{\infty} U(x)\, d\alpha(x) = 0 \tag{2.20}$$

(which follows from (2.16) with $u = 0$), we obtain

$$-\int_{-\infty}^{\infty} iu\, e^{iux}\left\{\int_{-\infty}^{x} U(y)\, d\alpha(y)\right\} dx = iu\int_{-\infty}^{\infty} e^{iux} S(x)\, d\beta(x)$$

and hence that

$$-\int_{-\infty}^{\infty} e^{iux}\left\{\int_{-\infty}^{x} U(y)\, d\alpha(y)\right\} dx = \int_{-\infty}^{\infty} e^{iux} S(x)\, d\beta(x), \tag{2.21}$$

$u \neq 0$. We can let $u \to 0$ to deduce that (2.21) holds even for $u = 0$. To see this it suffices to prove

$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{x} U(y)\, d\alpha(y)\right| dx < \infty, \tag{2.22}$$

and then to apply Lebesgue's Dominated Convergence Theorem. Now, since all moments of $d\alpha$ are finite, and since (2.20) holds, we see that as $x \to \infty$,

$$\int_{-\infty}^{x} U(y)\, d\alpha(y) = -\int_{x}^{\infty} U(y)\, d\alpha(y) = O(x^{-j}),$$

for each positive integer j. Further as $x \to -\infty$,

$$\int_{-\infty}^{x} U(y)\, d\alpha(y) = O(x^{-j})$$

for each positive integer j. Thus (2.22) holds and hence (2.21) is valid for all real u.

We may now use the uniqueness of Fourier transforms and the continuity of $F(u)$ in (2.19) (see Zygmund [21, p. 293, Thm. 10.15]) to deduce for all real u.

$$-\int_{-\infty}^{u}\left\{\int_{-\infty}^{x} U(y)\, d\alpha(y)\right\} dx = \int_{-\infty}^{u} S(x)\, d\beta(x). \tag{2.23}$$

It now follows that $S(x)\, d\beta(x)$ is absolutely continuous in $(-\infty, \infty)$. Thus $d\beta_s \equiv 0$ and $d\beta_j$ can have point masses only at zeros of S, so that with the notation of Theorem 1.1,

$$d\beta(x) = \beta'(x)\, dx + \sum_{k=1}^{N'} \mu_k \delta_{s_k}(x). \tag{2.24}$$

Further, from (2.23)

$$-\int_{-\infty}^{u} U(y)\, d\alpha(y) = S(u)\beta'(u), \qquad u \in (-\infty, \infty). \tag{2.25}$$

Next, using (2.15), in the same way as we used (2.21), we deduce that for all real u,

$$\begin{aligned}\int_{-\infty}^{u} V(x)\, d\alpha(x) &= \int_{-\infty}^{u} S(x)\, d\beta(x)\\ &= \int_{-\infty}^{u} S(x)\beta'(x)\, dx,\end{aligned} \tag{2.26}$$

by absolute continuity of $S\,d\beta$. We deduce that $V(x)\,d\alpha(x)$ is absolutely continuous in $(-\infty, \infty)$. Thus $d\alpha_s \equiv 0$ and $d\alpha_j$ can have point masses only at zeros of V, so that

$$d\alpha(x) = \alpha'(x)\,dx + \sum_{k=1}^{N} \lambda_k \delta_{v_k}(x), \tag{2.27}$$

with the notation of Theorem 1.1. Then (2.26) and (2.27) show that

$$V(x)\alpha'(x) = S(x)\beta'(x), \qquad x \in (-\infty, \infty). \tag{2.28}$$

Further, substituting (2.27) and (2.28) into (2.25), we obtain for all real $u \notin \{v_1, v_2 \cdots v_N\}$,

$$-\int_{-\infty}^{u} U(y)\alpha'(y)\,dy - \sum_{v_k \leqq u} \lambda_k U(v_k) = V(u)\alpha'(u). \tag{2.29}$$

We deduce that $V\alpha'$ is absolutely continuous in any interval $I_k = (v_k, v_{k+1})$, and

$$-U(u)\alpha'(u) = (V(u)\alpha'(u))', \qquad u \in I_k.$$

Hence in I_k,

$$(V\alpha')'/(V\alpha') = -U/V,$$

and integrating, we obtain

$$\alpha'(x) = (A_k/|V(x)|)\exp\left(-\int^{x} U(u)/V(u)\,du\right), \qquad x \in I_k,$$

where A_k is a nonnegative constant. Together with (2.27), this establishes (1.5) and (1.7). It is obvious that at least one A_k must be positive, since $\alpha(x)$ has infinitely many points of increase. Now consider an interval I_k in which α' is not identically zero. Then $A_k > 0$ and α' and V do not vanish in I_k. Hence the left member of (2.28) is of one sign and nonzero in I_k, so that S must have the same sign as V in I_k and does not vanish in I_k. Now (1.6) follows easily from (1.5), (2.24) and (2.28).

It remains to prove (1.8). Since the integral in (2.29) is continuous in $\mathbb{R}$, it suffices to show that $\alpha'(x)V(x)$ is continuous in $\mathbb{R}$. As $\alpha'V$ is continuous in each I_k, it suffices to show that

$$\lim_{x \to v_k} \alpha'(x)V(x) = 0, \qquad k = 1, 2, \cdots, N. \tag{2.30}$$

Fix k. If as $x \to v_k$, $\exp(-\int^x U(u)/V(u)\,du) \nrightarrow 0$, it will be bounded below by a positive number in either some left or right neighbourhood of v_k. Then if $\alpha'(x)$ is not identically zero in that neighbourhood, it follows from (1.5) that in that neighbourhood

$$\alpha'(x) \geqq C/|x - v_k|^l,$$

where l is a positive integer and $C > 0$. This contradicts the integrability of $d\alpha$. Thus either $\alpha' \equiv 0$ near v_k (in which case (2.30) is trivial), or $\exp(-\int^x U(u)/V(u)\,du) \to 0$ as $x \to v_k$ and then (1.5) yields (2.30). □

Proof of III⇒II. From (1.5) and (1.7) we see that in any interval I_k (given by (1.4)), α' is absoutely continuous and

$$(V(x)\alpha'(x))' = -U(x)\alpha'(x), \qquad x \in I_k. \tag{2.31}$$

Further, using the integrability of $d\alpha$ as in the proof of II⇒III, we see that (2.30) is true. Finally, from (1.5) and (1.6), we deduce that for all real $x \notin \{v_1, v_2, \cdots, v_N\}$,

$$V(x)\alpha'(x) = S(x)\beta'(x). \tag{2.32}$$

Let n be a nonnegative integer. As $d\beta$ has jumps only at zeros of S,

$$\begin{aligned}\int_{-\infty}^{\infty} Sp_n' q_k \, d\beta(x) &= \int_{-\infty}^{\infty} Sp_n' q_k \beta' \, dx \\ &= \int_{-\infty}^{\infty} p_n' q_k V\alpha' \, dx \quad (\text{by } (2.32)) \\ &= \sum_{k=0}^{N} \int_{I_k} p_n' q_k V\alpha' \, dx \\ &= -\sum_{k=0}^{N} \left\{ p_n q_k V\alpha' \Big|_{v_k}^{v_{k+1}} - \int_{I_k} p_n (q_k V\alpha')' \, dx \right\} \\ &= -\int_{-\infty}^{\infty} p_n \{ q_k' V\alpha' + q_k (V\alpha')' \} \, dx \quad (\text{by } (2.30)) \\ &= -\int_{-\infty}^{\infty} p_n \{ q_k' V - q_k U \} \alpha' \, dx \quad (\text{by } (2.31)) \\ &= -\int_{-\infty}^{\infty} p_n \{ q_k' V - q_k U \} \, d\alpha(x) + \sum_{k=1}^{N} \lambda_k p_n(v_k) \{ q_k' V - q_k U \}(v_k), \end{aligned} \tag{2.33}$$

by (1.5). Since $q_k' V - q_k U$ has degree at most $k+t+1$, the integral in the right member of (2.33) will vanish if $n > k+t+1$, that is if $k < n-1-t$. In view of (1.8), and as $v_1, v_2, \cdots, v_N$ are zeros of V, the sum in (2.33) is zero. Thus, expanding Sp_n' in terms of $q_0, q_1, \cdots, q_n$, we obtain (1.3). □

Thus the proof of Theorem 1.1 is complete. We turn to proving Theorem 1.2.

Proof of Theorem 1.2. Suppose that any of the assertions I, II or III of Theorem 1.1 hold, so that, in particular, (1.5) and (1.6) are valid. The partial fraction decomposition of U/V has the form

$$U(u)/V(u) = P'(u) - \sum_{k=1}^{N} \Gamma_k/(u - v_k) + \sum b_{ik}(u - v_i)^{-k-1} + U^*(u)/V^*(u), \tag{2.34}$$

where P is a polynomial, U^* and V^* are as in Theorem 1.2, and in the finite sum $\sum b_{ik}(u-v_i)^{-k-1}$, all the integers k are positive. Integrating (2.34), we obtain (1.13). The stated restrictions on $\{a_{ik}\}$ and P depending on the support of $d\alpha$, follow immediately from the integrability of $d\alpha$. □

REFERENCES

[1] W. A. AL-SALAM AND T. S. CHIHARA, *Another characterization of the classical orthogonal polynomials*, this Journal, 3 (1972), pp. 65-70.

[2] V. BADKOV, *Convergence in the mean and almost everywhere of Fourier series in polynomials orthogonal on an interval*, Math. USSR-Sb., 24 (1974), pp. 223-256.

[3] S. BONAN, *Applications of G. Freud's theory*. I., in Approximation Theory, vol. 4, C. K. Chui et al., eds., Academic Press, New York, 1984, pp. 347-351.

[4] S. BONAN AND P. NEVAI, *Orthogonal polynomials and their derivatives*, I, J. Approx. Theory, 40 (1984), pp. 134-147.

[5] G. FREUD, *Orthogonal Polynomials*, Akademiai Kiado, Pergamon Press, Budapest, 1971.

[6] ———, *On Markov-Bernstein type inequalities and their applications*, J. Approx. Theory, 19 (1977), pp. 22-37.

[7] L. YA. GERONIMUS, *Orthogonal Polynomials*, Consultants' Bureau, New York, 1961.

[8] E. HENDRIKSEN AND H. VAN ROSSUM, *Semiclassical orthogonal polynomials*, in Orthogonal Polynomials and their Applications, C. Brezinski et al., eds., Lecture Notes in Math., Springer, New York, Berlin, 1985.

[9] T. H. KOORNWINDER, *A further generalization of Krall's Jacobi type polynomials*, manuscript.

[10] D. S. LUBINSKY, *On Nevai's bounds for orthogonal polynomials associated with exponential weights*, J. Approx. Theory, 44 (1985), pp. 86-91.

[11] ———, *Estimates of Freud-Christoffel functions for some weights with the whole real line as support*, J. Approx. Theory, 44 (1985), pp. 343-379.

[12] AL. MAGNUS, *On Freud's equations for exponential weights*, J. Approx. Theory, 46 (1986).

[13] P. MARONI, *Une caractérisation des polynômes orthogonaux semi-classiques*, C.R. Acad. Sci. Paris Sér. I Math., 301 (1985), pp. 269-272.

[14] ———, *Prolégomènes a l'étude des polynômes orthogonaux semi-classiques*, to appear.

[15] P. NEVAI, *Orthogonal Polynomials*, Mem. Amer. Math. Soc., 18 (1979), pp. 1-185.

[16] ———, *Lagrange interpolation at zeros of orthogonal polynomials*, in Approximation Theory, Vol. 2, G. G. Lorentz et al., eds., Academic Press, New York, 1976, pp. 163-201.

[17] ———, *Mean convergence of Lagrange interpolation* III, Trans. Amer. Math. Soc., 282 (1984), pp. 669-698.

[18] ———, *Exact bounds for orthogonal polynomials associated with exponential weights*, J. Approx. Theory, 44 (1985), pp. 82-85.

[19] P. NEVAI, *G. Freud, Orthogonal polynomials and Christoffel functions* (*a case study*), J. Approx. Theory, to appear.

[20] A. RONVEAUX, *Polynômes orthogonaux dont les polynômes dérivés sont quasi orthogonaux*, C.R. Acad. Sci. Paris Sér. I Math., 289 (1979), pp. 433-436.

[21] A. ZYGMUND, *Trigonometric Series, Vol.* 2, Cambridge University Press, Cambridge, 1959.

[22] W. HAHN, *Über die Jacobischen Polynome und zwei verwandte polynomklassen*, Math. Z., 39 (1935), pp. 634-638.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

019

ON SIEVED ORTHOGONAL POLYNOMIALS. V: SIEVED POLLACZEK POLYNOMIALS*

JAIRO A. CHARRIS† AND MOURAD E. H. ISMAIL‡

This paper is dedicated to the memory of Jerry Fields, our teacher and our friend.

Abstract. The general Pollaczek polynomials and their sieved analogues are studied in detail. Basic analogues of the Pollaczek polynomials are introduced. The measures that these polynomials are orthogonal with respect to are obtained by applying Darboux's Method and Markov's Theorem.

1. Introduction. A distribution function $\psi(x)$ is a nondecreasing function defined on $(-\infty, \infty)$ and having infinitely many points of increase and finite moments of all orders. A sequence of polynomials $\{p_n(x)\}$ is orthogonal if $p_n(x)$ has precise degree n and there exists a distribution function $\psi(x)$ such that

$$\int_{-\infty}^{\infty} p_n(x)p_m(x)\,d\psi(x) = \lambda_n\delta_{mn}. \tag{1.1}$$

The support of $d\psi(x)$ is the spectrum of ψ and will be denoted by $\sigma(\psi)$. Let

$$p_0(x) = 1, \qquad p_1(x) = A_0x + B_0. \tag{1.2}$$

A set of polynomials $\{p_n(x)\}$ satisfying (1.2) is orthogonal if and only if it satisfies a three term recurrence relation

$$p_{n+1}(x) = (A_nx + B_n)p_n(x) - C_np_{n-1}(x), \qquad n > 0, \tag{1.3}$$

and a positivity condition

$$A_nA_{n-1}C_n > 0, \qquad n = 1, 2, \cdots. \tag{1.4}$$

This is known as Favard's Theorem [9], [30]. We shall normalize $\psi(x)$ by

$$\int_{-\infty}^{\infty} d\psi(x) = 1, \tag{1.5}$$

$$\psi(-\infty) = 0, \qquad \psi(x) = \lfloor \psi(x+0) + \psi(x-0) \rfloor /2. \tag{1.6}$$

It is easy to obtain

$$\lambda_0 = 1, \quad \lambda_n = A_0C_1 \cdots C_n/A_n, \quad n = 1, 2, \cdots. \tag{1.7}$$

A central problem in the theory of orthogonal polynomials is to determine the qualitative behavior of a distribution function from the qualitative behavior of the coefficients in the three term recurrence relation [10], [18]-[21]. These qualitative results are usually motivated by specific examples, or models, where both the polynomials and the distribution function are known explicitly. These models are usually hard to come by so new models that exhibit distinctly different qualitative behavior are of some interest and importance. The Szegö theory of orthogonal polynomials is modeled after the Chebyshev polynomials [10]-[12], [30]. Recently several good candidates of models for new theories have been discovered [4], [15].

* Received by the editors June 10, 1985; accepted for publication May 20, 1986.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287. The work of this author was supported in part by Arizona State University and the National University of Colombia, Bogotá, Colombia, South America.

‡ Department of Mathematics, Arizona State University, Tempe, Arizona 85287. The work of this author was supported in part by a grant from the National Science Foundation.

The second order difference equation (1.3) has two linearly independent solutions $\{p_n(x)\}$ and $\{p_n^*(x)\}$. The denominator polynomials $\{p_n(x)\}$ satisfy the initial conditions (1.2) while the numerator polynomials $\{p_n^*(x)\}$ are given initially by

$$p_0^*(x)=0, \qquad p_1^*(x)=A_0. \tag{1.8}$$

It can be proved that the boundedness conditions

$$|B_n/A_n|\leqq M \quad \text{and} \quad C_n/(A_nA_{n-1})<M, \quad n=1,2,\cdots, \tag{1.9}$$

imply the boundedness of $\sigma(\psi)$ (Chihara [9, pp. 67, 109]). When $\sigma(\psi)$ is bounded, the Stieltjes transform of the distribution function can be recovered from the asymptotic behavior of $p_n(x)$ and $p_n^*(x)$.

THEOREM 1.1 (Markov). *If $\sigma(\psi)$ is compact then*

$$\lim_{n\to\infty} p_n^*(z)/p_n(z)=\int_{-\infty}^{\infty}\frac{d\psi(t)}{z-t}, \qquad z\notin\sigma(\psi), \tag{1.10}$$

and the limit is uniform on compact subsets of $\mathbb{C}-\sigma(\psi)$.

The Perron-Stieltjes Inversion Formula is

$$F(z)=\int_{-\infty}^{\infty}\frac{d\psi(t)}{z-t} \quad \text{iff } \psi(t_2)-\psi(t_1)=\lim_{\varepsilon\to 0+}\int_{t_1}^{t_2}\frac{F(t-i\varepsilon)-F(t+i\varepsilon)}{2\pi i}\,dt. \tag{1.11}$$

In (1.11) it is assumed that $\sigma(\psi)$ is contained in a half line. The asymptotic behavior of $p_n(z)$ and $p_n^*(z)$ may be computed by applying Darboux's Method (Olver [23, § 8.9]).

THEOREM 1.2 (Darboux's Method). *Let $f(z)$ be analytic in $|z|<r$, $0<r<\infty$, and have a finite number of singularities on $|z|=r$. Assume that $g(z)$ is also analytic in $|z|<r$ and $f-g$ is continuous in $|z|=r$. If $f(z)=\sum_0^\infty f_nz^n$, $g(z)=\sum_0^\infty g_nz^n$, then $f_n=g_n+o(r^{-n})$.*

The function $g(z)$ is called a comparison function. The origin of the terminology "numerator" and "denominator" is that the continued fraction

$$\chi(z):=\frac{A_0|}{|A_0z+B_0}-\frac{C_1|}{|A_1z+B_1}-\cdots-\frac{C_n|}{|A_nz+B_n}-\cdots$$

is given by

$$\chi(z)=\lim_{n\to\infty}p_n^*(z)/p_n(z), \qquad z\notin\sigma(\psi).$$

Pollaczek's investigations of a stochastic model of the French telephones led him to a generalization of the Legendre polynomials [24]. The ultraspherical polynomials are generated by

$$C_0^\lambda(x)=1, \qquad C_1^\lambda(x)=2\lambda x, \tag{1.12}$$

$$(n+1)C_{n+1}^\lambda(x)=2x(n+\lambda)C_n^\lambda(x)-(n+2\lambda-1)C_{n-1}^\lambda(x), \qquad n>0. \tag{1.13}$$

The case $\lambda=\frac{1}{2}$ is the Legendre polynomials. Later, Szegö [29] extended Pollaczek's work by generalizing the ultraspherical polynomials in the way Pollaczek generalized the Legendre polynomials. Szegö considered the polynomials

$$(n+1)P_{n+1}^\lambda(x;a,b)=2[x(n+\lambda+a)+b]P_n^\lambda(x;a,b)-(n+2\lambda-1)P_{n-1}^\lambda(x;a,b) \tag{1.14}$$

with

$$P_0^\lambda(x;a,b)=1, \qquad P_1^\lambda(x;a,b)=2x(\lambda+a)+2b. \tag{1.15}$$

Clearly $P_n^\lambda(x; 0, 0) = C_n^\lambda(x)$. Szegö referred to $\{P_n^\lambda(x; a, b)\}$ as the Pollaczek polynomials. They do not belong to the Szegö class and the limiting distribution of their zeros is very different from the classical polynomials (Novikoff [22]). Both Pollaczek and Szegö only handled the case $\lambda > 0$, $a > |b|$, when ψ is absolutely continuous and $\sigma(\psi)$ is $[-1, 1]$. Askey and Ismail [4] treated the case $b = 0$ of the polynomials in (1.15) and (1.14). The case $b = 0$ contains subcases when the distribution function has infinitely many jumps. The jumps and the polynomials have been computed explicitly. This provides a model very different from the case $\lambda > 0$, $a > |b|$ of Pollaczek and Szegö. This raises the question of determining the distribution function of the general Pollaczek polynomials. This will follow as a corollary of some of the results of this work, see § 6. The special cases $b = \pm a$ were treated in Bank and Ismail [7]. This is equivalent to determining the spectral measure of the differential operator

$$-\frac{1}{2}\frac{d^2}{dr^2} + \frac{l(l+1)}{2r^2} + \frac{Z}{r},$$

which is the radial part of a Schrödinger wave equation with a Coulomb potential.

L. J. Rogers introduced continuous q-ultraspherical polynomials $\{C_n(x; \beta \mid q)\}$ and used them to prove the celebrated Rogers-Ramanujan identities. They satisfy

$$(1.16) \qquad (1 - q^{n+1})C_{n+1}(x; \beta \mid q) = 2x(1 - \beta q^n)C_n(x; \beta \mid q) - (1 - \beta^2 q^{n-1})C_{n-1}(x; \beta \mid q),$$

for $n > 0$ and

$$(1.17) \qquad C_0(x; \beta \mid q) = 1, \qquad C_1(x; \beta \mid q) = 2x(1 - \beta)/(1 - q),$$

where $-1 < q < 1$. The distribution function of these polynomials was computed in [2], [3], [6]. These polynomials generalize the ultraspherical polynomials in a direction different from Szegö's generalization. It is clear that $C_n(x; q^\lambda \mid q) \to C_n^\lambda(x)$ as $q \to 1$.

Al-Salam, Allaway and Askey [1] observed that other interesting polynomials arise as limiting cases of the continuous q-ultraspherical polynomials. They showed that

$$(1.18) \qquad B_n^\lambda(x; k) = \lim_{s \to 1} C_n(x; s^{\lambda k + 1}\omega \mid s\omega), \qquad \omega = \exp(2\pi i/k)$$

exists and satisfies

$$(1.19) \qquad \begin{cases} B_{n+1}^\lambda(x; k) = 2xB_n^\lambda(x; k) - B_{n-1}^\lambda(x; k), & k \nmid n+1, \\ mB_{mk}^\lambda(x; k) = 2x(m + \lambda)B_{mk-1}^\lambda(x; k) - (m + 2\lambda)B_{mk-2}^\lambda(x; k), & m > 0, \end{cases}$$

$$(1.20) \qquad B_0^\lambda(x; k) = 1, \quad B_1^\lambda(x; k) = 2x, \quad k > 1.$$

Al-Salam, Allaway and Askey named these polynomials "sieved ultraspherical polynomials of the second kind." They also showed that the sieved ultraspherical polynomials of the first kind

$$(1.21) \qquad c_n^\lambda(x; k) = \lim_{s \to 1} (\omega s; \omega s)_n C_n(x; s^{\lambda k} \mid \omega s)/(s^{2\lambda k}; s\omega)_n$$

satisfy

$$(1.22) \qquad \begin{cases} c_{n+1}^\lambda(x; k) = 2xc_n^\lambda(x; k) - c_{n-1}^\lambda(x; k), & k \nmid n, \\ (m + 2\lambda)c_{mk+1}^\lambda(x; k) = 2x(m + \lambda)c_{m_k}^\lambda(x; k) - mc_{mk-1}^\lambda(x; k), & m > 0, \end{cases}$$

$$(1.23) \qquad c_0^\lambda(x; k) = 1, \qquad c_1^\lambda(x; k) = x, \quad k > 1$$

where ω is as in (1.18). In (1.21) we used the notation

$$(1.24) \qquad (a; q)_0 = 1, \qquad (a; q)_n = \prod_{j=1}^{n} (1 - aq^{j-1}).$$

Al-Salam, Allaway and Askey let $q=\omega s$ and let $s\to 1$ formally in the orthogonality relation of the continuous q-ultraspherical polynomials after choosing β as in (1.18) and (1.21). They mentioned the orthogonality relations

$$\int_{-1}^{1} c_n^\lambda(x;k)c_m^\lambda(x;k)w_1(x)\,dx=\Gamma\left(\frac{1}{2}\right)\Gamma\left(\lambda+\frac{1}{2}\right)h_{n,1}\delta_{m,n}/\Gamma(\lambda+1), \tag{1.25}$$

and

$$\int_{-1}^{1} B_m^\lambda(x;k)B_n^\lambda(x;k)w_2(x)\,dx=\Gamma\left(\frac{1}{2}\right)\Gamma\left(\lambda+\frac{1}{2}\right)h_{n,2}\delta_{m,n}/\Gamma(\lambda+1), \tag{1.26}$$

where

$$w_1(x)=(1-x^2)^{\lambda-1/2}|U_{k-1}(x)|^{2\lambda},\qquad h_{n,1}=\frac{(\lambda)_{\lceil n/k\rceil}(1)_{\lfloor n/k\rfloor}}{(2\lambda)_{\lceil n/k\rceil}(\lambda+1)_{\lfloor n/k\rfloor}}, \tag{1.27}$$

and

$$w_2(x)=(1-x^2)^{\lambda+1/2}|U_{k-1}(x)|^{2\lambda},\qquad h_{n,2}=\frac{(\lambda+1)_{\lfloor n/k\rfloor}(2\lambda+1)_{\lfloor (n+1)/k\rfloor}}{2(1)_{\lfloor n/k\rfloor}(\lambda+1)_{\lfloor (n+1)/k\rfloor}}. \tag{1.28}$$

For proofs, see Askey and Shukla [5], Charris and Ismail [8] and Ismail [14].

This paper is part of a series of papers on sieved orthogonal polynomials, [8], [14]-[16]. In this part, we thoroughly investigate a sieved analogue of the Pollaczek polynomials. We start with an analogue of the continuous q-ultraspherical polynomials. The appropriate analogue is

$$\begin{cases} F_0(x)=1, \quad F_1(x)=2[(1-\Delta U)x+V]/(1-q),\\ (1-q^{n+1})F_{n+1}(x)=2[(1-U\Delta q^n)x+Vq^n]F_n(x)-(1-\Delta^2q^{n-1})F_{n-1}(x), \end{cases} \tag{1.29}$$

$$n>0.$$

We shall also use $F_n(x;U,V,\Delta;q)$ instead of $F_n(x)$ if we need to exhibit the dependence on the parameters U, V, Δ and q. In § 3 we obtain generating functions for $\{F_n(x)\}$ and $\{F_n^*(x)\}$. We then take

$$U=s^{ka},\quad V=\omega s^k(1-s^{kb}),\quad \Delta=\omega s^{k\lambda+1},\quad q=s\omega, \tag{1.30}$$

and let $s\to 1$. Set

$$B_n^\lambda(x;a,b;k)=\lim_{s\to 1}F_n(x;s^{ka},\omega s^k(1-s^{kb}),\omega s^{k\lambda+1};s\omega). \tag{1.31}$$

Writing $B_n^\lambda(x)$ for $B_n^\lambda(x;a,b;k)$ we get

$$B_0^\lambda(x)=1,\qquad B_1^\lambda(x)=2x, \tag{1.32}$$

and

$$\begin{cases} B_{n+1}^\lambda(x)=2xB_n^\lambda(x)-B_{n-1}^\lambda(x), \qquad k\nmid n+1,\\ mB_{mk}^\lambda(x)=2[(m+\lambda+a)x+b]B_{mk-1}^\lambda(x)-(2\lambda+m)B_{mk-2}^\lambda(x), \qquad m>0. \end{cases} \tag{1.33}$$

These are the sieved Pollaczek polynomials of the second kind. The case $b=0$ is in [14] when $\lambda>0$. Generating functions for the B_n^λ's and their numerators follow from the generating functions for the F_n's and F_n^*'s. This is also done in § 3. The q-binomial theorem

$$\frac{(az;q)_\infty}{(z;q)_\infty}=\sum_{n=0}^{\infty}\frac{(a;q)_n}{(q;q)_n}z^n \tag{1.34}$$

(Slater [28, p. 248]) will be used. The asymptotic behavior of $B_n^\lambda(x)$ and $B_n^{*\lambda}(x)$ are determined and the continued fraction whose numerators are $\{B_n^{*\lambda}(x)\}$ and denominators are $\{B_n^\lambda(x)\}$ is computed. Also, in § 4 analogous results are obtained for the sieved Pollaczek polynomials of the first kind which arise as follows. Let $\{G_n(x; U, V, \Delta; q)\}$ or simply $\{G_n(x)\}$ be

$$G_n(x)=\frac{(q;q)_nF_n(x)}{(\Delta^2;q)_n}. \tag{1.35}$$

The G_n's satisfy $G_0(x)=1$, $G_1(x)=2[(1-\Delta U)x+V]/(1-\Delta^2)$ and

$$(1-\Delta^2q^n)G_{n+1}(x)=2[(1-\Delta Uq^n)x+Vq^n]G_n(x)-(1-q^n)G_{n-1}(x), \qquad n>0. \tag{1.36}$$

Now let

$$U=s^{ka}, \quad V=s^k(1-s^{kb}), \quad \Delta=s^{k\lambda}, \quad q=s\omega, \tag{1.37}$$

and define the sieved Pollaczek polynomials of the first kind by

$$c_n^\lambda(x; a, b; k)=\lim_{s\to1} G_n(x; s^{ka}, s^k(1-s^{kb}), s^{k\lambda}; \omega s).$$

They satisfy

$$c_0^\lambda(x; a, b; k)=1, \qquad c_1^\lambda(x; a, b; k)=x, \tag{1.38}$$

and

$$\begin{cases} c_{n+1}^\lambda(x; a, b; k)=2xc_n^\lambda(x; a, b; k)-c_{n-1}^\lambda(x; a, b; k), & k\nmid n \\ (m+2\lambda)c_{mk+1}^\lambda(x; a, b; k)=2[x(m+a+\lambda)+b]c_{mk}^\lambda(x; a, b; k) \\ \qquad\qquad -mc_{mk-1}^\lambda(x; a, b; k), \end{cases} \tag{1.39}$$

for $m>0$. The distribution functions of both polynomials are determined in §§ 4 and 5. It turns out that the discrete spectrum (the closure of the isolated points of discontinuity of $\psi(x)$) is very hard to determine in this generality. The orthogonality relations are stated explicitly at the end of § 5. Section 2 contains a brief survey of the Hadamard integral, a very important technique in determining the leading term in the singular part of a complex valued function defined as a definite integral. This technique is used in §§ 4 and 5. In § 6, we treat the general Pollaczek polynomials. This is achieved by letting $k=1$ in our results on the sieved Pollaczek polynomials of the first kind. The measure that the q-Pollaczek polynomials $\{F_n(x)\}$ are orthogonal with respect to is also found in § 6.

The asymptotic formula

$$\frac{\Gamma(a+n)}{\Gamma(b+n)}\sim n^{a-b} \quad \text{as } n\to\infty \tag{1.40}$$

and the Chu-Vandermonde sum (Rainville [26, p. 69])

$${}_2F_1\left(\left.\begin{matrix}-n, b\\ c\end{matrix}\right|1\right)=\frac{(c-b)_n}{(c)_n} \tag{1.41}$$

will be used in the sequel. In § 5 we shall need the following lemma (Shohat and Tamarkin [27, pp. 45-46]).

LEMMA 1.3. *Let $\{P_n(x)\}$ be an orthonormal polynomial set and assume that the associated moment problem is determined. The corresponding distribution function has a jump at $x=\xi$ if and only if*

$$\sum_0^\infty |P_n(\xi)|^2<\infty. \tag{1.42}$$

2. The Hadamard integral. In this section we study some basic properties of the (simple) Hadamard integral (Hadamard [13]).

We say that an open subset Ω of the complex plane is a branched neighborhood of b if Ω contains a set of the form $D-R_b$, where D is an open disc such that $b\in\bar{D}$ and R_b is a half-line emanating at b and not bisecting D. We will usually assume that Ω is simply connected. Clearly, any open disc is a branched neighborhood of its boundary points. If D is the unit disc, $D-[0,\infty)$ is a branched neighborhood of 0.

Let Ω be a simply connected branched neighborhood of b and assume that ρ is a complex number which is not a negative integer, that $(t-b)^\rho$ is defined in Ω and that $g(t)$ is an analytic function having a power series development $\sum_{n=0}^\infty a_n(b-t)^n$ around b which holds in a neighborhood of $\Omega\cup\{b\}$. We define the Hadamard integral

$$\int_z^{\overline{b|}} (b-t)^\rho g(t)\,dt, \qquad z\in\Omega,$$

by the formula

$$\int_z^{\overline{b|}} (b-t)^\rho g(t)\,dt=\sum_{n=0}^\infty \frac{a_n}{\rho+n+1}(b-z)^{\rho+n+1}. \tag{2.1}$$

It is clear that when $\operatorname{Re}(\rho)>-1$, then

$$\int_z^{\overline{b|}} (b-t)^\rho g(t)\,dt=\int_z^b (b-t)^\rho g(t)\,dt, \tag{2.2}$$

where the integral on the right side is over any curve in Ω joining z and b.

More generally, if Ω' is a simply connected open set containing Ω, and g is analytic in Ω' and has a power series expansion around b which holds in a neighborhood of $\Omega\cup\{b\}$, we define

$$\int_a^{\overline{b|}} (b-t)^\rho g(t)\,dt=\int_a^z (b-t)^\rho g(t)\,dt+\int_z^{\overline{b|}} (b-t)^\rho g(t)\,dt, \qquad a\in\Omega', \tag{2.3}$$

where $z\in\Omega$. Furthermore

$$\int_{\underline{|b}}^a (b-t)^\rho g(t)\,dt=-\int_a^{\overline{b|}} (b-t)^\rho g(t)\,dt. \tag{2.4}$$

If Ω is also a branched neighborhood of a and Ω' is a neighborhood of $\Omega\cup\{a\}$, we define, for $g(t)$ analytic in Ω' and $\rho,\sigma\neq-1,-2,\cdots$,

$$\begin{aligned}\int_{\underline{|a}}^{\overline{b|}} (t-a)^\sigma(b-t)^\rho g(t)\,dt&=\int_{\underline{|a}}^z (t-a)^\sigma(b-t)^\rho g(t)\,dt\\&\quad+\int_z^{\overline{b|}} (t-a)^\sigma(b-t)^\rho g(t)\,dt,\end{aligned} \tag{2.5}$$

where z is any point in Ω'.

The integral $\int_{\underline{|a}}^{\overline{b|}} (t-a)^\sigma(b-t)^\rho g(t)\,dt$ is an extension of the integral $\int_a^b (t-a)^\sigma(b-t)^\rho g(t)\,dt$ from the proper cases $\operatorname{Re}(\sigma)>-1$, $\operatorname{Re}(\rho)>-1$ to the case $\sigma,\rho\neq-1,-2,-3,\cdots$.

The definition of the Hadamard integral can be extended to a function $f(t)$ of the form

$$f(t)=\sum_{n=0}^{\infty} C_n(b-t)^{\rho+n}, \qquad t\in\Omega. \tag{2.6}$$

Let g satisfy the same assumptions as in (2.3). We now define the extended Hadamard integral by

$$\int_a^{\overline{b|}} f(t)g(t)\,dt=\sum_{n=0}^{N} C_n\int_a^{\overline{b|}}(b-t)^{\rho+n}g(t)\,dt+\int_a^b h(t)g(t)\,dt, \tag{2.7}$$

where

$$h(t)=\sum_{n=N+1}^{\infty} C_n(b-t)^{\rho+n}$$

and $\operatorname{Re}(\rho+n)>-1$ for $n>N$. Functions defined by (2.6) are said to have an algebraic branch singularity at $t=b$. When f is given by (2.6), Ω is a branched neighborhood of a, and

$$g(t)=\sum_{n=0}^{\infty} a_n(t-a)^{\sigma+n} \tag{2.8}$$

with $\operatorname{Re}(\sigma)\neq -1,-2,\cdots$, we define

$$\int_{\underline{|a}}^{\overline{b|}} f(t)g(t)\,dt=\int_{\underline{|a}}^{z} f(t)g(t)\,dt+\int_z^{\overline{b|}} f(t)g(t)\,dt, \qquad z\in\Omega'. \tag{2.9}$$

It is not difficult to prove the following.

THEOREM 2.1. *Let f be an analytic function in the simply connected branched neighborhood Ω of the point b, and assume that f has an algebraic branch singularity at b. Let $\{g_n\}$ be a sequence of analytic functions in a neighborhood Ω' of $\Omega\cup\{b\}$ converging uniformly to zero on compact subsets of Ω'. Then, for all $a\in\Omega'$ we have*

$$\lim_{n\to\infty}\int_a^{\overline{b|}} f(t)g_n(t)\,dt=0.$$

COROLLARY 2.2. *Let f, Ω, $\{g_n\}$ and Ω' be as in Theorem 2.1 but assume that $\{g_n\}$ converges to g on compact sets. Then*

$$\lim_{n\to\infty}\int_a^{\overline{b|}} f(t)g_n(t)\,dt=\int_a^{\overline{b|}} f(t)g(t)\,dt. \tag{2.10}$$

COROLLARY 2.3. *Let f, Ω, Ω' be as in the theorem, and assume that*

$$g(t)=\sum_{n=0}^{\infty} a_n(t-a)^n \tag{2.11}$$

holds for $a\in\Omega$ and all $t\in\Omega'$. Then

$$\int_a^{\overline{b|}} f(t)g(t)\,dt=\sum_{n=0}^{\infty} a_n\int_a^{\overline{b|}} f(t)(t-a)^n\,dt. \tag{2.12}$$

Since uniform convergence on compact subsets is sometimes difficult to check, the following corollary is often useful.

COROLLARY 2.4. *Let f, Ω, Ω', $\{g_n\}$ and g be as in Theorem* 2.1, *but assume only that $\{g_n\}$ is uniformly bounded on compact subsets of Ω' and that $\{g_n(t)\}$ converges to $g(t)$ for each t in a subset S of Ω' having a limit point in Ω'. Then*

$$\lim_{n\to\infty}\int_a^{\overline{b}\rceil} f(t)g_n(t)\,dt=\int_a^{\overline{b}\rceil} f(t)g(t)\,dt. \tag{2.13}$$

We now study Hadamard integrals of functions that will arise in this work. These integrals are related to certain analytic functions in the cut plane $\mathbb{C}-[-1,1]$ that we will now introduce.

Let $\sqrt{z+1}$ be the branch of the square root of $z+1$ in $\mathbb{C}-(-\infty,-1]$ that makes $\sqrt{z+1}>0$ if $z>-1$, and $\sqrt{z-1}$ be the branch of the square root of $z-1$ in $\mathbb{C}-(-\infty,1]$ with $\sqrt{z-1}>0$ for $z>1$. Both $\sqrt{z+1}$ and $\sqrt{z-1}$ are single valued in the cut plane $\mathbb{C}-(-\infty,1]$. Let

$$\tau(z)=\sqrt{z+1}\sqrt{z-1},\qquad z\in\mathbb{C}-(-\infty,1]. \tag{2.14}$$

Observe that when $x<-1$ we have

$$\lim_{\substack{y\to0\\ y>0}}\sqrt{x+iy+1}\,\sqrt{x+iy-1}=i\sqrt{-x-1}\cdot i\sqrt{-x+1}=-\sqrt{x^2-1} \tag{2.15}$$

and

$$\lim_{\substack{y\to0\\ y<0}}\sqrt{x+iy+1}\,\sqrt{x+iy-1}=(-i\sqrt{-x-1})\cdot(-i\sqrt{-x+1})=-\sqrt{x^2-1}. \tag{2.16}$$

We now extend τ, by continuity, to the cut plane $\mathbb{C}-[-1,1]$. In order to do so we define

$$\tau(z)=-\sqrt{z^2-1},\qquad z<-1. \tag{2.17}$$

Clearly, $\tau(z)$ is analytic in $\mathbb{C}-[-1,1]$. In what follows we shall simply write

$$\tau(z)=\sqrt{z^2-1}. \tag{2.18}$$

We now define the following analytic functions in $\mathbb{C}-[-1,1]$

$$\alpha(z)=z+\tau(z)=z+\sqrt{z^2-1},\qquad \beta(z)=z-\tau(z)=z-\sqrt{z^2-1} \tag{2.19}$$

and

$$A(z)=-\lambda+\frac{az+b}{\tau(z)}=-\lambda+\frac{az+b}{\sqrt{z^2-1}},\qquad B(z)=-\lambda-\frac{az+b}{\tau(z)}=-\lambda-\frac{az+b}{\sqrt{z^2-1}}. \tag{2.20}$$

Here, a, b, λ are real numbers and

$$\lambda>-\tfrac{1}{2}, \tag{2.21}$$

$$a-\lambda\neq0,1,2,\cdots. \tag{2.22}$$

We note that

$$\alpha(x)=x+\sqrt{x^2-1},\qquad \beta(x)=x-\sqrt{x^2-1}\quad \text{if } x>1, \tag{2.23}$$

$$\alpha(x)=x-\sqrt{x^2-1},\qquad \beta(x)=x+\sqrt{x^2+1}\quad \text{if } x<-1, \tag{2.24}$$

$$A(x)=-\lambda\pm\frac{ax+b}{\sqrt{x^2-1}},\qquad B(x)=-\lambda\mp\frac{ax+b}{\sqrt{x^2-1}},\qquad \pm x>1, \tag{2.25}$$

$$\lim_{y\to0\pm}\tau(x+iy)=\pm i\sqrt{1-x^2},\qquad -1\leqq x\leqq1. \tag{2.26}$$

The following functions are continuous on their domain of definition

$$\tau_+(x+iy)=\begin{cases}\tau(x+iy), & y>0, \quad \tau(x), \quad |x|>1, \quad y=0,\\ i\sqrt{1-x^2}, & |x|\leqq 1, \quad y=0,\end{cases} \tag{2.27}$$

$$\tau_-(x+iy)=\begin{cases}\tau(x+iy), & y<0, \quad \tau(x), \quad |x|>1, \quad y=0,\\ -i\sqrt{1-x^2}, & |x|\leqq 1, \quad y=0,\end{cases} \tag{2.28}$$

$$\alpha_\pm(z)=z+\tau_\pm(z), \qquad \beta_\pm(z)=z-\tau_\pm(z), \tag{2.29}$$

$$A_\pm(z)=-\lambda+\frac{az+b}{\tau_\pm(z)}, \qquad \beta_\pm(z)=-\lambda-\frac{az+b}{\tau_\pm(z)}. \tag{2.30}$$

Observe that for $-1\leqq x\leqq 1$ we have

$$\alpha_-(x)=\overline{\beta_+(x)}, \qquad \overline{\beta_-(x)}=\alpha_+(x), \tag{2.31}$$

and

$$A_-(x)=\overline{B_+(x)}, \qquad \overline{B_-(x)}=A_+(x). \tag{2.32}$$

To simplify the notation we will write when $-1<x<1$

$$\alpha_+(x)=\alpha(x), \quad \beta_+(x)=\beta(x); \quad A_+(x)=A(x), \quad B_+(x)=B(x). \tag{2.33}$$

The following elementary result will be very useful.

LEMMA 2.5. *For each z in $\mathbb{C}$, $\alpha(z)$ and $\beta(z)$ are the solutions of the equation*

$$t^2-2zt+1=0 \tag{2.34}$$

that satisfy

$$\alpha(z)+\beta(z)=2z, \quad \alpha(z)-\beta(z)=2\tau(z)=2\sqrt{z^2-1}, \quad \alpha(z)\beta(z)=1. \tag{2.35}$$

Furthermore, $|\beta(z)|\leqq|\alpha(z)|$, with $|\alpha(z)|=|\beta(z)|$ if and only if $-1\leqq z\leqq 1$.

Now let

$$\Omega=\{z\notin[-1,1]\colon B(z)\neq 1,2,\cdots\}, \qquad \Omega^*=\{z\notin[-1,1]\colon B(z)\neq 0,1,\cdots\}. \tag{2.36}$$

LEMMA 2.6. *For $z\in\Omega$ (respectively $z\in\Omega^*$) and all integers $n\geqq 0$,*

$$\int_0^1(1-u)^{-B(z)}u^n\,du=\frac{n!}{(-B+1)_{n+1}}, \qquad z\in\Omega, \tag{2.37}$$

$$\int_0^1(1-u)^{-B(z)-1}u^n\,du=\frac{n!}{(-B)_{n+1}}, \qquad z\in\Omega^*. \tag{2.38}$$

The next theorem gives a series expansion for a Hadamard integral.

THEOREM 2.7. *For every $z\in\Omega$, define $F(z)$ by*

$$F(z)=\int_0^1\left(1-\frac{\beta^k}{\alpha^k}u\right)^{-A(z)-1}(1-u)^{-B(z)}\,du. \tag{2.39}$$

Then

$$F(z)=\sum_{n=0}^{\infty}\frac{(A+1)_n}{(-B+1)_{n+1}}\left(\frac{\beta^k}{\alpha^k}\right)^n \tag{2.40}$$

and is analytic in Ω.

We observe that

$$F(z)=\frac{-1}{B-1}{}_2F_1\left(\begin{matrix}A+1,1\\-B+2\end{matrix}\,\middle|\,\frac{\beta^k}{\alpha^k}\right). \tag{2.41}$$

For points in Ω^* Theorem 2.7 takes the following form.

THEOREM 2.8. *For each $z\in\Omega^*$, let*

$$G(z)=\int_0^{\rceil 1}\left(1-\frac{\beta^k}{\alpha^k}u\right)^{-A-1}(1-u)^{-B-1}\,du. \tag{2.42}$$

Then the function $G(z)$ is analytic in Ω^* and is given by

$$G(z)=-\frac{1}{B}\sum_{n=0}^{\infty}\frac{(A+1)_n}{(-B+1)_n}\left(\frac{\beta^k}{\alpha^k}\right)^n=-\frac{1}{B}{}_2F_1\left(\begin{matrix}A+1,1\\-B+1\end{matrix}\,\middle|\,\frac{\beta^k}{\alpha^k}\right). \tag{2.43}$$

The next theorem relates a Hadamard beta integral to an ordinary beta integral.

THEOREM 2.9. *For $-1<x<1$, we have*

$$\int_{\lfloor 0}^{\rceil 1}(1-u)^{-B(x)-1}u^{-A(x)-1}\,du=\frac{\Gamma(-A(x))\Gamma(-B(x))}{\Gamma(2\lambda)},\qquad \lambda\neq 0, \tag{2.44}$$

and

$$\int_{\lfloor 0}^{\rceil 1}(1-u)^{-B(x)}u^{-A(x)-1}\,du=\frac{\Gamma(-B(x)+1)\Gamma(-A(x))}{\Gamma(2\lambda+1)}. \tag{2.45}$$

Proof. Note in the first place that $-A-B=2\lambda$. We shall only give a proof of (2.44) because (2.45) can be proved similarly. When $-1<x<1$, we have

$$A(x)=-\lambda-i\frac{ax+b}{\sqrt{1-x^2}},\qquad B(x)=-\lambda+i\frac{ax+b}{\sqrt{1-x^2}}, \tag{2.46}$$

so that $\operatorname{Re}(A(x))=\operatorname{Re}(B(x))=-\lambda$. If $\lambda>0$, (2.44) and (2.45) are just the beta integral. Now, assume $-\frac{1}{2}<\lambda<0$ and $0<z<1$. Clearly

$$\int_{\lfloor 0}^{\rceil 1}(1-u)^{-B-1}u^{-A-1}\,du=\int_{\lfloor 0}^{z}(1-u)^{-B-1}u^{-A-1}\,du+\int_{z}^{\rceil 1}(1-u)^{-B-1}u^{-A-1}\,du. \tag{2.47}$$

By the definition of the Hadamard integral,

$$\int_{\lfloor 0}^{z}(1-u)^{-B-1}u^{-A-1}\,du=z^{-A}\sum_{n=0}^{\infty}\frac{(B+1)_n}{n!}\cdot\frac{z^n}{n-A}. \tag{2.48}$$

For the time being we let λ be a complex number in the domain U given by $\operatorname{Re}(\lambda)>-\frac{1}{2}$, $\lambda\neq 0$. Then, the right side of (2.44) is an analytic function of λ in this domain, and an argument based on (2.48) shows that

$$f(\lambda)=\int_0^z(1-u)^{-B-1}u^{-A-1}\,du$$

is analytic in U. On the other hand,

$$g(\lambda)=\int_z^{\rceil 1}(1-u)^{-B-1}u^{-A-1}\,du=\int_{\lfloor 0}^{1-z}u^{-B-1}(1-u)^{-A-1}\,du$$

is also analytic in U. Since, from (2.47),

$$f(\lambda)+g(\lambda)=\frac{\Gamma(-A)\Gamma(-B)}{\Gamma(2\lambda)}$$

for $\operatorname{Re}(\lambda) > 0$, the above equality also holds in U and, in particular, for $-\frac{1}{2} < \lambda < 0$. This completes the proof of the theorem.

3. Generating functions and asymptotics. We first evaluate the generating function

$$F(x, t) = \sum_{n=0}^{\infty} F_n(x; U, V, \Delta; q) t^n. \tag{3.1}$$

Multiply (1.29) by t^{n+1}, $n = 1, 2, \cdots$, and add the resulting equations to obtain the q-difference equation

$$(t^2 - 2tx + 1) F(x, t) = [\Delta^2 t^2 - 2(U\Delta x - V)t + 1] F(x, qt). \tag{3.2}$$

Now, $\alpha(x)$, $\beta(x)$, as given by (2.19) are roots of

$$t^2 - 2xt + 1 = (1 - t/\alpha)(1 - t/\beta). \tag{3.3}$$

The roots of $\Delta^2 t^2 - 2(U\Delta x - V)t + 1 = 0$ are

$$\xi(x) = \frac{(U\Delta x - V) + \sqrt{(U\Delta x - V)^2 - \Delta^2}}{\Delta^2}, \quad \zeta(x) = \frac{(U\Delta x - V) - \sqrt{(U\Delta x - V)^2 - \Delta^2}}{\Delta^2}. \tag{3.4}$$

Clearly

$$\xi(x)\zeta(x) = \frac{1}{\Delta^2}, \qquad \Delta^2 t^2 - 2(U\Delta x - V)t + 1 = (1 - t/\xi)(1 - t/\xi). \tag{3.5}$$

Iteration of (3.2) gives the functional equation

$$F(x, t) = \frac{(t/\xi; q)_n (t/\zeta; q)_n}{(t/\alpha; q)_n (t/\beta; q)_n} F(x, q^n t), \tag{3.6}$$

whose solution, since $F(x, tq^n) \to F(x, 0) = 1$ as $n \to \infty$, is

$$F(x, t) = \frac{(t/\xi; q)_\infty (t/\zeta; q)_\infty}{(t/\alpha; q)_\infty (t/\beta; q)_\infty}. \tag{3.7}$$

We now determine the generating function of $\{F_n^*(x)\}$, namely

$$F^*(x, t) = \sum_{n=0} F_n^*(x) t^n.$$

In this case, $F^*(x; q^n t^n) \to 0$ as $n \to \infty$, since $F_0^*(x) = 0$.

Replace the F_n's in (1.29) by F_n^*'s then multiply by t^{n+1}, $n = 1, 2, \cdots$ and add to get

$$F^*(x, t) = \frac{2(1 - U\Delta)t}{(1 - t/\alpha)(1 - t/\beta)} + \frac{(1 - t/\xi)(1 - t/\zeta)}{(1 - t/\alpha)(1 - t/\beta)} F^*(x, qt), \tag{3.8}$$

where α, β are as in (3.3) and ξ, ζ as in (3.4). Iterating (3.8) gives

$$F^*(x, t) = 2t(1 - U\Delta) \sum_{n=0}^{\infty} \frac{(t/\xi; q)_n (t/\zeta; q)_n}{(t/\alpha; q)_{n+1} (t/\beta; q)_{n+1}} q^n. \tag{3.9}$$

The generating function (3.9) can be written in the form

$$F^*(x, t) = 2t(1 - U\Delta) F(x, t) \sum_{n=0}^{\infty} \frac{(q^{n+1} t/\alpha; q)_\infty (q^{n+1} t/\beta; q)_\infty}{(q^n t/\xi; q)_\infty (q^n t/\zeta; q)_\infty} q^n.$$

The q-binomial theorem (1.34) implies

$$F^*(x,t)=2t(1-u\Delta)F(x,t)\sum_{m,j=0}^{\infty}\frac{(q\xi/\alpha;q)_j(q\zeta/\beta;q)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\xi}\right)^j\left(\frac{t}{\zeta}\right)^m\sum_{n=0}^{\infty}q^{(m+j+1)n}.$$

Therefore,

$$\begin{aligned}F^*(x,t)=2tF(x,t)&\sum_{m,j=0}^{\infty}\frac{(q^2\xi/\alpha;q)_j(q^2\zeta/\beta;q)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\xi}\right)^j\left(\frac{t}{\zeta}\right)^m\\&\cdot\frac{(1-\xi q/\alpha)}{(1-\xi q^{j+1}/\alpha)}\frac{(1-\zeta q/\beta)}{(1-\zeta q^{j+1}/\beta)}\frac{1-U\Delta}{(1-q^{m+j+1})}.\end{aligned}\tag{3.10}$$

We obtain a generating function for the B_n^λ's by taking

$$U=s^{ka},\quad V=\omega s^k(1-s^{kb}),\quad \Delta=\omega s^{k\lambda+1},\quad q=s\omega,\tag{3.11}$$

and letting $s\to 1$ in (3.7). Let

$$B^\lambda(x,t)=\sum_{n=0}^{\infty}B_n^\lambda(x)t^n=\lim_{s\to 1}F(x,t),\tag{3.12}$$

where U, V, Δ, q are given as in (3.11).

The q-binomial theorem (1.34) yields

$$\frac{(t/\xi;q)_\infty}{(t/\alpha;q)_\infty}=\sum_{n=0}^{\infty}\frac{(\alpha/\xi;q)_n}{(q;q)_n}\left(\frac{t}{\alpha}\right)^n.$$

We write

$$\frac{(\alpha/\xi;q)_n}{(q;q)_n}=\prod_{j=0}^{n-1}\frac{1-q^j\alpha/\xi}{1-q^{j+1}},$$

and examine the behavior as $s\to 1$ of each factor in the above product. Set

$$\eta(s):=\omega(xs^{k\lambda+ka+1}-s^k(1-s^{kb})).\tag{3.13}$$

It is easy to see that when Δ is given by (3.11)

$$\xi=\frac{\eta(s)+\sqrt{\eta^2(s)-\Delta^2}}{\Delta^2},\qquad \zeta=\frac{\eta(s)-\sqrt{\eta^2(s)-\Delta^2}}{\Delta^2}.\tag{3.14}$$

When $s\to 1$, $\eta(s)\to\eta(1)=\omega x$ and $\alpha/\xi\to\omega$. Hence, if $k\nmid j+1$, we obtain

$$\lim_{s\to 1}\frac{1-q^j\alpha/\xi}{1-q^{j+1}}=\frac{1-\omega^{j+1}}{1-\omega^{j+1}}=1.\tag{3.15}$$

We now consider the case $k|j+1$. Let

$$f_j(s):=q^j\alpha/\xi,\qquad j=0,\pm1,\pm2,\cdots,\tag{3.16}$$

i.e.,

$$f_j(s)=\alpha q^j\Delta^2\zeta=\alpha\omega^js^j[\eta(s)-\sqrt{\eta(s)^2-\Delta^2}],\qquad f_j(1)=1.$$

Then

$$f_j'(s)=\alpha j\omega^js^{j-1}[\eta(s)-\sqrt{\eta(s)^2-\Delta^2}]+\alpha\omega^js^j\left[\eta'(s)-\frac{\eta(s)\eta'(s)-\Delta\Delta'}{\sqrt{\eta(s)^2-\Delta^2}}\right],$$

so that

$$f_j'(1)=j+\alpha\omega^j\left[\eta'(1)-\frac{\eta(1)\eta'(1)-\Delta(1)\Delta'(1)}{\sqrt{\eta(1)^2-\Delta(1)^2}}\right].$$

Since $\eta(1)=\omega x$, $\Delta(1)=\omega$ and $\Delta'(1)=\omega(k\lambda+1)$, we have

$$f_j'(1)=j+\alpha\omega^j\frac{\eta'(1)(-\beta)+(k\lambda+1)\omega}{\sqrt{x^2-1}}$$

$$=j-\frac{\omega^j[\eta'(1)-(k\lambda+1)\alpha\omega]}{\sqrt{x^2-1}}.$$

It is easy to see that $\eta'(1)=\omega[(k\lambda+ka+1)x+bk]$ and

(3.17) $$f_j'(1)=k\lambda+j+1-k\frac{ax+b}{\sqrt{x^2-1}},\qquad k|j+1.$$

If $g_j(s)=\omega^{j+1}s^{j+1}$, $j=0,\pm1,\pm2,\cdots$,

$$g_j'(s)=(j+1)\omega^{j+1}s^j,\qquad g_j'(1)=(j+1)\omega^{j+1}.$$

When $k|j+1$, this yields

$$\lim_{s\to1}\frac{1-q^j\alpha/\xi}{1-q^{j+1}}=\frac{f_j'(1)}{g_j'(1)}=\frac{-A(x)+m}{m},\qquad j+1=km,$$

where $A(x)$ is given by (2.20). Similarly

(3.18) $$\lim_{s\to1}\frac{1-q^j\beta/\zeta}{1-q^{j+1}}=\frac{-B(x)+m}{m},\qquad j+1=km,$$

with $B(x)$ given by (2.20). Hence, using the q-binomial theorem, we get

$$\lim_{s\to1}(t/\xi;q)_\infty/(t/\alpha;q)_\infty=\lim_{s\to1}\sum_{\substack{0\le l<k\\ m\ge0}}\frac{(\alpha/\xi;q)_{mk+l}}{(q;q)_{mk+l}}\left(\frac{t}{\alpha}\right)^{mk+l}$$

$$=\sum_{m=0}^{\infty}\frac{(-A+1)_m}{m!}\left(\frac{t}{\alpha}\right)^{mk}\sum_{l=0}^{k-1}\left(\frac{t}{\alpha}\right)^l$$

$$=\sum_{m=0}^{\infty}\frac{(-A+1)_m}{m!}\left(\frac{t}{\alpha}\right)^{mk}\frac{1-(t/\alpha)^k}{1-t/\alpha},$$

so that

$$\lim_{s\to1}(t/\xi;q)_\infty/(t/\alpha;q)_\infty=(1-t/\alpha)^{-1}(1-t^k/\alpha^k)^A.$$

Similarly

$$\lim_{s\to1}(t/\zeta;q)_\infty/(t/\beta;q)_\infty=(1-t/\beta)^{-1}(1-t^k/\beta^k)^B;$$

hence for $x\in C$ and $|t|<|\beta(x)|$ we have

(3.19) $$B^\lambda(x,t)=(1-2xt+t^2)^{-1}(1-t^k/\alpha^k)^A(1-t^k/\beta^k)^B.$$

We now proceed to similarly evaluate the generating function

(3.20) $$B^{*\lambda}(x,t)=\sum_{n=0}^{\infty}B_n^{*\lambda}(x)t^n=\lim_{s\to1}F^*(x,t),$$

with U, V, Δ, q as given by (3.11) and $F^{\lambda}(x, t)$ is as in (3.10). Clearly

$$\lim_{s\to 1}\frac{1-q^{r+1}\xi/\alpha}{1-q^{r}}=\lim_{s\to 1} q\xi/\alpha\,\frac{1-q^{-r-1}\alpha/\xi}{1-q^{-r}}=\lim_{s\to 1}\frac{1-q^{-r-1}\alpha/\xi}{1-q^{-r}}$$

$$=1 \text{ or } \frac{f'_{-r-1}(1)}{g'_{-r-1}(1)}, \quad \text{accordingly as } k\nmid r \text{ or } k\mid r.$$

Here, f_j, g_j are as before. Hence

$$\lim_{s\to 1}\frac{1-q^{r+1}\xi/\alpha}{1-q^{r}}=\frac{A+r/k}{r/k} \text{ or } 1 \quad \text{if } k\mid r \text{ or } k\nmid r. \tag{3.21}$$

Similarly

$$\lim_{s\to 1}\frac{1-q^{r+1}\xi/\beta}{1-q^{r}}=\frac{B+r/k}{r/k} \text{ or } 1 \quad \text{if } k\mid r \text{ or } k\nmid r. \tag{3.22}$$

On the other hand,

$$\lim_{s\to 1}(1-q\xi/\alpha)/(1-q^{1+kr}\xi/\alpha)=\lim_{s\to 1}(1-q^{-1}\alpha/\xi)/(1-q^{-1-kr}\alpha/\xi)$$

and (3.16) imply

$$\lim_{s\to 1}\frac{1-q\xi/\alpha}{1-q^{1+kr}\xi/\alpha}=\frac{f'_{-1}(1)}{f'_{-1-kr}(1)}=\frac{A}{A+r}, \tag{3.23}$$

and similarly

$$\lim_{s\to 1}\frac{1-q\zeta/\beta}{1-q^{1+kr}\zeta/\beta}=\frac{B}{B+r}. \tag{3.24}$$

Clearly

$$\lim_{s\to 1}\frac{1-\xi q/\alpha}{1-q^{m+j+1}}=0 \text{ or } \frac{1}{m+j+1} \quad \text{if } k\nmid m+j+1 \text{ or } k\mid m+j+1,$$

$$\lim_{s\to 1}\frac{1-U\Delta}{1-\xi/\alpha q^{j+1}}=\frac{1-\omega}{1-\omega^{l}} \quad \text{if } j=ks+l, \quad 0<l<k,$$

and

$$\lim_{s\to 1}\frac{1-U\Delta}{1-\zeta/\beta q^{m+1}}=\frac{1-\omega}{1-\omega^{l}} \quad \text{if } m=kr+l, \quad 0<l<k,$$

imply that the summands in (3.10) vanish except in three cases:

(I) $k\mid j$ and $k\mid m$,
(II) $k\mid j$ and $k\mid j+m+1$,
(III) $k\mid m$ and $k\mid j+m+1$.

Since in case (I) $(1-U\Delta)/(1-q^{m+j+1})\to 1$ as $s\to 1$ we get

$$\lim_{s\to 1}\frac{F^{*}}{F}=\Sigma_1(x)+\Sigma_2(x)+\Sigma_3(x), \tag{3.25}$$

where k, j belong to case (I), (II), (III), respectively, in $\Sigma_1, \Sigma_2, \Sigma_3$. Using (3.21), (3.22), (3.23) and (3.24), we obtain

$$(3.26)\qquad \begin{cases} \Sigma_1(x)=2t(1-t^k/\alpha^k)^{-A}(1-t^k/\beta^k)^{-B}, \\ \Sigma_2(x)=-2t^k\alpha^{k-1}B\displaystyle\sum_{j,m}\frac{(A)_j}{j!}\frac{(B+1)_m}{m!}\left(\frac{t}{\alpha}\right)^{kj}\left(\frac{t}{\beta}\right)^{km}\frac{1}{j+m+1}, \\ \Sigma_3(x)=-2t^k\beta^{k-1}A\displaystyle\sum_{j,m}\frac{(A+1)_j}{j!}\frac{(B)_m}{m!}\left(\frac{t}{\alpha}\right)^{kj}\left(\frac{t}{\beta}\right)^{km}\frac{1}{j+m+1}. \end{cases}$$

Therefore

$$\begin{aligned} B^{*\lambda}(x,t)&=\frac{2t}{1-2xt+t^2}+2B^\lambda(x,t) \\ &\cdot\left\{(-B)\alpha^{k-1}\int_0^{t^k}\left(1-\frac{u}{\alpha^k}\right)^{-A}\left(1-\frac{u}{\beta^k}\right)^{-B-1}du\right. \\ &\left.+(-A)\beta^{k-1}\int_0^{t^k}\left(1-\frac{u}{\alpha^k}\right)^{-A-1}\left(1-\frac{u}{\beta^k}\right)^{-B}du\right\}. \end{aligned}$$

Integration by parts gives

$$\begin{aligned} B^{*\lambda}(x,t)&=\frac{2}{t-\alpha}+2B^\lambda(x,t) \\ &\cdot\left\{\beta+A\beta^k(\beta-\alpha)\int_0^{t^k}\left(1-\frac{u}{\alpha^k}\right)^{-A-1}\left(1-\frac{u}{\beta^k}\right)^{-B}du\right\}, \end{aligned}$$

which, after the change of variables $u=vt^k$, becomes

$$\begin{aligned} (3.27)\qquad B^{*\lambda}(x,t)&=\frac{2}{t-\alpha}+2B^\lambda(x,t) \\ &\cdot\left\{\beta+A(\beta-\alpha)\left(\frac{t}{\alpha}\right)^k\int_0^1\left(1-v\frac{t^k}{\alpha^k}\right)^{-A-1}\left(1-v\frac{t^k}{\beta^k}\right)^{-B}dv\right\}. \end{aligned}$$

This is an analytic function of x, t for all x in the cut plane $\mathbb{C}-[-1,1]$ and $|t|<|\beta(x)|$.

We now study the asymptotic behavior of $\{B_n^\lambda(x)\}$ and $\{B_n^{*\lambda}(x)\}$ for large n. We use Darboux's Method, Theorem 1.2. Let $x\in\mathbb{C}-[-1,1]$ be fixed. The generating function $B^\lambda(x,t)$, as given by (3.19), is an analytic function of t, for $|t|<|\beta(x)|$, and has an algebraic branch singularity of order $-B(x)+1$ at $\beta(x)$. The leading term in a comparison function is

$$\begin{aligned} \left(1-\frac{t}{\beta}\right)^{B-1}\lim_{t\to\beta}\frac{B^\lambda(x,t)}{(1-t/\beta)^{B-1}}&=\left(1-\frac{t}{\beta}\right)^{B-1}\left(1-\frac{\beta}{\alpha}\right)^{-1}\left(1-\left(\frac{\beta}{\alpha}\right)^k\right)^A\lim_{t\to\beta}\frac{(1-(t/\beta)^k)^B}{(1-t/\beta)^B} \\ &=k^B\left(1-\frac{\beta}{\alpha}\right)^{-1}\left(1-\frac{\beta^k}{\alpha^k}\right)^A\sum_{n=0}^{\infty}\frac{(-B+1)_n}{n!}\left(\frac{t}{\beta}\right)^n. \end{aligned}$$

Using (1.40) we obtain

$$(3.28)\qquad B_n^\lambda(x)\sim k^B\frac{\alpha}{\alpha-\beta}\left(1-\frac{\beta^k}{\alpha^k}\right)^A\frac{\beta^{-n}}{\Gamma(-B+1)}n^{-B},\qquad n\to\infty.$$

This formula gives the asymptotic behavior of the $B_n^\lambda(x)$'s for $x \in \mathbb{C}-[-1,1]$. For $-1 \leqq x \leqq 1$ then $|\alpha(x)|=|\beta(x)|$, and $\alpha(x), \beta(x)$ are both branched singularities of $B^\lambda(x,t)$. There is an additional difficulty. If $\operatorname{Re}(A)<0$ and $x=\xi_j$, where

$$\xi_j=\cos\left(\frac{j\pi}{k}\right), \qquad 0 \leqq j \leqq k, \tag{3.29}$$

then $(\beta/\alpha)^k=1$ (and also $\beta/\alpha=1$ if $j=0,k$), so that $(1-(\beta/\alpha)^k)^A$ is meaningless (also $(1-\beta/\alpha)^{-1}$ is meaningless if $j=0,k$). This makes it difficult to establish the asymptotic behavior of the $B_n^\lambda(x)$ at $x=\xi_j$. For the sake of completeness, we state the asymptotic behavior of $B_n^\lambda(x)$ for $x \neq \xi_j$, although we will not make any use of it. In § 6 we will study the asymptotic behavior at the points ξ_j.

Assume $x \neq \xi_j$, $j=0,1,2,\cdots,k$. Clearly

$$\begin{aligned} B_n^\lambda(x) &\sim k^B\frac{\alpha}{\alpha-\beta}\left(1-\left(\frac{\beta}{\alpha}\right)^k\right)^A\frac{\beta^{-n}}{\Gamma(-B+1)}n^{-B} \\ &+k^A\frac{\beta}{\beta-\alpha}\left(1-\left(\frac{\alpha}{\beta}\right)^k\right)^B\frac{\alpha^{-n}}{\Gamma(-A+1)}n^{-A} \qquad \text{as } n\to\infty. \end{aligned} \tag{3.30}$$

Recalling that $\overline{\alpha(x)}=\beta(x)$ and $\overline{A(x)}=B(x)$ when $-1\leqq x\leqq 1$, we have that

$$B_n^\lambda(x)\sim 2\operatorname{Re}\left\{k^B\frac{\alpha}{\alpha-\beta}\left(1-\left(\frac{\beta}{\alpha}\right)^k\right)^A\frac{\beta^{-n}n^{-B}}{\Gamma(-B+1)}\right\}. \tag{3.31}$$

This can be put in the form

$$\begin{aligned} B_n^\lambda(x) &\sim \frac{(2k)^{-\lambda}n^\lambda}{\Gamma(-B+1)|(\alpha-\beta)/2i||(\alpha^k-\beta^k)/2i|\lambda} \\ &\times\exp\left(\Phi(x)\arg i\alpha^{-k}\left(\frac{\alpha^k-\beta^k}{2i}\right)\right)\cos\varepsilon_n(x), \end{aligned} \tag{3.32}$$

where $-\pi<\arg z\leqq\pi$,

$$\Phi(x)=\frac{ax+b}{\sqrt{1-x^2}}$$

and

$$\begin{aligned} \varepsilon_n(x)&=n(\arg\alpha)-\Phi(x)\left[\ln\left(\frac{n}{k}\right)+\ln|2\sin(k(\arg\alpha))|\right]-\lambda\arg(i\alpha^{-k}\sin(k\arg\alpha)) \\ &-\arg(\Gamma(-B+1))-\arg(i\alpha^{-1}\sin(\arg\alpha)). \end{aligned}$$

If $x=\cos\Theta$, $(j-1)/k\pi<\Theta<j\pi/k$, $j=1,2,\cdots,k$, we can write (3.32) in the form

$$B_n^\lambda(\cos\Theta)\sim\frac{(2k)^{-\lambda}n^\lambda\cos\varepsilon_n(\Theta)}{|\Gamma(-B+1)|\sin\Theta|\sin k\Theta|^\lambda}\exp\left[\left(j\pi-k\Theta-\frac{\pi}{2}\right)\Phi(\cos\Theta)\right], \tag{3.33}$$

where

$$\Phi(\cos\Theta)=a\cot\Theta+b\csc\Theta,$$

and

$$\varepsilon_n(\Theta)=(n+k\lambda+1)\Theta+\frac{\pi}{2}(\lambda(1-2j)-1)-\Phi(\cos\Theta)\left[\ln\left(\frac{n}{k}\right)+\ln|2\sin k\Theta|\right].$$

We now determine the asymptotic behavior of $B_n^{*\lambda}(x)$ for $x \notin [-1, 1]$. Note that

$$B(x) = B_+(x) = -\lambda + i\Phi(x) = -\lambda + i\frac{ax+b}{\sqrt{1-x^2}}, \qquad -1 \leqq x \leqq 1,$$

implies $\operatorname{Re}(B(x)) = -\lambda < -\frac{1}{2}$ for $-1 \leqq x \leqq 1$. Let $\delta > 0$ be such that $\operatorname{Re}(B(z)) < 1$ for $z \in (x-\delta, x+\delta) \times (0, \delta)$, and call this last set K. For $z \in K$, $\operatorname{Re}(B) > 0$,

$$(3.34) \qquad \left|\left(1-\frac{t^k}{\alpha^k}u\right)^{-A-1}\left(1-\frac{t^k}{\beta^k}u\right)^{-B}\right| \leqq C(1-u)^{-\operatorname{Re}(B)}, \qquad 1 \leqq u < 1,$$

where C depends only on z. Since $\operatorname{Re}(B) < 1$, the function $(1-u)^{-\operatorname{Re}(B)}$ is integrable in $[0, 1]$. The Lebesgue Dominated Convergence Theorem implies

$$(3.35) \qquad \lim_{t \to \beta(z)} \int_0^1 \left(1-\frac{t^k}{\alpha^k}u\right)^{-A-1}\left(1-\frac{t^k}{\beta^k}u\right)^{-B} du = \int_0^1 \left(1-\frac{\beta^k}{\alpha^k}u\right)^{-A-1}(1-u)^{-B}\,du.$$

The relationship (3.35) trivially holds if $\operatorname{Re}(B) \leqq 0$. Hence, the leading term in a comparison function for $B^{*\lambda}(z, t)$ is

$$(3.36) \quad 2\tilde{B}^\lambda(z,t)\left\{\beta + A(\beta-\alpha)\left(\frac{\beta}{\alpha}\right)^k \int_0^1 \left(1-u\frac{\beta^k}{\alpha^k}\right)^{-A-1}(1-u)^{-B}\,du\right\}, \qquad z \in K$$

where $\tilde{B}^\lambda(z, t)$ is the dominant term in a comparison function for $B^\lambda(z, t)$. Thus

$$(3.37) \quad B_n^{*\lambda}(z) - 2B_n^\lambda(z)\left\{\beta + A(\beta-\alpha)\left(\frac{\beta}{\alpha}\right)^k \int_0^1 \left(1-u\frac{\beta^k}{\alpha^k}\right)^{-A-1}(1-u)^{-B}\,du\right\},$$

as $n \to \infty$ and the associated continued fraction is

$$(3.38) \quad \chi(z) = 2\left\{\beta + A(\beta-\alpha)\beta^{2k}\int_0^1 \left(1-\frac{\beta^k}{\alpha^k}u\right)^{-A-1}(1-u)^{-B}\,du\right\}, \qquad z \in K.$$

The right-hand side of (3.38) is analytic on K. The left-hand side of (3.38) fails to be analytic only on the support of $d\phi$, ϕ being the normalized distribution function for $\{B_n^\lambda(x)\}$. From the properties of the Hadamard Integral of § 2

$$\int_0^{\overline{1}\rceil} \left(1-\frac{\beta^k}{\alpha^k}u\right)^{-A-1}(1-u)^{-B}\,du$$

is an analytic continuation of $\int_0^1 (1-(\beta^k/\alpha^k)u)^{-A-1}(1-u)^{-B}\,du$ to the set Ω of (2.36). Thus, we proved that $\Omega \subseteq \mathbb{C} - \sigma(\phi)$ and

$$(3.39) \quad \chi(z) = 2\left\{\beta + A(\beta-\alpha)\beta^{2k}\int_0^{\overline{1}\rceil} \left(1-u\frac{\beta^k}{\alpha^k}\right)^{-A-1}(1-u)^{-B}\,du\right\}, \qquad z \in \Omega.$$

In the next section, we study the behavior of $\chi(z)$ on the set $D = \{z\colon B(z) = 1, 2, \cdots\}$ in order to determine the conditions under which it is part of $\sigma(\phi)$.

We now establish a generating function for the polynomials of the first kind $\{c_n^\lambda(x; a, b; k)\}$. For brevity, we will write $c_n^\lambda(x)$ for $c_n^\lambda(x; a, b; k)$. Recall

$$(3.40) \qquad c_n^\lambda(x) = \lim_{s \to 1} G_n(x; U, V, \Delta; q)$$

where

$$(3.41) \qquad U = s^{ka}, \quad V = s^k(1-s^{kb}), \quad \Delta = s^{k\lambda}, \quad q = s\omega.$$

Now (3.3), (3.4) and (3.7) give

$$G(x, t) := \sum_{n=0}^{\infty} \frac{1-U\Delta q^n}{1-U\Delta} F_n(x) = h(s, t) \frac{(qt/\xi; q)_\infty (qt/\zeta; q)_\infty}{(t/\alpha; q)_\infty (t/\beta; q)_\infty} \tag{3.42}$$

where

$$g(s, t) := 1-\Delta U + (\Delta - U)\Delta t^2 + 2Vt, \qquad h(s, t) := g(s, t)/(1-\Delta U). \tag{3.43}$$

As $s \to 1$, $h(s, t) \to 1 + \lfloor 2bt - (\lambda - a)t^2 \rfloor / (\lambda + a)$. Furthermore,

$$\lim_{s\to 1} \frac{(1-q^j\alpha/\xi)}{1-q^j} = \frac{A+m}{m} \text{ or } 1 \quad \text{if } j = km \text{ or } k \nmid j. \tag{3.44}$$

A calculation similar to what we used to prove (3.9) yields

$$C^\lambda(x, t) = \frac{[\lambda + a + 2bt - (\lambda - a)t^2]}{(1-2xt+t^2)(\lambda + a)} (1-t^k\beta^k)^A (1-t^k\alpha^k)^B, \tag{3.45}$$

where $C^\lambda(x, t)$ is the limit of $G(x, t)$ as $s \to 1$. In fact

$$C^\lambda(x, t) = \sum_{n=0}^{\infty} b_n c_n^\lambda(x) t^n, \tag{3.46}$$

where b_n is the limit of $(1-U\Delta q^n)(\Delta^2; q)_n / [(1-\Delta U)(q; q)_n]$. Thus

$$b_n = \frac{(\lambda + a + 1)_{\lfloor n/k \rfloor} (2\lambda)_{\lceil n/k \rceil}}{\lfloor n/k \rfloor ! (\lambda + a)_{\lceil n/k \rceil}}, \tag{3.47}$$

where $\lfloor x \rfloor$ and $\lceil x \rceil$ denote, respectively, the largest integer $<x$ and the smallest integer $>x$.

From (3.45) and Darboux's Method we readily obtain

$$b_n c_n^\lambda(x) \sim -k^B \frac{B}{\lambda + a} (1-\beta^{2k})^A \frac{\beta^{-n}}{\Gamma(-B+1)} n^{-B}, \qquad n \to \infty, \tag{3.48}$$

holding in the complex plane outside the interval $[-1, 1]$.

We wish to determine the asymptotic behavior of $c_n^{*\lambda}(x)$ in order to find the continued fraction whose denominators are $\{c_n^\lambda(x)\}$. The first step is to derive a generating function for $\{c_n^{*\lambda}(x)\}$. Let

$$f(s, t) := 1 + \Delta^2 t^2 + 2(V - xU\Delta)t, \tag{3.49}$$

$$G^*(x, t) := \sum_{n=0}^{\infty} \frac{1-U\Delta q^n}{1-U\Delta} F_n^*(x) t^n. \tag{3.50}$$

Formulas (3.9), (3.42), (3.43), (3.49) and (3.50) lead to

$$G^*(x, t) = \frac{2t}{f(s, t)} + G(x, t) \frac{F^*(x, t)}{F(x, t)}. \tag{3.51}$$

Letting $s \to 1$ in (3.51) and taking (3.41) into account we obtain

$$\sum_{n=0}^{\infty} b_n c_n^{*\lambda}(x) t^n = \frac{2t}{1-2xt+t^2} + C^\lambda(x, t) \lim_{s\to 1} \frac{F^*(x, t)}{F(x, t)}, \tag{3.52}$$

and $C^\lambda(x, t)$ and b_n are given by (3.45) and (3.47), respectively. The following lemma generalizes results of [1] and [14].

LEMMA 3.1. *With $F(x, t)$ given by* (3.7) *and $F^*(x, t)$ by* (3.9), *we have*

$$\lim_{s\to 1}\frac{F^*(x,t)}{F(x,t)}=2(\lambda+a)\frac{\alpha^k-\beta^k}{\alpha-\beta}\int_0^{t^k}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B-1}\,du. \tag{3.53}$$

Proof. Using

$$F^*(x,t)=2tF(x,t)\sum_{m,j=0}^{\infty}\frac{(q\xi/\alpha;q)_j(q\zeta/\beta;q)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\xi}\right)^j\left(\frac{t}{\zeta}\right)^m\frac{1-U\Delta}{1-q^{m+j+1}},$$

we easily see that

$$\lim_{s\to 1}\frac{1-q^j\xi/\alpha}{1-q^j}=1\left(\text{or }\frac{A+r}{r}\right)\text{ if } k\nmid j \text{ (or } j=kr),$$

$$\lim_{s\to 1}\frac{1-q^m\zeta/\beta}{1-q^m}=1\left(\text{or }\frac{B+r}{r}\right)\text{ accordingly as } k\nmid m \text{ or } m=kr.$$

Furthermore,

$$\lim_{s\to 1}\frac{1-U\Delta}{1-q^{m+j+1}}\text{ is 0 if } k\nmid j+m+1 \text{ and is } \frac{\lambda+a}{r} \text{ if } j+m+1=kr.$$

Using the above limiting relationships we see that

$$\lim_{s\to 1}\frac{F^*(x,t)}{F(x,t)}=2k(\lambda+a)\sum_{\substack{j,m=0\\ k|j+m+1}}^{\infty}\frac{(A+1)_{\lfloor j/k\rfloor}(B+1)_{\lfloor m/k\rfloor}}{\lfloor j/k\rfloor!\,\lfloor m/k\rfloor!\,(m+j+1)}t^{j+m+1}\alpha^{m-j}.$$

Let $j=kj_1+l$, $m=km_1+k-l-1$, where $0\leqq l<k$, $j_1\geqq 0$, $m_1\geqq 0$. Then

$$\lim_{s\to 1}\frac{F^*(x,t)}{F(x,t)}=2(\lambda+a)\sum_{j_1,m_1=0}^{\infty}\frac{(A+1)_{j_1}(B+1)_{m_1}}{j_1!\,m_1!\,(j_1+m_1+1)k}t^{(m_1+j_1+1)k}$$
$$\cdot\,\beta\alpha^{(m_1-j_1+1)k}\sum_{l=0}^{k-1}\beta^{2l}.$$

Therefore

$$\lim_{s\to 1}\frac{F^*(x,t)}{F(x,t)}=2(\lambda+a)\frac{\alpha^k-\beta^k}{\alpha-\beta}\int_0^{t^k}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B-1}\,du.$$

This proves the lemma.

This shows that a generating function of the numerator polynomials is

$$\begin{aligned}\sum_{n=0}^{\infty}b_nc_n^{*\lambda}(x)t^n={}&\frac{2t}{t^2-2xt+1}+2(\lambda+a)\frac{\alpha^k-\beta^k}{\alpha-\beta}\left\{1+\frac{2b}{\lambda+a}t-\frac{\lambda-a}{\lambda+a}t^2\right\}\\ &\cdot\frac{(1-t^k\beta^k)^A(1-t^k\alpha^k)^B}{t^2-2xt+1}\int_0^{t^k}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B-1}\,du.\end{aligned}\tag{3.54}$$

Integration by part gives

$$\begin{aligned}\int_0^{t^k}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B-1}\,du={}&\frac{\beta^k}{B}(1-t^k\beta^k)^{-A-1}(1-t^k\alpha^k)^{-B}-\frac{\beta^k}{B}\\ &-\beta^{2k}\frac{A+1}{B}\int_0^{t^k}(1-u\beta^k)^{-A-2}(1-u\alpha^k)^{-B}\,du.\end{aligned}$$

An argument similar to the one used to prove (3.38) gives

$$\lim_{t\to\beta}\int_0^{t^k}(1-u\beta^k)^{-A-2}(1-u\alpha^k)^{-B}\,du=\int_0^{\overline{\beta^k}\rceil}(1-u\beta^k)^{-A-2}(1-u\alpha^k)^{-B}\,du.$$

The above calculations, (3.54) and Darboux's Method establish

$$b_nc_n^{*\lambda}(x)\sim 2(\lambda+a)\frac{(\beta^k-\alpha^k)}{(\beta-\alpha)}\frac{\beta^k}{B}\left\{1+\frac{A+1}{\alpha^k}\int_0^{\overline{\beta^k}\rceil}(1-u\beta^k)^{-A-2}(1-u\alpha^k)^{-B}\,du\right\}b_nc_n^{\lambda}(x),$$

as $n\to\infty$, when $x\in\mathbb{C}-[a,b]$. This proves that the continued fraction is

$$\chi^*(z)=-2\frac{\lambda+a}{\alpha-\beta}\left(1-\frac{\beta^k}{\alpha^k}\right)\frac{1}{B}\left\{1+\frac{(A+1)}{\alpha^k}\int_0^{\overline{\beta^k}\rceil}(1-u\beta^k)^{-A-2}(1-u\alpha^k)^{-B}\,du\right\},\tag{3.55}$$

$x\in\mathbb{C}-[a,b]$, which can also be written

$$\chi^*(z)=-2\frac{(\lambda+a)}{\alpha-\beta}\left(1-\frac{\beta^k}{\alpha^k}\right)\frac{1}{B}\left\{1+(A+1)\frac{\beta^k}{\alpha^k}\int_0^{\overline{1}\rceil}\left(1-u\frac{\beta^k}{\alpha^k}\right)^{-A-2}(1-u)^{-B}\,du\right\}.\tag{3.56}$$

This is the form of the continued fraction for the polynomials $c_n^\lambda(x)$ that we will use henceforth. It is an analytic function of z in the subset Ω^* of $\mathbb{C}-[-1,1]$ where $B(z)\neq 0,1,2,\cdots$, i.e., its only possible singularities outside $[-1,1]$ are on the set $D^*=\{z\colon B(z)=0,1,2,\cdots\}$.

4. The continued fractions $\chi(z)$ and $\chi^*(z)$. The continued fractions $\chi(z)$ and $\chi^*(z)$ are analytic function of z in the cut plane $\mathbb{C}-[-1,1]$, except possibly on the sets D and D^* where $B(z)$ is, respectively, a positive and a nonnegative integer. We now identify these sets and the type of singularities $\chi(z)$ and $\chi^*(z)$ have on such sets.

Let Ω be, as in (2.36), the set of points in $\mathbb{C}-[-1,1]$ where $B(z)\neq 1,2,\cdots$, and Ω^* be the subset of Ω where $B(z)\neq 0$. The function $\chi(z)$ is analytic in Ω and

$$\begin{aligned}\int_0^{\overline{1}\rceil}(1-u\beta^{2k})^{-A-1}(1-u)^{-B}\,du&=(1-\beta^{2k})^{-A-1}\sum_{n=1}^{\infty}\frac{(A+1)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^n\cdot\frac{1}{n+1-B}\\&=-\frac{\alpha^k}{A\beta^k}(1-\beta^{2k})^{-A}\sum_{n=1}^{\infty}\frac{(A)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^n\frac{n}{n-B}\end{aligned}$$

for $z\in\Omega$. Therefore

$$\chi(z)=2\left\{\beta+(\alpha-\beta)(1-\beta^{2k})^{-A}\sum_{n=1}^{\infty}\frac{(A)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^n\frac{n}{n-B}\right\}.\tag{4.1}$$

This series representation of $\chi(z)$ will be used to determine the type of the singularities of $\chi(z)$ in D. For $z\in\Omega^*$ we have

$$\begin{aligned}&(A+1)\int_0^{\overline{1}\rceil}(1-u\beta^{2k})^{-A-2}(1-u)^{-B}\,du\\&=(1-\beta^{2k})^{-A-2}\sum_{n=1}^{\infty}\frac{(A+1)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^{n-1}\cdot\left(\frac{n}{n-B}\right),\end{aligned}$$

which leads to the series representation

$$\begin{aligned}\chi^*(z)=&-\frac{2(\lambda+a)}{(\alpha-\beta)B}\left(1-\frac{\beta^k}{\alpha^k}\right)\\&\cdot\left\{1-\left(1-\frac{\beta^k}{\alpha^k}\right)^{-A-1}\sum_{n=1}^{\infty}\frac{(A+1)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^n\frac{n}{n-B}\right\}.\end{aligned}\tag{4.2}$$

We now analyze the singularities of $\chi(z)$ and $\chi^*(z)$. Markov's Theorem gives

$$(4.3) \qquad \chi(z)=\int_{-\infty}^{+\infty}\frac{d\phi(t)}{z-t}\,dt, \qquad \chi^*(z)=\int_{-\infty}^{+\infty}\frac{d\psi(t)}{z-t},$$

where $\phi(t)$, $\psi(t)$ are, respectively, the normalized distribution functions of $\{B_n^\lambda(x)\}$ and $\{c_n^\lambda(x)\}$. From Markov's Theorem it also follows that the set of singularities of $\chi(z)$ (respectively, $\chi^*(z)$) coincides with the support $\sigma(\phi)$ of $d\phi$ (respectively, $\sigma(\psi)$ of $d\psi$). Hence, the singularities of both $\chi(z)$ and $\chi^*(z)$ are real numbers. The singularities of $\chi(z)$ are located on $[-1,1]$ and on the set $D=\{z\in\mathbb{C}-[-1,1]\colon B(z)=1,2,\cdots\}$, and those of $\chi^*(z)$, on $[-1,1]$ and on the set $D^*=\{z\in\mathbb{C}-[-1,1]\colon B(z)=0,1,2,\cdots\}$. It follows that both D and D^* are subsets of $\mathbb{R}$. In fact $D\subseteq D^*\subseteq(-\infty,-1)U(1,+\infty)$. Furthermore $\sigma(\phi)\subseteq DU[-1,1]$ and $\sigma(\psi)\subseteq D^*U[-1,1]$ hold. To determine if $D\subseteq\sigma(\phi)$, or if $D^*\subseteq\sigma(\psi)$, we have to determine the sets of D and D^* and the nature of the singularities $\chi(z)$, and $\chi^*(z)$ have on these sets. We will do this in this part of the paper, leaving for § 5 the study of the singularities of $\chi(z)$ and $\chi^*(z)$ in $[-1,1]$.

Recall that $B(x)$ is defined in (2.20) and (2.25). To determine D and D^* we solve the equations $B(x)=n$. In doing so, we will first consider the case $n+\lambda>0$; i.e., we will assume $n\geqq 1$ in the general case $\lambda>-\frac{1}{2}$ (this determines D) and also allow $n=0$ if $\lambda>0$. The solutions of $B(x)=n$ satisfy

$$(4.4) \qquad (a^2-(n+\lambda)^2)x^2+2abx+b^2+(n+\lambda)^2=0, \qquad x\neq\pm1.$$

Let

$$(4.5)\quad \Delta_n=(n+\lambda)^2+b^2-a^2,\quad x_n=\frac{-ab+(n+\lambda)\sqrt{\Delta_n}}{a^2-(n+\lambda)^2},\quad y_n=\frac{-ab-(n+\lambda)\sqrt{\Delta_n}}{a^2-(n+\lambda)^2}.$$

The solutions of the quadratic equation in (4.4) are $x=x_n$, $x=y_n$. Those solutions contain the solutions of $B(x)=n$.

In order to obtain a sufficient condition for x_n, y_n to belong to D^*, observe that

$$(4.6)\quad ax_n+b=(n+\lambda)\frac{a\sqrt{\Delta_n}-b(n+\lambda)}{a^2-(n+\lambda)^2},\qquad ay_n+b=-(n+\lambda)\frac{a\sqrt{\Delta_n}+b(n+\lambda)}{a^2-(n+\lambda)^2},$$

$$(4.7)\qquad \sqrt{x_n^2-1}=\pm\frac{a\sqrt{\Delta_n}-b(n+\lambda)}{a^2-(n+\lambda)^2},\qquad \sqrt{y_n^2-1}=\pm\frac{a\sqrt{\Delta_n}+b(n+\lambda)}{a^2-(n+\lambda)^2}.$$

The fact that the numbers in (4.7) are positive numbers when x_n, $y_n\in D^*$ will be important in our considerations. Now, if $B(x_n)=n$ and $x_n<-1$, then $ax_n+b=(n+\lambda)\sqrt{x_n^2-1}$ and (4.6) yields

$$(4.8)\qquad \sqrt{x_n^2-1}=\frac{[a\sqrt{\Delta_n}-b(n+\lambda)]}{[a^2-(\lambda+n)^2]}.$$

Similarly, when $B(x_n)=n$ but $x_n>1$, then

$$(4.9)\qquad \sqrt{x_n^2-1}=\frac{[-a\sqrt{\Delta_n}+b(n+\lambda)]}{\lfloor a^2-(n+\lambda)^2\rfloor}.$$

We prove in the same manner that if $B(y_n)=n$ then

$$(4.10)\qquad \frac{ay_n+b}{\sqrt{y_n^2-1}}=n+\lambda,\qquad \sqrt{y_n^2-1}=-\frac{a\sqrt{\Delta_n}+b(n+\lambda)}{a^2-(n+\lambda)^2}\quad \text{if } y_n<-1,$$

$$(4.11)\qquad -\frac{ay_n+b}{\sqrt{y_n^2-1}}=n+\lambda,\qquad \sqrt{y_n^2-1}=\frac{a\sqrt{\Delta_n}+b(n+\lambda)}{a^2-(n+\lambda)^2}\quad \text{if } y_n>1.$$

Now we give a condition guaranteeing that $x_n < -1$ and $y_n > 1$. An obvious necessary condition is that $\Delta_n \geqq 0$.

LEMMA 4.1. *Assume $n+\lambda > 0$, $a < n+\lambda$. Then $x_n < -1$ for all $b \neq a$ and $y_n > 1$ for all $b \neq -a$.*

Proof. The above conditions make $(n+\lambda)^2 - a^2 > 0$. In fact $a+n+\lambda > 0$ follows from the positivity condition and $n+\lambda-a > 0$ by assumption. Thus $\Delta_n > 0$ for all $b \in \mathbb{R}$. Consider x_n, y_n as functions of the parameter b. Both $x_n(b)$ and $y_n(b)$ are continuous functions of b for all $b \in \mathbb{R}$ and $x_n(b) = -1$ if and only if $b = a$, $x_n(b) = 1$ if and only if $b = -a$. Since $x_n(a) = 1$ and $x_n(b)$ cannot take values in $(-1, 1)$, it follows that $x_n(b) < -1$ for all $b \in \mathbb{R}$, $b \neq a$. From $y_n(-a) = 1$ we conclude that $y_n(b) > 1$ for all $b \in \mathbb{R}$, $b \neq -a$. This proves the lemma.

Remark 4.2. The assumptions of the lemma are automatically satisfied if $a < 0$.

The next theorem characterizes the solutions of $B(x) = n$ among $\{x_n, y_n\}$.

THEOREM 4.3. *Assume $n+\lambda > 0$, $a < n+\lambda$. Then, $x_n \in D^*$ if and only if $b \neq a$ and*

$$a\sqrt{\Delta_n} - b(n+\lambda) < 0,$$

and, under these conditions $x_n < -1$. Furthermore, $y_n \in D^*$ if and only if $b \neq -a$, and

$$a\sqrt{\Delta_n} + b(n+\lambda) < 0,$$

in which case $y_n > 1$.

Proof. If $n+\lambda > 0$, $a < n+\lambda$ we have, by Lemma 4.1, $x_n \leqq -1$ with $x_n = -1$ if and only if $b = a$. Similarly $y_n \geqq 1$, with $y_n = 1$ if and only if $b = -a$. Hence, if $x_n \in D^*$ we must have $b \neq a$ and, since

$$\sqrt{x_n^2 - 1} = \frac{a\sqrt{\Delta_n} - b(n+\lambda)}{a^2 - (n+\lambda)^2}, \qquad a^2 - (n+\lambda)^2 < 0,$$

we must have $a\sqrt{\Delta_n} - b(n+\lambda) < 0$. In the same manner we see that if $y_n \in D^*$ then $b \neq -a$ and

$$\sqrt{y_n^2 - 1} = \frac{a\sqrt{\Delta_n} + b(n+\lambda)}{a^2 - (b+\lambda)^2},$$

then $a\sqrt{\Delta_n} + b(n+\lambda) < 0$. Conversely, if $b \neq a$ we have $x_n < -1$; and if $a\sqrt{\Delta_n} - b(n+\lambda) < 0$, $\sqrt{x_n^2 - 1}$ is given by (4.7). Then we immediately obtain that $B(x_n) = n$, and thus $x_n \in D^*$. In the same way we prove that if $b \neq -a$ then $y_n > 1$, and if $a\sqrt{\Delta_n} + b(n+\lambda) < 0$ then $\sqrt{y_n^2 - 1}$ is given by (4.11). Thus $B(y_n) = n$ and $y_n \in D^*$. This completes the proof.

Remark 4.4. Even under the conditions of Theorem 4.3, $x = -1$ (respectively, $y = 1$) is not in D^* if $a = b$ (respectively, if $b = -a$). They are not even solutions of $B(x) = n$. However, $x_n(-a)$ and $y_n(a)$ are in D^*.

Now let

$$X_n(b) = a\sqrt{\Delta_n} - b(n+\lambda), \qquad Y_n(b) = a\sqrt{\Delta_n} + b(n+\lambda). \tag{4.12}$$

Theorem 4.3 points out the importance of knowing the signs of $X_n(b)$ and $Y_n(b)$. The sign of $X_n(b)$ and $Y_n(b)$ can often be determined from their asymptotic behavior as $b \to \pm\infty$, which is, in general, easier than struggling with the inequalities. This is so because $X_n(b) = 0$ if and only if $b = a$, while $Y_n(b) = 0$ if and only if $b = -a$. We can say a little more. In fact, in any interval I in which $\Delta_n > 0$, we have

$$X_n'(b) = \frac{ab}{\sqrt{\Delta_n}} - n + \lambda, \qquad Y_n'(b) = \frac{ab}{\sqrt{\Delta_n}} + (n+\lambda),$$

so $X_n'(b)\neq 0$, $Y_n'(b)\neq 0$ for all $b\in I$. It follows that X_n and Y_n are strictly monotonic in I. For example, if $I=(a,+\infty)$ and $\Delta_n\geqq 0$ for all $b\in I$, we can deduce that the sign of $X_n(b)$ will be that of $a-n-\lambda$ for all $b\in I$. In fact

$$X_n(b)\sim b(a-b-\lambda)\quad \text{as } b\to+\infty$$

and $X_n(b)$ does not change sign in $(a,+\infty)$. Similarly we conclude that the sign of $Y_n(b)$ in $(-\infty,-\mathrm{a})$, agrees with the sign of $a-n-\lambda$ (provided that $\Delta_n\geqq 0$ in $(-\infty,-a)$).

We now establish two corollaries to Theorem 4.3.

COROLLARY 4.5. *Assume* $0\leqq a<n+\lambda$. *If* $b>a$, *then* $x_n<-1$ *and is in* D^* *but* y_n *is not a solution of* $B(x)=n$.

Proof. Clearly $a\sqrt{\Delta_n}-b(n+\lambda)\sim b(a-n-\lambda)$ as $b\to+\infty$ and does not change sign in $(a,+\infty)$. Thus $a\sqrt{\Delta_n}-b(n+\lambda)<0$. On the other hand, $a\sqrt{\Delta_n}+b(n+\lambda)>0$. By Theorem 4.3, $x_n\in D^*$ and y_n is not a solution of $B(x)=n$.

COROLLARY 4.6. *Let* $n+\lambda>0$, $a<0$, $b\geqq 0$. *Then*

(i) *if* $-b\leqq a$, $x_n<-1$ *and belongs to* D^* *but* y_n *is not a solution of* $B(x)=n$;

(ii) *if* $a<-b$, *both* x_n *and* y_n *are in* D^*, $x_n<-1$ *and* $y_n>1$.

Proof. (i) If $-b<a$, we have $a<n+\lambda$, $a\sqrt{\Delta_n}-b(n+\lambda)<0$, $a\sqrt{\Delta_n}+b(n+\lambda)>0$. By Theorem 4.3, $x_n\in D^*$ and y_n does not satisfy $B(x)=n$. If $a=-b$ we still have $a\sqrt{\Delta_n}-b(n+\lambda)=2a(n+\lambda)<0$, and $x_n\in D^*$. On the other hand, $y_n=1\notin D^*$.

(ii) In this case $a<n+\lambda$, $a\sqrt{\Delta_n}-b(n+\lambda)<0$, $a\sqrt{\Delta_n}+b(n+\lambda)<0$. By Theorem 4.3, x_n, $y_n\in D^*$. This completes the proof.

THEOREM 4.7. *Assume* $0<n+\lambda<a$. *Then, for all* $b>a$, $x_n<-1$ *and is in* D^* *but* y_n *is not a solution of* $B(x)=n$.

Proof. In this case $b^2>a^2$, hence $\Delta_n>0$ for $b\in[a,\infty)$. Since $x_n(a)=-1$, the continuity of $x_n(b)$ in the interval $[a,\infty)$ forces $x_n(b)<-1$ for all $b\in(a,\infty)$. Since $a^2-(n+\lambda)^2>0$, $x_n(b)$ will be in D^* if and only if $X_n(b)>0$. But this is the case, as

$$X_n(b)-b(a-(n+\lambda))>0,\qquad b\to\infty.$$

Now, $y_n(a)=-(a^2+(n+\lambda)^2)/(a^2-(n+\lambda)^2)<-1$, and the continuity of $y_n(b)$ in the interval $[a,\infty)$ implies $y_n(b)<-1$ for all $b>a$. If $y_n(b)$ were a solution of $B(x)=n$ we would have

$$\sqrt{y_n^2(b)-1}=-\frac{Y_n(b)}{a^2-(n+\lambda)^2}=\frac{Y_n(b)}{(n+\lambda)^2-a^2},$$

so that $Y_n(b)<0$. But obviously $Y_n(b)>0$ under the assumptions. Hence, $y_n(b)$ is not a solution of $B(x)=n$. This proves the theorem.

The results contained in Theorems 4.3 and 4.7 and their corollaries provide the information needed to characterize the set D. To determine D^* we need to consider the additional case $n=0$, $\lambda<0$. So, assume $\lambda<0=n$, and consider x_0, y_0 as function of the parameter b (for fixed a). Therefore

$$x_0(b)=\frac{-ab+\lambda\sqrt{\Delta_0}}{a^2-\lambda^2},\qquad y_0(b)=-\frac{ab+\lambda\sqrt{\Delta_0}}{a^2-\lambda^2}. \tag{4.13}$$

Observe that $\sqrt{\lambda^2}=-\lambda$, and now $x_0(b)=1$ if and only if $b=-a$, $y_0(b)=-1$ if and only if $b=a$. Again, $x_0(b)$ and $y_0(b)$ cannot take values in the interval $(-1,1)$.

Let

$$X_0(b)=a\sqrt{\Delta_0}-b\lambda,\qquad Y_0(b)=a\sqrt{\Delta_0}+b\lambda. \tag{4.14}$$

If x_0, y_0 are solutions of $B(x)=0$ then,

$$\sqrt{x_0^2-1}=\pm\frac{X_0(b)}{a^2-\lambda^2}, \qquad \sqrt{y_n^2-1}=\pm\frac{Y_0(b)}{a^2-\lambda^2}.$$

Therefore,

$$\sqrt{x_0^2-1}=\frac{X_0(b)}{a^2-\lambda^2},\quad x_0<-1, \qquad \sqrt{x_0^2-1}=-\frac{X_0(b)}{a^2-\lambda^2},\quad x_0>1, \tag{4.15}$$

while

$$\sqrt{y_0^2-1}=-\frac{Y_0(b)}{a^2-\lambda^2},\quad y_0<-1, \qquad \sqrt{y_0^2-1}=\frac{Y_0(b)}{a^2-\lambda^2},\quad y_0>1. \tag{4.16}$$

We now have that $X_0(b)=0$ if and only if $b=-a$, $Y_0(b)=0$ if and only if $b=a$. As before, it is easy to check that $X_0'(b)$, $Y_0'(b)$ never vanish, so that $X_0(b)$, $Y_0(b)$ are strictly monotonic in any interval I where $\Delta_0>0$; therefore, $X_0(b)$ never changes sign in such an interval if $-a\notin I$ and the same is true of $Y_0(b)$ if $a\notin I$. The signs of $X_0(b)$, $Y_0(b)$ can then be obtained, as before, from their asymptotic behavior.

Remark 4.8. Observe that the behavior of $X_0(b)$, $Y_0(b)$ when $\lambda>0$ is included in the case $n+\lambda>0$.

LEMMA 4.9. *If $a>\lambda$, $\lambda<0$, then $x_0>1$ for $b\neq -a$ and $y_0<-1$ when $b\neq a$.*

Proof. The proof is similar to that of Lemma 4.1. Observe that $\lambda^2-a^2=(\lambda+a)(\lambda-a)>0$ (since $a+\lambda<0$ if $\lambda<0$), so that $\Delta_0>0$ for all $b\in\mathbb{R}$. Considering again x_0 as a function of b we see that $x_0(-a)=1$, and therefore $x_0(b)>1$ if $b\neq -a$. The same type of argument proves the assertions concerning y_0.

THEOREM 4.10. *If $a>\lambda$, $\lambda<0$, then $x_0\in D^*$ if and only if $b\neq -a$ and $a\sqrt{\Delta_0}-b\lambda>0$. In this case $x_0>1$. Furthermore $y_0\in D^*$ if and only if $b\neq a$ and $a\sqrt{\Delta_0}+b\lambda>0$, in which case $y_0<-1$.*

Proof. The proof is essentially the same as that of Theorem 4.3.

Remark 4.11. Observe that $x_0(-a)$, $y_0(a)$ are not solutions of $B(x)=0$. However, both $x_0(a)$ and $y_0(-a)$ are in D^*.

COROLLARY 4.12. *Assume $a>b>0$, $\lambda<0$. Then x_0 and y_0 are solutions of $B(x)=0$ in D^*. Furthermore, $x_0>1$ and $y_0<-1$.*

Proof. We have $a>\lambda$ and obviously $a\sqrt{\Delta_0}-b\lambda>0$. Theorem 4.10 shows that $x_0\in D^*$ and $x_0>1$. If $b=0$, we trivially have $a\sqrt{\Delta_0}+b\lambda>0$. Therefore, we also have $a\sqrt{\Delta_0}+b\lambda>0$ if $b>0$. Thus, $y_0<-1$ and is in D^*.

THEOREM 4.13. *Assume $a<\lambda<0$. Then*

(i) *if $-b<a$, $x_0>1$ and is in D^* but y_0 does not make $B(x)=0$;*

(ii) *if $a\leqq -b$, $b>0$, neither x_0 nor y_0 is a solution of $B(x)=0$.*

Proof. (i) In this case $\Delta_0\geqq b^2-a^2>0$ and $b>-a$. Moreover $x_0(b)>1$ for $b>-a$, because $x_0=1$ when $b=-a$. Since $a\sqrt{\Delta_0}-b\lambda\sim b(a-\lambda)$ when $b\to+\infty$, we have $a\sqrt{\Delta_0}-b\lambda<0$. Thus $\sqrt{x_0^2-1}=[b\lambda-a\sqrt{\Delta_0}]/[a^2-\lambda^2]$ and $x_0(b)$ is in fact a solution of $B(x)=0$. On the other hand, $a\sqrt{\Delta_0}+b\lambda<0$ and $y_0(b)=-[ab+\lambda\sqrt{\Delta_0}]/[a^2-\lambda^2]$ contradict each other, as it is impossible to have $(a^2-\lambda^2)\sqrt{y_0^2-1}=a\sqrt{\Delta_0}+b\lambda$. Therefore, y_0 is not a solution of $B(x)=0$.

(ii) The conclusion is obviously true if $\Delta_0<0$. So, assume $\Delta_0>0$. We have $b<-a$ and from $a\sqrt{\Delta_0}-b'\lambda\sim b'(a-\lambda)<0$, as $b'\to+\infty$, we conclude that $a\sqrt{\Delta_0}-b\lambda>0$. Also, $x_0=1$ when $b=-a$, so that $x_0(b)>1$. This leads to the contradiction $\sqrt{x_0^2-1}=(b\lambda-a\sqrt{\Delta_0})/(a^2-\lambda^2)<0$. Hence, x_0 is not a solution of $B(x)=0$. Since $a\sqrt{\Delta_0}+b\lambda<0$ and $y_0(b)>1$ are contradictory, the proof is complete.

The proof of the following theorem requires only a slight modification of the proof of Theorem 4.13.

THEOREM 4.14. *Assume $\lambda < a < 0$. Then*

(i) *if $-b < a$, $x_0 > 1$ and is in D^* but y_0 is not a solution of $B(x) = 0$;*

(ii) *if $a < -b$, $b > 0$, neither x_0 nor y_0 is a solution of $B(x) = 0$.*

We shall now study the structure of D and the types of singularities of $\chi(z)$ on this set. We recall that D and D^* are subsets of $\mathbb{R} - [-1, 1]$ defined by

$$D = \{z \in \mathbb{C} - [-1, 1]: B(z) = 1, 2, \cdots\}, \qquad D^* = DU\{z: B(z) = 0\}. \tag{4.17}$$

THEOREM 4.15. *If $a \geqq |b|$, the set D is empty.*

Proof. When $x < -1$, $B(x)$ equals n iff $(n+\lambda)\sqrt{x^2-1} = ax + b$. When $n \geqq 1$, $\lambda > -\frac{1}{2}$ in order for the above inequality to hold $ax + b$ must be positive. But $ax + b < b - a < 0$ when $x < -1$ and we conclude that D has no elements less than -1. Similarly, when $x > 1$, $ax + b$ is positive and $B(x) = n$ iff $(n+\lambda)\sqrt{x^2-1} = -(ax+b)$, which cannot hold since $a \geqq |b|$.

We now consider the cases $a \leqq |b|$. Figure 4.1 shows a subdivision of the (λ, a) plane into five regions by the line $\lambda = -\frac{1}{2}$, the half line $a + \lambda + 1 = 0$, $\lambda > -\frac{1}{2}$ and the λ-axis $a = 0$. The first two lines do not belong to any region while the last one is assumed to be part of Region I for $\lambda > -\frac{1}{2}$. Notice that $\lambda = 0$ is not a division line. The positivity conditions $\lambda + a + 1 > 0$, $\lambda > -\frac{1}{2}$ restrict us to Regions I and II of the (λ, a) plane. We now characterize the set D when λ and a belong to either region. Consider first the case $b \geqq 0$.

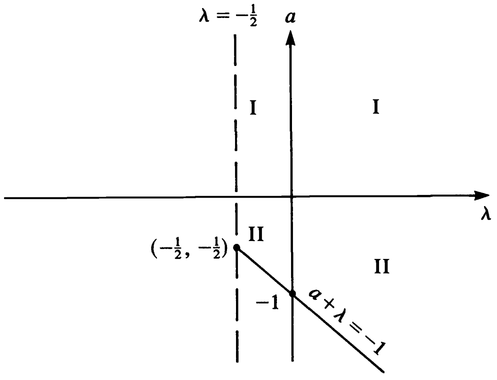


FIG. 4.1

THEOREM 4.16. *Let $a \leqq b$, $b \geqq 0$. Then, the set D is given, in each one of the two possible regions, as follows.*

Region I (i) *$a < b$. Then $D = \{x_n: n \geqq 1\}$.*
(ii) *$a = b$. Then $D = \varnothing$.*

Region II (i) *$-b \leqq a < b$. Then $D = \{x_n: n \geqq 1\}$.*
(ii) *$a < -b$. Then $D = \{x_n: n \geqq 1\} \cup \{y_n: n \geqq 1\}$.*

In the two regions x_n, y_n are given by (4.5) and we have $x_n < -1$ and $y_n > 1$ for all $n \geqq 1$.

Proof. I. Part (i) follows from Corollary 4.5 and Theorem 4.7. Part (ii) has been treated in Theorem 4.15.

II. Part (i) follows immediately from part (i) of Corollary 4.6. Part (ii) follows at once from Corollary 4.6 (ii).

Remark 4.17. The case $b=0$ was studied in Ismail [14]. It corresponds to the symmetric sieved Pollaczek polynomials of the second kind. Notice that in this case we are restricted to Region II with the constraint $a<-b$. By Theorem 4.16, D is $\{x\colon x=x_n \text{ or } y_n,\ n>0\}$.

We now come to the case $b<0$, $a\leqq -b$. Note that

$$B_n^\lambda(x;\, a, b;\, k)=(-1)^n B_n^\lambda(-x;\, a, -b;\, k) \tag{4.18}$$

follows from the three term recurrence relation. Thus we have the following.

THEOREM 4.18 *If $a\leqq -b$, $b<0$, the set D is given, in each one of the Regions* I, II *in the following manner.*

Region I (i) $a<-b$. *Then,* $D=\{y_n\colon n\geqq 1\}$.
(ii) $a=-b$. *Then,* $D=\varnothing$.
Region II (i) $b\leqq a<-b$. *Then,* $D=\{y_n\colon n\geqq 1\}$.
(ii) $a<b$. *Then,* $D=\{x_n\colon n\geqq 1\}\cup\{y_n\colon n\geqq 1\}$.

Here, x_n, y_n are also given by (4.7) *and $x_n<-1$, $y_n>1$ for all $n\geqq 1$.*

Now we study the nature of singularities of $\chi(z)$ on the set D. Recall that

$$\chi(z)=2\left\{\beta+(\alpha-\beta)(1-\beta^{2k})^{-A}\sum_{n=1}^{\infty}\frac{(A)_n}{n!}\left(\frac{\beta^k}{\beta^k-\alpha^k}\right)^n\frac{n}{n-B}\right\}.$$

Let $x_n\in D$ with $x_n<-1$ and $B(x_n)=n$. A simple calculation gives

$$\lim_{z\to x_n}(z-x_n)\chi(z)=2(-1)^{n+1}(\alpha-\beta)\beta^{2kn}(1-\beta^{2k})^{2\lambda}\frac{B(x_n)}{B'(x_n)}.$$

We used $B(x_n)=n$, $-A-B=2\lambda$, $\alpha\beta=1$ and $B'(x_n)=\lim_{z\to x_n}(B-n)/(z-x_n)$. Hence, if $B'(x_n)\neq 0$, x_n will be a simple pole of $\chi(z)$. We now have

$$B'(x_n)=\frac{d}{dx}\left[-\lambda+\frac{ax+b}{\sqrt{x^2-1}}\right]\Bigg|_{x=x_n}=\frac{a\sqrt{x_n^2-1}-(n+\lambda)x_n}{x_n^2-1}.$$

Using (4.5) one can derive

$$B'(x_n)=\frac{\sqrt{\Delta_n}[\alpha^2-(n+\lambda)^2]^2}{[a\sqrt{\Delta_n}-b(n+\lambda)]^2}. \tag{4.19}$$

In the same way when $y_n>1$ and $B(y_n)=n$, we obtain

$$B'(y_n)=-\frac{\sqrt{\Delta_n}[a^2-(n+\lambda)^2]^2}{[a\sqrt{\Delta_n}+b(n+\lambda)]^2}. \tag{4.20}$$

Therefore $B'(x)\neq 0$ in all cases.

Remark 4.19. Note that the denominator of (4.19) vanishes when $a=b$. This does not affect our results, because in this case $D=\varnothing$ in Region I, while $D\subseteq(1,+\infty)$ in Region II. The denominator of (4.20) is zero when $b=-a$, but in this case D is either empty or contained in $(-\infty,-1)$.

We now record the residues of $\chi(z)$ on D. Recall that $(-1)^n(A)_n=(-A-n+1)_n$, $-A-B=2\lambda$ and $\alpha-\beta=-2\sqrt{x_n^2-1}$ when $x_n<-1$, $\alpha-\beta=2\sqrt{y_n^2-1}$ when $y_n>1$. The use of (4.8), (4.19) and (4.20) lead to

$$\operatorname{Res}(\chi(z), x_n)=4\beta^{2kn}\frac{(2\lambda+1)_n}{(n-1)!\sqrt{\Delta_n}}(1-\beta^{2k})^{2\lambda}\left[\frac{a\sqrt{\Delta_n}-b(n+\lambda)}{a^2-(n+\lambda)^2}\right]^3,\qquad x_n<-1, \tag{4.21}$$

$$(4.22)\quad \operatorname{Res}(\chi(z), y_n) = 4\beta^{2kn}\frac{(2\lambda+1)_n}{(n-1)!\sqrt{\Delta_n}}(1-\beta^{2k})^{2\lambda}\left[\frac{a\sqrt{\Delta_n}+b(n+\lambda)}{a^2-(n+\lambda)^2}\right]^3, \qquad y_n > 1.$$

The last two equations could be written somewhat more explicitly by noticing that

$$(4.23)\qquad \beta(x_n) = \frac{\sqrt{\Delta_n}-b}{a-\lambda-n}, \qquad \alpha(x_n) = -\frac{\sqrt{\Delta_n}+b}{a+\lambda+n}, \quad x_n < -1,$$

$$(4.24)\qquad \beta(y_n) = -\frac{\sqrt{\Delta_n}+b}{a-n-\lambda}, \qquad \alpha(y_n) = \frac{\sqrt{\Delta_n}-b}{a+\lambda+n}, \qquad y_n > 1.$$

In view of Markov's Theorem we have the following.

THEOREM 4.20. *The function $\chi(z)$ has a simple pole at every point of D and $D \subseteq \sigma(\phi)$.*

Remark 4.21. Note that since

$$\lim_{n\to\infty} x_n = -1 \quad \text{and} \quad \lim_{n\to\infty} y_n = 1,$$

$-1, 1$ are the only possible limit points of D in $\mathbb{C}$ (or in $\mathbb{R}$).

In § 6 we shall characterize the singularities of $\chi(z)$ in $[-1, 1]$.

We next determine the structure of D^* as a function of a, b, λ. First we treat the case $b \geqq 0$.

THEOREM 4.22. *Let $a > b$, $b \geqq 0$. Then*

(i) *if $\lambda > 0$, $D^* = \emptyset$;*

(ii) *if $\lambda < 0$, $D^* = \{x_0, y_0\}$. Here $x_0 > 1$, $y_0 < -1$.*

Proof. The proof of (i) is similar to the proof of Theorem 4.15 and will be omitted. Part (ii) follows from Corollary 4.12.

Remark 4.23. The case $b = 0$ is also in Ismail [14]. If $a = b > 0$ we still have $x_0 \in D^*$, but $y_0 = -1 \in D^*$. The identity

$$(4.25)\qquad c_n^\lambda(x; a, b; k) = (-1)^n c_n^\lambda(-x; a, -b; k)$$

can be easily proved from the three term recurrence relation. Thus (4.25) and Theorem 4.22 give the following.

THEOREM 4.24. *If $a > -b$, $b < 0$, then $D^* = \emptyset$ when $\lambda > 0$ and $D^* = \{x_0, y_0\}$ when $\lambda < 0$. Here, $x_0 > 1$ and $y_0 < -1$.*

We now consider the case $a \leqq |b|$. In doing so it is convenient to divide the (λ, a)-plane into the ten regions shown in Fig. 4.2. The divisions are determined by the lines $\lambda = \frac{1}{2}$ and $\lambda = 0$; the half lines $\lambda + a = 0$, $\lambda > -\frac{1}{2}$ and $\lambda + a + 1 = 0$, $\lambda > -\frac{1}{2}$, as well as by the λ-axis, $a = 0$, for $\lambda < -\frac{1}{2}$ and $\lambda > 0$. The division between Regions III* and IV* is determined by the line segment $\lambda - a = 0$, $-\frac{1}{2} < \lambda < 0$. The positivity conditions

$$\lambda + a + 1 > 0, \quad \lambda(\lambda + a) > 0, \quad \lambda > -\tfrac{1}{2},$$

restrict cases of orthogonality to Regions I*–IV*. We assume that the λ-axis belong to Region I for $\lambda > 0$. None of the other division lines is part of any region. Note that $\lambda = 0$ is now a dividing line.

THEOREM 4.25. *The set D^* can be described as follows when $a \leqq b$, $b \geqq 0$.*

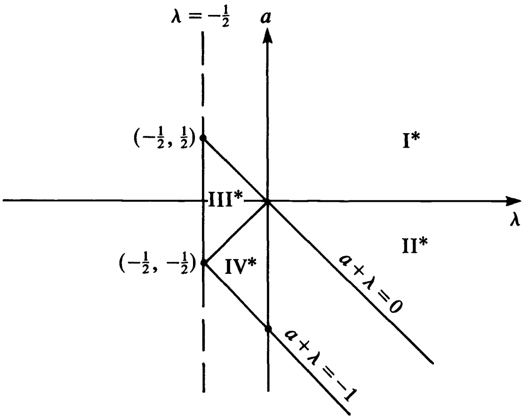


FIG. 4.2

Region I* (i) *$a<b$. Then, $D^*=\{x_n\colon n\geqq 0\}$.*
(ii) *$a=b$. Then, $D^*=\varnothing$.*
Region II* (i) *$-b\leqq a<b$. Then, $D^*=\{x_n\colon n\geqq 0\}$.*
(ii) *$a<-b$. Then, $D^*=\{x_n\colon n\geqq 0\}\cup\{y_n\colon n\geqq 0\}$.*
Region III* (i) *$-b<a$. Then, $D^*=\{x_n\colon n\geqq 0\}$, $x_0>1$.*
(ii) *$-b=a\neq 0$. Then, $D^*=\{x_n\colon n\geqq 1\}$.*
(iii) *$a<-b$. Then, $D^*=\{x_n\colon n\geqq 1\}\cup\{y_n\colon n\geqq 1\}$.*
(iv) *$a=b>0$ $(=0)$. Then, $D^*=\{x_0\}$ $(=\varnothing)$, $x_0>1$.*
Region IV* (i) *$-b<a$. Then, $D^*=\{x_n\colon n\geqq 0\}$, $x_0>1$.*
(ii) *$b=-a$. Then, $D^*=\{x_n\colon n\geqq 1\}$.*
(iii) *$a<-b$. Then, $D^*=\{x_n\colon n\geqq 1\}\cup\{y_n\colon n\geqq 1\}$.*

In all regions, $x_n<-1$ and $y_n>1$ for $n\geqq 1$. Also, $x_0<-1$, $y_0>1$ if $\lambda>0$.

Proof. Regions I*, II*. The proof in these cases is as the proof of I, II in Theorem 4.16, because Theorems 4.3, 4.7 and their corollaries hold for $n=0$ when $\lambda>0$.

III* (i) follows immediately from Theorem 4.16 for $n\geqq 1$, and from Theorem 4.14, (i), for $n=0$.
(ii) follows from Theorem 4.16 for $n\geqq 1$ and from Theorem 4.14, (ii), for $n=0$.
(iii) follows from Theorem 4.16 for $n\geqq 1$ and from Theorem 4.14, (ii), for $n=0$.
(iv) If $b=0$, it is impossible to have $-\lambda=n$, $n\geqq 0$, since $0>2\lambda>-1$. If $b>0$, use Theorem 4.16 and Remark 4.23.

IV* (i) follows from Theorem 4.16 for $n\geqq 1$ and from Theorem 4.13, (i), for $n=0$.
(ii) follows from Theorem 4.16 for $n\geqq 1$ and from Theorem 4.13, (ii), for $n=0$.
(iii) follows from Theorem 4.16 for $n\geqq 1$ and from Theorem 4.13, (ii), for $n=0$.

The proof of the theorem is now complete.

Now consider the case $b<0$. In view of (4.25) we easily obtain the following.

THEOREM 4.26. *If $a\leqq -b$, $b<0$, the composition of D^* in different regions of the a, b, λ parameter space is as follows.*

Region I* (i) *$a<-b$. Then, $D^*=\{y_n\colon n\geqq 0\}$.*
(ii) *$a=-b$. Then, $D^*=\varnothing$.*

Region II* (i) $b \leqq a < -b$. *Then*, $D^* = \{y_n \colon n \geqq 0\}$.
(ii) $a < b$. *Then*, $D^* = \{x_n \colon n \geqq 0\} \cup \{y_n \colon n \geqq 0\}$.
Region III* (i) $b < a < -b$. *Then*, $D^* = \{y_n \colon n \geqq 0\}$, $y_0 > 1$.
(ii) $a = b$. *Then*, $D^* = \{y_n \colon n \geqq 1\}$.
(iii) $a < b$. *Then*, $D^* = \{x_n \colon n \geqq 1\} \cup \{y_n \colon n \geqq 1\}$.
(iv) $a = -b > 0\ (=0)$. *Then*, $D^*\{y_0\}$ $(D^* = \varnothing)$.
Region IV* (i) $b < a$. *Then*, $D^* = \{y_n \colon n \geqq 0\}$, $y_0 < -1$.
(ii) $a = b$. *Then*, $D^* = \{y_n \colon n \geqq 1\}$.
(iii) $a < b$. *Then*, $D^* = \{x_n \colon n \geqq 1\} \cup \{y_n \colon n \geqq 1\}$.

In all regions we always have $x_n < -1$ *and* $y_n > 1$ *for* $n \geqq 1$. *Furthermore*, $x_0 < -1$, $y_0 > 1$, *if* $\lambda > 0$.

Remark 4.27. Note that $-1, 1$ are the only possible limit points of D^* in $\mathbb{C}$ (or in $\mathbb{R}$).

Now we study the nature of the singularities of $\chi^*(z)$ in D^*. If $x \in D^*$ and $B(x) = n > 1$, we have, from (4.2),

$$\lim_{z \to x} (z-x)\chi^*(z) = -2\frac{\lambda + a}{\alpha - \beta}\beta^{2kn}(1-\beta^2)^{2\lambda}\frac{(2\lambda)_n}{n!}\frac{n}{B'(x)}. \tag{4.26}$$

Here we used $B = n$, $-A - B = 2\lambda$ and $\alpha\beta = 1$. If $B(x) = 0$ then

$$\lim_{z \to x} (z-x)\chi^*(z) = -2\frac{(\lambda + a)}{\alpha - \beta}(1-\beta^{2k})^{2\lambda}\frac{1}{B'(x)}. \tag{4.27}$$

Now, a calculation similar to that used to obtain (4.19) and (4.20) shows that

$$B'(x_0) = \pm\frac{\sqrt{\Delta_0}(a^2 - \lambda^2)^2}{(a\sqrt{\Delta_0} - b\lambda)^2}. \tag{4.28}$$

The sign is according to whether $\lambda > 0$ or $\lambda < 0$ (i.e., according to whether $x_0 < -1$ or $x_0 > 1$). Furthermore

$$B'(y_0) = \mp\frac{\sqrt{\Delta_0}(a^2 - \lambda^2)^2}{(a\sqrt{\Delta_0} - b\lambda)^2} \tag{4.29}$$

with the same determination of the sign. Hence, all the singularities of $\chi^*(z)$ in D^* are simple poles.

Recall that $\alpha(x_n) - \beta(x_n) = -2\sqrt{x_n^2 - 1}$, $\alpha(y_n) - \beta(y_n) = 2\sqrt{y_n^2 - 1}$ for $n \geqq 1$, and also for $n = 0$ if $\lambda > 0$. If $\lambda < 0$ then $\alpha(x_0) - \beta(x_0) = 2\sqrt{x_0^2 - 1}$, $\alpha(y_0) - \beta(y_0) = -2\sqrt{y_0^2 - 1}$. This information coupled with (4.19), (4.20), (4.27), (4.28) and (4.29) establish

$$\operatorname{Res}(\chi(z), x_n) = (\lambda + a)\beta^{2kn}(1 - \beta^{2k})^{2\lambda}\frac{(2\lambda)_n}{n!\sqrt{\Delta_n}}\left[\frac{a\sqrt{\Delta_n} - b(n+\lambda)}{a^2 - (n+\lambda)^2}\right], \tag{4.30}$$

$$\operatorname{Res}(\chi(z), y_n) = (\lambda + a)\beta^{2kn}(1 - \beta^{2k})^{2\lambda}\frac{(2\lambda)_n}{n!\sqrt{\Delta_n}}\left[\frac{a\sqrt{\Delta_n} + b(n+\lambda)}{a^2 - (n+\lambda)^2}\right], \tag{4.31}$$

for $n > 1$, and also, according to $\lambda \lessgtr 0$,

$$\operatorname{Res}(\chi(z), x_0) = \pm\frac{(a + \lambda)}{\sqrt{\Delta_0}}(1 - \beta^{2k})\left[\frac{a\sqrt{\Delta_0} - b\lambda}{a^2 - \lambda^2}\right], \tag{4.32}$$

$$\operatorname{Res}(\chi(z), y_0) = \mp\frac{a + \lambda}{\sqrt{\Delta_0}}(1 - \beta^{2k})\left[\frac{a\sqrt{\Delta_0} + b\lambda}{a^2 - \lambda^2}\right]. \tag{4.33}$$

The quantities α and β are given by (4.23) and (4.24) when $n>1$ or $n=1$ and $\lambda>0$. Furthermore

$$\beta(x_0)=-\frac{\sqrt{\Delta_0}+b}{a-\lambda},\qquad \alpha(x_0)=\frac{\sqrt{\Delta_0}-b}{a+\lambda},\qquad \lambda<0, \tag{4.34}$$

$$\beta(y_0)=\frac{\sqrt{\Delta_0}-b}{a-\lambda},\qquad \alpha(y_0)=-\frac{\sqrt{\Delta_0}+b}{a+\lambda},\qquad \lambda<0. \tag{4.35}$$

In view of Markov's Theorem we have shown the following.

THEOREM 4.28. *The function* $\chi^*(z)$ *has a simple pole at each point of* D^* *and* $D^*\subseteq\sigma(x)$.

Note, again, that the only possible limit points of D^* in $\mathbb{R}$ are $-1, 1$.

5. Orthogonality relations. We saw in § 4 that

$$D\subseteq\sigma(\phi)\subseteq[-1,1]\cup D,\qquad D^*\subseteq\sigma(\psi)\subseteq[-1,1]\cup D^*.$$

The purpose of the present section is to compute the orthogonality relation for our polynomials. This will follow from analyzing the singularities of $\chi(z)$ and $\chi^*(z)$ in $[-1,1]$. Observe that both $\sigma(\phi)$ and $\sigma(\psi)$ are compact subset of $\mathbb{R}$.

We start by determining the measure $d\phi$.

Remark 5.1. Observe that when $\lambda>0$, (3.38) implies the more symmetric form

$$\begin{aligned}\chi(z)=-2\Big\{A\beta^{k-1}&\int_0^{\overline{\beta^k}|}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B}\,du\\ &+B\alpha^{k-1}\int_0^{\overline{\beta^k}|}(1-u\beta^k)^{-A}(1-u\alpha^k)^{-B-1}\,du\Big\}.\end{aligned} \tag{5.1}$$

In fact,

$$A\beta^k\int_0^{\overline{\beta^k}|}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B}\,du=-1-B\alpha^k\int_0^{\overline{\beta^k}|}(1-u\beta^k)^{-A}(1-u\alpha^k)^{-B-1}\,du$$

for z in a neighborhood of $[-1,1]$, where $\operatorname{Re}(B(z))>0$ and, hence for all z, by analytic continuation.

Now, the restriction of the function $\beta(x)/\alpha(x)=\beta_+(x)/\alpha_+(x)$, to $[-1,1]$ is continuous. Let

$$\xi_j=\cos(\pi j/k),\qquad j=1,2,\cdots,k-1. \tag{5.2}$$

Clearly the points $x\in[-1,1]$, $x\neq\xi_j$, for any j, make $|\beta(x)/\alpha(x)|=1$ and $\operatorname{Im}\{\beta(x)/\alpha(x)\}^k\neq 0$, thus $\operatorname{Re}\{\beta(x)/\alpha(x)\}^k<1$ and

$$1-\operatorname{Re}\left\{u\left(\frac{\beta(x)}{\alpha(x)}\right)^k\right\}>0,\qquad 0\leqq u\leqq 1.$$

Hence, there are $\gamma,\delta>0$ such that

$$1-\operatorname{Re}\left(u\left(\frac{\beta(z)}{\alpha(z)}\right)^k\right)\geqq\gamma,\qquad 0\leqq u\leqq 1+\delta,$$

for all z in a compact set $K_x=[x-\delta,x+\delta]\times[0,\delta]$. It follows that if $0<\varepsilon_n<\delta$ and $\varepsilon_n\to 0$ as $n\to\infty$, then

$$\lim_{n\to\infty}\left[1-u\left\{\frac{\beta(t+i\varepsilon_n)}{\alpha(t+i\varepsilon_n)}\right\}^k\right]^{-A(t+i\varepsilon_n)-1}=\left[1-u\left\{\frac{\beta(t)}{\alpha(t)}\right\}^k\right]^{-A(t)-1}$$

holds uniformly for all $t\in[x-\delta,x+\delta]$ and all $u\in[0,1+\delta]$. As in § 2, we let

$$F(z)=\int_0^{1\rceil}\left[1-u\left\{\frac{\beta(z)}{\alpha(z)}\right\}^k\right]^{-A(z)-1}(1-u)^{-B(z)}\,du,\qquad z\in\Omega.$$

From the above argument

$$F_+(t):=\lim_{n\to\infty}F(t+i\varepsilon_n)=\int_0^{1\rceil}\left[1-u\left\{\frac{\beta(t)}{\alpha(t)}\right\}^k\right]^{-A(t)-1}(1-u)^{-B(t)}\,du,\tag{5.3}$$

the limit being uniform for $t\in[x-\delta,x+\delta]$.

Recall that $\overline{\beta_-(x)}=\alpha(x)$, $\overline{\alpha_-(x)}=\beta(x)$, $\overline{B_-(x)}=A(x)$ and $\overline{A_-(x)}=B(x)$. It is now easy to get, uniformly on $[x-\delta,x+\delta]$

$$F_-(t):=\lim_{n\to\infty}F(t-i\varepsilon_n)=\int_0^{1\rceil}\left[1-u\left\{\frac{\alpha(t)}{\beta(t)}\right\}^k\right]^{-B(t)-1}(1-u)^{-A(t)}\,du.\tag{5.4}$$

This shows that both $F_+(x)$ and $F_-(x)$ are continuous functions of x, $-1\leqq x\leqq 1$, $x\neq\xi_j$. Now

$$\chi_\pm(t):=\lim_{\varepsilon\to 0+}\chi(t+i\varepsilon)=2\{\beta+A(\beta-\alpha)\beta^{2k}F_\pm(t)\}\tag{5.5}$$

uniformly in $[x-\delta,x+\delta]$. This shows that the limit

$$\tilde{\chi}(t)=\lim_{\varepsilon\to 0+}(\chi(t-i\varepsilon)-\chi(t+i\varepsilon))\tag{5.6}$$

is uniform in $[x-\delta,x+\delta]$. This and (5.5) prove that $\tilde{\chi}(x)$

$$\begin{aligned}\tilde{\chi}(x)=2(\alpha-\beta)\biggl\{1+A\beta^k&\int_0^{\beta^k\rceil}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B}\,du\\&+B\alpha^k\int_0^{\alpha^k\rceil}(1-u\beta^k)^{-A}(1-u\alpha^k)^{-B-1}\,du\biggr\}\end{aligned}\tag{5.7}$$

is a continuous function of x, for $-1<x<1$, $x\neq\xi_j$, $j=1,2,\cdots,k-1$. The uniform convergence in (5.6) and the Perron-Stieltjes Inversion Formula imply

$$\phi(x)-\phi(x')=\frac{1}{2\pi i}\int_{x'}^{x}\tilde{\chi}(t)\,dt,\qquad x'\in[x-\delta,x+\delta],$$

so that the measure $d\phi$ is absolutely continuous in each one of the subintervals (ξ_j,ξ_{j-1}), and is given on such an interval by

$$d\phi=\phi'(x)\,dx=\frac{1}{2\pi i}\tilde{\chi}(x)\,dx.\tag{5.8}$$

Now, the right-hand side of (5.7) is an analytic function of λ for $\operatorname{Re}(\lambda)>-\frac{1}{2}$. If $\lambda\geqq 0$, we can use integration by parts because the first integral is a proper integral. Thus

$$\int_0^{\beta^k\rceil}(1-u\beta^k)^{-A-1}(1-u\alpha^k)^{-B}\,du=-\frac{1}{A\beta^k}-\frac{B\alpha^k}{A\beta^k}\int_0^{\beta^k\rceil}(1-u\beta^k)^{-A}(1-u\alpha^k)^{-B-1}\,du.$$

Therefore

$$\tilde{\chi}(x)=-2(\alpha-\beta)\alpha^kB\int_{\lfloor\alpha^k}^{\beta^k\rceil}(1-u\beta^k)^{-A}(1-u\alpha^k)^{-B-1}\,du,\qquad\lambda\geqq 0,$$

which is equivalent to

$$\tilde{\chi}(x)=-2(\alpha-\beta)B\left(1-\frac{\beta^k}{\alpha^k}\right)^{-A}\left(1-\frac{\alpha^k}{\beta^k}\right)^{-B}\int_{\lfloor 0}^{1\rceil}(1-v)^{-B-1}v^{-A}\,dv.\tag{5.9}$$

The integral on the right-hand side is, in fact, proper. Therefore

$$\tilde{\chi}(x)=2(\alpha-\beta)\left(1-\frac{\beta^k}{\alpha^k}\right)^{-A}\left(1-\frac{\alpha^k}{\beta^k}\right)^{-B}\frac{\Gamma(-B+1)\Gamma(-A+1)}{\Gamma(2\lambda+1)},\tag{5.10}$$

since $A+B=-2\lambda$. Since the right-hand side is also an analytic function of λ for $\operatorname{Re}(\lambda)>-\frac{1}{2}$, the identity (5.10) holds for all $\lambda>-\frac{1}{2}$. From (5.8) and (5.10) we obtain

$$\phi'(x)=\frac{\alpha-\beta}{\pi i}\left(1-\frac{\beta^k}{\alpha^k}\right)^{-A}\left(1-\frac{\alpha^k}{\beta^k}\right)^{-B}\frac{\Gamma(-B+1)\Gamma(-A+1)}{\Gamma(2\lambda+1)}\tag{5.11}$$

holding for $x\neq\xi_j$. Clearly (5.11) determines $d\phi$ as an absolutely continuous measure in each one of the intervals (ξ_j,ξ_{j-1}), $j=1,2,\cdots,k$.

Now write $x=\cos\theta$, $(j-1)\pi/k<\theta<j\pi/k$, $j=1,2,\cdots,k$, and recall that $\alpha(x)$, $\beta(x)$ are complex conjugate and so are $A(x)$, $B(x)$. Since $\alpha(\cos\theta)=e^{i\theta}$, $\beta(\cos\theta)=e^{-i\theta}$, $A=-\lambda-i\Phi(x)$ and $B=-\lambda+i\Phi(x)$, where $\Phi(\cos\theta)=a\cot\theta+b\csc\theta$. We thus obtain

$$\left(1-\frac{\beta^k}{\alpha^k}\right)^{-A}\left(1-\frac{\alpha^k}{\beta^k}\right)^{-B}=2^{2\lambda}|\sin k\theta|^{2\lambda}\exp\{\Phi(\cos\theta)(2k-2j\pi+\pi]\},$$

where we used the identities

$$\arg(i\alpha^{-k}\sin k\theta)=\pi j-k\theta-\frac{\pi}{2},\qquad \arg(-i\alpha^k\sin k\theta)=k\theta-j\pi+\frac{\pi}{2}.\tag{5.12}$$

For $(j-1)\pi<k\theta<j\pi$, $j=1,2,\cdots,k$, (5.11) is

$$\begin{aligned}\phi'(\cos\theta)=2\frac{2\lambda+1\sin\theta}{\pi\Gamma(2\lambda+1)}&|\sin(k\theta)|^{2\lambda}|\Gamma(\lambda+1+i\Phi(\cos\theta))|^2\\ &\cdot\exp\{2\Phi(\cos\theta)[k\theta-j\pi+\pi/2]\}.\end{aligned}\tag{5.13}$$

Note that if $\lambda\geqq 0$, ϕ' is continuous in $(-1,1)$, and $d\phi$ is absolutely continuous in $(-1,1)$. Clearly ϕ' vanishes at $t=\xi$ if $\lambda>0$.

We now apply Lemma 1.3 to show that the points $-1, 1, \xi_j$, $0<j<k$ do not support discrete masses. In the present case the coefficients A_n and C_n of (1.2) and (1.3) are

$$A=\begin{cases}2 & \text{if } k\nmid n+1,\\ \dfrac{2(\lambda+a+m)}{m} & \text{if } n+1=km,\end{cases}\qquad C_n=\begin{cases}1 & \text{if } k\nmid n+1,\\ \dfrac{2\lambda+m}{m} & \text{if } n+1=mk.\end{cases}\tag{5.14}$$

Therefore, from (1.1) and (1.6),

$$\lambda_n=\begin{cases}\dfrac{m}{\lambda+a+m}\cdot\dfrac{(2\lambda+1)_m}{m!} & \text{if } n+1=km,\\ \dfrac{(2\lambda+1)_p}{p!}, & p=[(n+1)/k]\quad\text{if } k\nmid n+1.\end{cases}\tag{5.15}$$

This and (1.40) show that $\Gamma(2\lambda+1)\lambda_{mk}\sim m^{2\lambda}$ as $m\to\infty$. Now we determine the asymptotic behavior of the polynomials $\{B_n^\lambda(x)\}$ at $x=\pm1$, ξ_j, $0<j<k$. Clearly $A(\xi_j)$, $B(\xi_j)$ are well defined. At $\alpha=\alpha(\xi_j)$ and $\beta=\beta(\xi_j)$, $B^\lambda(x,t)$ has an algebraic branch singularity of order $-A-B+1=2\lambda+1$. The dominant term in a comparison function is

$$\tilde{B}^\lambda(\xi_j,t)=k^{-2\lambda}\frac{\alpha}{\alpha-\beta}(1-t/\beta)^{-2\lambda-1}+k^{-2\lambda}\frac{\beta}{\beta-\alpha}(1-t/\alpha)^{-2\lambda-1}.\tag{5.16}$$

Darboux's Method readily gives

$$B_n^\lambda(\xi_j) \sim k^{-2\lambda} \frac{(\alpha^{n+1}-\beta^{n+1})(2\lambda+1)_n}{(\alpha-\beta)n!}.$$

Using (1.40) and $U_n(\varepsilon_j) = (\alpha^{n+1}-\beta^{n+1})/(\alpha-\beta) = \pm 1$, we obtain $[\Gamma(2\lambda+1)B_{mk}^\lambda(\xi_j)]^2 \sim m^{4\lambda}$, which implies

$$\frac{[B_{mk}^\lambda(\xi_j)]^2}{\lambda_{mk}} \sim \frac{m^{2\lambda}}{\Gamma(2\lambda+1)}. \tag{5.17}$$

Therefore the series $\sum_{n=0}^\infty [B_n^\lambda(\xi_j)]^2/\lambda_n$ is divergent, and $d\phi$ has no masses at the points $\xi_j, j=1, 2, \cdots, k-1$.

Now we consider the points $x=\pm 1$. In this case A and B are undefined, and we have to evaluate $B^\lambda(\pm 1, t)$ by a limiting process. In view of (2.25) and (3.31) the generating function (3.19) can be expressed in the form

$$B^\lambda(z, t) = \frac{(1-t^k\alpha^k)^{-\lambda}(1-t^k\beta^k)^{-\lambda}}{(1-t\alpha)(1-t\beta)} \exp\left\{ i\Phi(z) \log \frac{1-t^k\beta^k}{1-t^k\alpha^k} \right\}.$$

As $z \to \pm 1$, $a(z)$, $\beta(z) \to \pm 1$, and we apply L'Hôpital's rule to determine the limit under the exponential. Now (2.19) yields

$$\alpha'(z) = 2\alpha(z)/[\alpha(z)-\beta(z)], \qquad \beta'(z) = -2\beta(z)/[\alpha(z)-\beta(z)],$$

and an elementary calculation establishes the representation

$$B^\lambda(1, t) = \left(\frac{1-t^k}{1-t}\right)^2 (1-t^k)^{-2\lambda-2} \exp\left\{ 2(a+b)k \frac{t^k}{1-t^k} \right\}, \tag{5.18}$$

or, equivalently,

$$B^\lambda(1, t) = \left(\frac{1-t^k}{1-t}\right)^2 L^{(2\lambda+1)}(-2k(a+b), t^k), \tag{5.19}$$

where

$$L^{(2\lambda+1)}(x, t) := \sum_{n=0}^\infty L_n^{(2\lambda+1)}(x) t^n$$

is a generating function for the Laguerre polynomials of order $2\lambda+1$. Similarly

$$B^\lambda(-1, t) = \left(\frac{1 \pm t^k}{1+t}\right)^2 L^{(2\lambda+1)}(-2k(a-b), \pm t^k), \quad k = \text{odd or even}. \tag{5.20}$$

This and Darboux's Method establish

$$(\pm 1)^{km} B_{km}^\lambda(\pm 1) \sim k^2 L_m^{(2\lambda+1)}(-2k(a \pm b)). \tag{5.21}$$

Now, if $a=-b$, then $B_{km}^\lambda(1) \sim k^2 L_m^{(2\lambda+1)}(0) = k^2\Gamma(m+2\lambda+2)/[m!\Gamma(2\lambda+2)]$ (Lebedev [17, p. 85]) and $[B_{km}^\lambda(1)]^2/\lambda_{km} \sim m^{2\lambda+2}/(2\lambda+1)$. This argument shows that the series $\sum_{m=0}^\infty [B_{km}^\lambda(1)]^2/\lambda_{km}$ diverges and we conclude that $d\phi$ has no mass at 1 if $a+b=0$. If $a+b>0$, then Fejér's Formula for the Laguerre polynomials (Szegö [29, p. 198]) gives

$$B_{km}^\lambda(1) \sim \sqrt{\frac{e^{-2k(a+b)}}{\pi}}\, m^{\lambda+1/4}[-2k(a+b)]^{-\lambda-3/4} \cos\left(2\sqrt{-2km(a+b)} - \left(\lambda+\frac{3}{4}\right)\pi\right),$$

so that

$$\frac{[B_{km}^{\lambda}(1)]^2}{\lambda_{km}} \sim \frac{\sqrt{m}}{\pi} e^{-2k(a+b)}(-2k(a+b))^{-2\lambda-3/2}\Gamma(2\lambda+1)$$
$$\cdot \cos^2\left(2\sqrt{-2km(a+b)}-\left(\lambda+\frac{3}{4}\right)\pi\right).$$

Since the mth terms of the series $\sum_{m=0}^{\infty}[B_{km}^{\lambda}(1)]^2/\lambda_{km}$ does not tend to zero as $m\to\infty$, the series is divergent and 1 does not support any mass of $d\phi$. In case $a+b<0$, there is a positive constant C such that

$$B_{km}^{\lambda}(1)\sim C\sqrt{\frac{e^{-2k(a+b)}}{\pi}}(2k(a+b))^{-\lambda-3/4}m^{\lambda+1/2}\, e^{2\sqrt{2km(a+b)}}.$$

This is Perron's Formula for the Laguerre polynomials (Szegö [30, p. 199]). As before

$$\frac{[B_{km}^{\lambda}(1)]^2}{\lambda_{km}} \sim C\frac{e^{-2k(a+b)}(2k(a+b))^{-2\lambda-3/2}}{\pi}\Gamma(2\lambda+1)\sqrt{m}\, e^{4\sqrt{2km(a+b)}},$$

and the series $\sum_{n=0}^{\infty}[B_n^{\lambda}(1)]^2/\lambda_n$ also diverges.

A similar argument based on (5.21) shows that -1 supports no masses of $d\phi$. Thus we have proved the following.

THEOREM 5.2. *The measure $d\phi$ is absolutely continuous in the interval* $[-1,1]$ *and has masses only at the points of the discrete set D where B is a positive integer. In the interval* $[-1,1]$, *$d\phi=\phi'\,dx$, where ϕ' is given by*

$$\phi'(\cos\theta)=2^{2\lambda+1}\frac{(\sin\theta)^{2\lambda+1}}{\pi\Gamma(2\lambda+1)}|U_{k-1}(\cos\theta)|^{2\lambda}|\Gamma(\lambda+1+i\Phi(\cos\theta))|^2$$
$$\cdot\exp\left[2(k\theta-j\pi+\pi/2)\Phi(\cos\theta)\right] \tag{5.22}$$

and $\Phi(\cos\theta)=a\cot\theta+b\csc\theta$, $(j-1)\pi/k<\theta<j\pi/k$. *Furthermore, the support of $d\phi$ is the set*

$$\sigma(\phi)=DU[-1,1], \tag{5.23}$$

and $-1, 1$ are the only possible limit points of D in R. With the weight function

$$w(x)=\frac{\pi\Gamma(2\lambda+1)}{2^{2\lambda+1}}\phi'(x), \tag{5.24}$$

we have the orthogonality relation

$$\int_{-\infty}^{+\infty}B_n^{\lambda}(x)B_m^{\lambda}(x)w(x)\,dx+\sum_{\xi\in D}J_{\xi}B_n^{\lambda}(\xi)\beta_m^{\lambda}(\xi)=\frac{\pi\Gamma(2\lambda+1)}{2^{2\lambda+1}}\lambda_n\delta_{mn} \tag{5.25}$$

where

$$J_{\xi}=\frac{\pi\Gamma(2\lambda+1)}{2^{2\lambda+1}}\operatorname{Res}(\chi(z),\xi) \tag{5.26}$$

and Res $(\chi(z),\xi)$ *is as in* (4.21) *or* (4.22).

Finally, the set D is empty if and only if $a\geqq|b|$, and is an infinite countable set if $a<|b|$. Every point of $[-1,1]$ *is a continuous singularity of $\chi(z)$.*

We now wish to determine ψ' and show that the measure $d\psi$ has no discrete masses in $[-1, 1]$. The argument is similar to what we used to analyze $\chi(z)$. When $\lambda = 0$, (3.55) and integration by parts give

$$\chi^*(z) = 2(\lambda + a)\frac{\alpha^k - \beta^k}{\alpha - \beta}\int_0^{\overline{\beta^k}\rceil}(1 - u\beta^k)^{-A-1}(1 - u\alpha^k)^{-B-1}\,du. \tag{5.27}$$

Let $G(z) := \int_0^{\overline{1}\rceil}(1 - u\beta^{2k})^{-A-2}(1-u)^{-B}\,du$. Therefore

$$G_+(x) := \lim_{y\to 0+} G(x+iy) = \int_0^{\overline{1}\rceil}(1 - u\beta^{2k})^{-A-2}(1-u)^{-B}\,du,$$

$$G_-(x) := \lim_{y\to 0+} G(x-iy) = \int_0^{\overline{1}\rceil}(1 - u\beta^{2k})^{-B-2}(1-u)^{-A}\,du,$$

the limits being uniform in any closed subinterval of (ξ_j, ξ_{j-1}), $0 < j < k$. This implies, as before, that the limits

$$\chi^*_+(x) = \lim_{\varepsilon\to 0+}\chi^*(x + i\varepsilon), \qquad \chi^*_-(x) = \lim_{\varepsilon\to 0+}\chi^*(x - i\varepsilon)$$

hold uniformly on compact subintervals of (ξ_j, ξ_{j-1}). The function

$$\tilde{\chi}^*(x) = \chi^*_-(x) - \chi^*_+(x) = \lim_{\varepsilon\to 0+}(\chi^*(x - i\varepsilon) - \chi^*(x + i\varepsilon))$$

satisfies

$$\tilde{\chi}^*(x) = 2(\lambda + a)\frac{\alpha^k - \beta^k}{\alpha - \beta}\int_{\underline{\lfloor\alpha^k}}^{\overline{\beta^k}\rceil}(1 - u\beta^k)^{-A-1}(1 - u\alpha^k)^{-B-1}\,du, \tag{5.28}$$

from which we conclude that $\tilde{\chi}^*(x)$ is continuous in each one of the intervals (ξ_j, ξ_{j-1}). Furthermore, in (ξ_j, ξ_{j-1}), $d\psi$ is absolutely continuous with $2\pi i\,dx = \tilde{\chi}^*(x)\,dx$. This shows that

$$\pi i\psi'(x) = (\lambda + a)\frac{\alpha^k - \beta^k}{\alpha - \beta}\int_{\underline{\lfloor\alpha^k}}^{\overline{\beta^k}\rceil}(1 - u\beta^k)^{-A-1}(1 - u\alpha^k)^{-B-1}\,du,$$

that is

$$\pi i\psi'(x) = \frac{\lambda + a}{\alpha - \beta}\left(1 - \frac{\beta^k}{\alpha^k}\right)^{-A}\left(1 - \frac{\alpha^k}{\beta^k}\right)^{-B}\int_{\underline{\lfloor 0}}^{\overline{1}\rceil}(1 - \nu)^{-B-1}\nu^{-A-1}\,d\nu.$$

This establishes the explicit formula

$$\pi i\psi'(x) = \frac{\lambda + a}{\alpha - \beta}(1 - \beta^{2k})^{-A}(1 - \alpha^{2k})^{-B}\frac{\Gamma(-A)\Gamma(-B)}{\Gamma(2\lambda)}, \tag{5.29}$$

holding for $x \neq \xi_j$. Set $x = \cos\theta$, $(j-1)\pi < k\theta < j\pi$. The observations $\alpha(\cos\theta) = e^{i\theta}$, $\beta(\cos\theta) = e^{-i\theta}$ and $A = -\lambda - i\Phi(x)$, $B = -\lambda + i\Phi(x)$, where $\Phi(x)$ is as in (3.32), lead to

$$\begin{aligned}\psi'(\cos\theta) &= 2^{2\lambda-1}(\lambda + a)\frac{(\sin\theta)^{2\lambda-1}}{\pi\Gamma(2\lambda)}|U_{k-1}(\cos\theta)|^{2\lambda}\\ &\quad\cdot|\Gamma(\lambda + i\Phi(\cos\theta))|^2\exp(2\Phi(\cos\theta)(k\theta - j\pi + \pi/2)).\end{aligned} \tag{5.30}$$

Again, $\psi'(x)$ is continuous in $(-1, 1)$ if $\lambda > 0$, and vanishes at $x = \xi_j$, $j = 1, 2, \cdots, k-1$.

We now examine the possibility of $d\psi$ having discrete masses at the points $x = \pm 1$, ξ_j. We recall from formulas (3.19) and (3.45) that

$$C^\lambda(x, t) = \left(1 + \frac{2b}{\lambda + a}t - \frac{\lambda - a}{\lambda + a}t^2\right)B^\lambda(x, t). \tag{5.31}$$

If $x=\xi_j$ we have, from (2.19) and (2.20)

$$\lim_{t\to\beta}\left(1+\frac{2b}{\lambda+a}t-\frac{\lambda-a}{\lambda+a}t^2\right)=\frac{(\beta-\alpha)B}{\alpha(\lambda+a)},\qquad \lim_{t\to\alpha}\left(1+\frac{2b}{\lambda+a}t-\frac{\lambda-a}{\lambda+a}t^2\right)=\frac{(\alpha-\beta)A}{(\lambda+a)\beta}.$$

Hence, from (5.16)

$$\begin{aligned}\frac{1}{2}b_nc_n^\lambda(x)&\sim-\frac{(A+B)(n/k)^{2\lambda}}{\Gamma(2\lambda+1)(\lambda+a)}\cos(jn\pi/k)\\&=\frac{(n/k)^{2\lambda}}{\Gamma(2\lambda)(\lambda+a)}\cos(jn\pi/k).\end{aligned}$$

From (3.47), we have

$$b_{mk}=\frac{(\lambda+a+1)_m(2\lambda)_m}{m!(\lambda+a)_m}\sim\frac{m^{2\lambda}}{(\lambda+a)\Gamma(2\lambda)}. \tag{5.32}$$

This shows that

$$c_{mk}^\lambda(\xi_j)\sim 2(-1)^m,\qquad j=1,2,\cdots,k-1, \tag{5.33}$$

as can also be checked from the recurrence relation.

On the other hand $A_0=1$,

$$A_n=\begin{cases}2 & \text{if } k\nmid n,\\ \dfrac{2(\lambda+a+m)}{2\lambda+m} & \text{if } n=mk,\end{cases}\qquad C_n=\begin{cases}1 & \text{if } k\nmid n,\\ \dfrac{m}{2\lambda+m} & \text{if } n=mk,\end{cases}$$

so that

$$\lambda_{mk+l}=\frac{\lambda(m!)}{(2\lambda)_{m+1}},\quad 0<l<k;\qquad \lambda_{mk}=\frac{\lambda(m!)}{(\lambda+a+m)(2\lambda)_{m+1}}. \tag{5.34}$$

Clearly (5.34) gives

$$\lambda_{mk}\sim\lambda\Gamma(2\lambda)m^{-2\lambda}, \tag{5.35}$$

which when combined with (5.33) and the fact $2\lambda>-1$ will prove the divergence of the series $\sum_{n=0}^\infty[c_n^\lambda(\xi_j)]^2/\lambda_n$. Hence ξ_j, $0<j<k$ is not a discrete mass point.

Now we consider the cases $x=\pm1$. Note that the function

$$h(t):=1-\frac{\lambda-a}{\lambda+a}t^2+\frac{2b}{\lambda+a}t=1-t^2+2t\frac{at+b}{\lambda+a}$$

vanishes at $t=1$ when $a+b=0$ and at $t=-1$ if $a-b=0$. The functions $C^\lambda(\pm1,t)$ behave differently according to whether $a=\pm b$ or $|a|\neq|b|$. The reason is that $\lim_{t\to\pm1}h(t)=2(a\pm b)/(\lambda+a)$. Using the asymptotic properties of Laguerre polynomials, one can show that $x=\pm1$ are not discrete mass points of $d\psi$ if $|a|\neq|b|$. The proof is similar to the way we handled the same points in relation to the measure $d\phi$. The same conclusion holds for $x=1$ if $a=b$ and for $x=-1$ if $a=-b$. Now, assume $a=-b$. Clearly

$$C^\lambda(1,t)=\frac{h(t)}{1-t}\frac{1-t^k}{1-t}L^{(2\lambda)}(0,t^k) \tag{5.36}$$

implies $(l+a)b_{mk}c_{mk}^\lambda(1)\sim 2k\lambda L_m^{(2\lambda)}(0)$. This, (5.32) and (5.33) ensure the divergence of the series $\sum_{n=0}^\infty[c_n^\lambda(1)]/\lambda_n$. The case $a=b$ and $x=-1$ can be handled in a similar way. This shows that neither $x=1$ nor $x=-1$ support discrete masses for dx. We summarize these results in the next theorem.

THEOREM 5.3. *The measure $d\psi$ is absolutely continuous in $[-1, 1]$ and its only masses are located on the discrete set D^* where $B(z)=0, 1, 2, \cdots$. On $[-1, 1]$, $d\psi=\psi'\,dx$, where $\psi'(x)$ is given by* (5.29) *or* (5.30). *The support of the measure $d\psi$ is*

$$\sigma(\psi)=[-1, 1]UD^*, \tag{5.37}$$

and the set D^* is empty if and only if $a \geqq |b|$ and $\lambda>0$, and a countable subset of $\mathbb{R}-[-1, 1]$, having at most 1 or -1 as limit points, in any case. With the weight function

$$W^*(x)=\pi\frac{\Gamma(2\lambda)}{\lambda+a}2^{-2\lambda+1}\psi'(x) \tag{5.38}$$

we obtain the orthogonality relation

$$\int_{-\infty}^{+\infty} c_n^\lambda(x)c_m^\lambda(x)W^*(x)\,dx+\sum_{\xi\in D^*} J_\xi c_n^\lambda(\xi)c_m^\lambda(\xi)=\lambda_n\delta_{mn}, \tag{5.39}$$

where λ_n is given by (5.34), *and*

$$J_\xi=\pi\frac{\Gamma(2\lambda)}{\lambda+a}2^{-2\lambda+1}\,\mathrm{Re}\, s(\chi^*(z), \xi), \tag{5.40}$$

and Res $(\chi(z), \xi)$ *is as in* (4.30), (4.31), (4.32) *and* (4.33).

6. The Pollaczek polynomials and their q-analogues. The recurrence relation (1.14) and the initial conditions (1.15) define the general Pollaczek polynomials $\{P_n^\lambda(x, a, b)\}$, or $\{P_n^\lambda(x)\}$ for short. The case $\lambda=\frac{1}{2}$ was studied in Pollaczek [24]. Szegö [29] studied the general nonsingular case $a \geqq |b|$, $\lambda>0$, and showed that the polynomials are orthogonal with respect to an absolutely continuous measure. He also found the measure. Special instances of the singular case have been studied by Bank and Ismail [7]. They covered the cases $a=\pm b$. In this section, we first study the general case, and especially the singular cases among $a<|b|$. It turns out that this is essentially the case $k=1$ of the sieved Pollaczek polynomials of the first kind previously studied. We also compute the continued fraction whose denominators are the q-Pollaczek polynomials $\{F_n(x)\}$, then determine the corresponding distribution function when it is continuous. The results of § 2 may be used to completely analyze the singular cases, if needed, but we will spare the reader the details. The q-Pollaczek polynomials arose recently in Al-Salam and Chihara's solution of a characterization problem of Andrews and Askey.

We denote by $P_n^{*\lambda}(x; a, b)$, or simply by $P_n^{*\lambda}(x)$, the Pollaczek numerator polynomials, and by $\chi_0(z)$ the continued fraction

$$\chi_0(z)=\lim_{n\to\infty} P_n^{*\lambda}(x)/P_n^\lambda(x). \tag{6.1}$$

The numerator polynomials satisfy (1.14) and the initial conditions

$$P_0^{*\lambda}(x)=0, \qquad P_1^{*\lambda}(x)=2(\lambda+a). \tag{6.2}$$

The positivity conditions are

$$(a+\lambda+n-1)(a+\lambda+n)(2\lambda+n-1)>0, \qquad n=1, 2, \cdots, \tag{6.3}$$

which evidently reduce to

$$\lambda>0 \text{ and } \lambda+a>0 \text{ or } -\tfrac{1}{2}<\lambda<0 \text{ and } 0<\lambda+a+1<1. \tag{6.4}$$

We will use the generating functions

$$P(x,t)=\sum_{n=0}^{\infty}P_n^{\lambda}(x)t^n,\qquad P^*(x,t)=\sum_{n=0}^{\infty}P_n^{*\lambda}(x)t^n. \tag{6.5}$$

They are solutions of the initial value problems

$$(t^2-2xt+1)\frac{\partial P}{\partial t}-2[(a+\lambda)x-\lambda t+b]P=0,\qquad P(x,0)=1, \tag{6.6}$$

$$(t^2-2xt+1)\frac{\partial P^*}{\partial t}-2[(a+\lambda)x-\lambda t+b]P^*=2(\lambda+a),\qquad P^*(x,0)=0. \tag{6.7}$$

From the partial fraction decomposition

$$\frac{2[(a+\lambda)x-\lambda t+b]}{t^2-2xt+1}=\frac{A(x)}{t-\alpha(x)}+\frac{B(x)}{t-\beta(x)} \tag{6.8}$$

where $\alpha(x)$, $\beta(x)$ are given by (2.19) and $A(x)$, $B(x)$ by (2.20), we readily obtain

$$P(x,t)=(1-t/\alpha)^A(1-t/\beta)^B \tag{6.9}$$

and

$$\begin{aligned} P^*(x,t)&=2(\lambda+a)(1-t/\alpha)^A(1-t/\beta)^B\int_0^t(1-u/\alpha)^{-A-1}(1-u/\beta)^{-B-1}\,du\\ &=2\frac{\lambda+a}{B(\alpha-t)}-\frac{2(\lambda+a)\beta}{B}\left[1+\frac{A+1}{\alpha}\int_0^t(1-u/\alpha)^{-A-2}(1-u/\beta)^{-B}\,du\right]\\ &\quad\cdot(1-t/\alpha)^A(1-t/\beta)^B. \end{aligned} \tag{6.10}$$

For fixed $x\in\mathbb{C}-[-1,1]$ and as $n\to\infty$, Darboux's Method gives

$$P_n^{\lambda}(x)\sim(1-\beta^2)^A\beta^{-n}n^{-B-1}/\Gamma(-B) \tag{6.11}$$

and

$$\begin{aligned} P_n^{*\lambda}(x)&\sim2\frac{\lambda+a}{\Gamma(-B+1)}\beta^{-n+1}n^{-B+1}(1-\beta/\alpha)^A\\ &\quad\cdot\left[1+\frac{A+1}{\alpha}\int_0^{\overline{\beta|}}(1-u/a)^{-A-2}(1-u/\beta)^{-B}\,du\right]. \end{aligned} \tag{6.12}$$

When $\lambda>0$, (6.12) takes the more symmetric form

$$P_n^{*\lambda}(x)\sim2(\lambda+a)(1-\beta/\alpha)^A\beta^{-n}\frac{n^{-B-1}}{\Gamma(-B)}\int_0^{\overline{\beta|}}(1-u/\alpha)^{-A-1}(1-u/\beta)^{-B-1}\,du. \tag{6.13}$$

Therefore

$$\chi_0(x)=-2\frac{(\lambda+a)\beta}{B}\left[1+\frac{A+1}{\alpha}\int_0^{\overline{\beta|}}(1-u\beta)^{-A-2}(1-u\alpha)^{-B}\,du\right] \tag{6.14}$$

or equivalently

$$\chi_0(x)=-2\frac{(\lambda+a)\beta}{B}\left[1+(A+1)\frac{\beta}{\alpha}\int_0^{\overline{1|}}(1-u\beta^2)^{-A-2}(1-u)^{-B}\,du\right], \tag{6.15}$$

and, when $\lambda>0$,

$$\chi_0(x)=2(\lambda+a)\beta\int_0^{1}(1-u\beta^2)^{-A-1}(1-u)^{-B}\,du, \tag{6.16}$$

$$\chi_0(x)=2(\lambda+a)\int_0^{\beta}(1-u\beta)^{-A-1}(1-u\alpha)^{-B-1}\,du. \tag{6.17}$$

As in the case of the continued fraction $\chi^*(x)$ we can show that $\chi_0(z)$ is analytic on Ω^* of (2.36). The series representation

$$\chi_0(x)=-2\frac{\lambda+a}{B}\beta\left\{1-(1-\beta^2)^{-A-1}\sum_{n=1}^{\infty}\frac{(A+1)_n}{n!}\left(\frac{\beta}{\beta-\alpha}\right)^n\frac{n}{n-B}\right\} \tag{6.18}$$

follows from (6.15) and proves that the singularities of $\chi_0(z)$ belong to D^* of (4.17) and are simple poles. Furthermore,

$$\operatorname{Res}(\chi_0, x)=-2\frac{\lambda+a}{B'(x)}\beta(x) \quad \text{if } B(x)=0, \tag{6.19}$$

$$\operatorname{Res}(\chi_0, x)=-2(\lambda+a)\beta^{2n+1}(1-\beta^2)^{2\lambda-1}\frac{(2\lambda)_n}{n!}\frac{n}{B'(x)} \tag{6.20}$$

if $B(x)=n\geqq 1$. Using (4.19), (4.20), we get

$$\operatorname{Res}(\chi_0, x_n)=(\lambda+a)\beta^{2n}(1-\beta^2)^{2\lambda}\frac{(2\lambda)_n[a\sqrt{\Delta_n}-b(n+\lambda)]}{n!\sqrt{\Delta_n}[a^2-(n+\lambda)^2]}, \tag{6.21}$$

and

$$\operatorname{Res}(\chi_0, y_n)=(\lambda+a)\beta^{2n}(1-\beta^2)^{2\lambda}\frac{(2\lambda)_n[a\sqrt{\Delta_n}+b(n+\lambda)]}{n!\sqrt{\Delta_n}[a^2-(n+\lambda)^2]}. \tag{6.22}$$

which hold for $n+\lambda>0$. Here, x_n, y_n are given by (4.5). Furthermore, from (4.28) and (4.29), we obtain

$$\operatorname{Res}(\chi_0, x_0)=-2(\lambda+a)\beta(x_0)\frac{[a\sqrt{\Delta_0}-b\lambda]^2}{\sqrt{\Delta_0}(a^2-y^2)^2} \tag{6.23}$$

and

$$\operatorname{Res}(\chi_0, y_0)=2(\lambda-a)\beta(y_0)\frac{[a\sqrt{\Delta_0}+b\lambda]^2}{\sqrt{\Delta_0}(a^2-\lambda^2)^2} \tag{6.24}$$

when $\lambda<0$, $n=0$, in which case $x_0>1$ (and $\beta>0$) and $y_0<-1$ (with $\beta<0$). Notice that $\lambda+a<0$ in the last two formulas. Again, the last four formulas could be made more explicit by using (4.23), (4.24), (4.34) and (4.35). We have the following theorem, which is nothing but a restatement of Theorems 4.22 and 4.24 for the Pollaczek polynomials.

THEOREM 6.1. *If $a>|b|$ then $D^*=\phi$ when $\lambda>0$. If $\lambda<0$ then $D^*=\{x_0, y_0\}$, and $x_0>1$, $y_0<-1$.*

With the subdivision of the (λ, a) plane shown in Fig. 4.2, Theorem 4.25 in the case of Pollaczek polynomials becomes

THEOREM 6.2. *When $b\geqq 0$ and $a\leqq b$, the set D^* is as follows:*

Region I* (i) *$a<b$. Then $D^*=\{x_n\colon n\geqq 0\}$.*
(ii) *$a=b$. Then $D^*=\varnothing$.*

Region II* (i) $-b \leqq a < b$. *Then* $D^* = \{x_n\colon n \geqq 0\}$.
(ii) $a < -b$. *Then* $D^* = \{x_n\colon n \geqq 0\} \cup \{y_n\colon n \geqq 0\}$.
Region III* (i) $-b < a < b$. *Then* $D^* = \{x_n\colon n \geqq 0\}$, $x_0 > 1$.
(ii) $a = -b \neq 0$. *Then* $D^* = \{x_n\colon n \geqq 1\}$.
(iii) $a < -b$. *Then* $D^* = \{x_n\colon n > 1\} \cup \{y_n\colon n > 1\}$.
(iv) $a = b > 0\ (=0)$. *Then* $D^* = \{x_0\}\ (=\varnothing)$.
Region IV* (i) $-b < a$. *Then* $D^* = \{x_n\colon n \geqq 0\}$, $x_0 > 1$.
(ii) $b = -a$. *Then* $D^* = \{x_n\colon n \geqq 1\}$.
(iii) $a < -b$. *Then* $D^* = \{x_n\colon n \geqq 1\} \cup \{y_n\colon n \geqq 1\}$.

In all the regions $x_n < -1$ *and* $y_n > 1$ *for* $n \geqq 1$. *Also,* $x_0 < -1$ *and* $y_0 > 1$ *if* $\lambda > 0$.

The symmetry relation

$$(-1)^n P_n^\lambda(x; a, b) = P_n^\lambda(-x; a, -b) \tag{6.25}$$

follows from (1.14) and (1.15). It shows that the case $a \leqq -b$, $b \leqq 0$, can be obtained from Theorem 6.2 interchanging x_n and y_n, $n \geqq 0$.

A special case of the results derived in § 5 is the following theorem:

THEOREM 6.3. *The measure* $d\psi$ *the polynomials* $\{P_n^\lambda(x)\}$ *are orthogonal with respect to is absolutely continuous in* $\lfloor -1, 1 \rfloor$ *and its only masses belong to* D^*. *Furthermore,* ψ' *is continuous in* $(-1, 1)$ *and*

$$\begin{aligned}\psi'(\cos\theta) = 2^{2\lambda-1}(\lambda+a)\frac{(\sin\theta)^{2\lambda-1}}{\pi}\,\frac{\Gamma(\lambda + i(a\cot\theta + b\csc\theta))^2}{\Gamma(2\lambda)}\\ \cdot \exp\lfloor (a\cot\theta + b\csc\theta)(2\theta - \pi)\rfloor.\end{aligned} \tag{6.26}$$

The support of the measure $d\psi$ *is*

$$\sigma(\psi) = \lfloor -1, 1 \rfloor\, UD^* \tag{6.27}$$

and the set D^* *is empty if and only if* $a \geqq |b|$ *and* $\lambda > 0$, *a finite set with one point if* $a = b > 0$ *or* $a = -b$, $b < 0$ *and two points if* $a > |b|$, *for* $\lambda < 0$; *and a countable infinite subset of* $(-\infty, -1) \cup (1, \infty)$ *having* -1 *or* 1 *as limit points in the other cases. With the weight function*

$$w(x) = \pi\frac{\Gamma(2\lambda)}{\lambda + a}2^{-2\lambda+1}\psi'(x) \tag{6.28}$$

the orthogonality relation is

$$\int_{-\infty}^{\infty} P_n^\lambda(x)P_m^\lambda(x)w(x)\,dx + \sum_{\xi\in D^*} J_\xi P_n^\lambda(\xi)P_m^\lambda(\xi) = h_n\delta_{mn} \tag{6.29}$$

where

$$h_n = \frac{2\pi\Gamma(n+2\lambda)}{2^{2\lambda}(n+\lambda+a)n!}, \qquad J_\xi = \pi\frac{\Gamma(2\lambda)}{\lambda+a}2^{-2\lambda+1}\,\mathrm{Re}\ s(\chi_0, \xi) \tag{6.30}$$

and Res (χ_0, ξ) *is given by* (6.21), (6.22), (6.23) *and* (6.24).

We now turn to the q-Pollaczek polynomials $\{F_n(x)\}$. The generating functions (3.7) and (3.9) and Darboux's Method establish

$$F_n(x) \sim \beta^{-n}(\beta/\xi; q)_\infty(\beta/\zeta; q)_\infty / \lfloor (q; q)_\infty(\beta^2; q)_\infty \rfloor, \tag{6.31}$$

$$F_n^*(x) \sim 2\beta^{1-n}\frac{(1-U\Delta)}{1-\beta^2}\,{}_2\phi_1\!\left(\begin{matrix}\beta/\xi, \beta/\zeta\\ q\beta^2\end{matrix}; q, q\right), \qquad z \in \mathbb{C} - \lfloor -1, 1 \rfloor, \tag{6.32}$$

where ${}_2\phi_1$ is the basic hypergeometric function

$$ {}_2\phi_1\left(\begin{matrix} a, b \\ c \end{matrix}; q, z\right) = \sum_{n=0}^{\infty} \frac{(a;q)_n(b;q)_n}{(c;q)_n(q;q)_n} z^n. $$

The corresponding continued fraction is

$$ \chi(x) = \frac{2\beta(q;q)_\infty(q\beta^2;q)_\infty(1-\Delta U)}{(\beta/\xi;q)_\infty(\beta/\zeta;q)_\infty} {}_2\phi_1\left(\begin{matrix} \beta/\xi, \beta/\zeta \\ q\beta^2 \end{matrix}; q, q\right). \tag{6.33} $$

It is easy to see that $F(x)$ is single-valued whenever it is defined outside $[-1, 1]$. This shows that the continuous spectrum is $[-1, 1]$ and the discrete spectrum coincides with the closure of the solutions of

$$ \beta = \xi q^n \quad \text{or} \quad \beta = \zeta q^n, \qquad n = 0, 1, \cdots. \tag{6.34} $$

One can also show that $[-1, 1]$ does not intersect the discrete spectrum. The absolutely continuous component of the spectral measure can be obtained from (6.33) via the Perron–Stieltjes Inversion Formula but it is easier to use the following theorem of Nevai [20, pp. 141-143].

THEOREM 6.4. *If* $\sum_1^\infty \{|B_n/A_n| + |\sqrt{C_n/(A_nA_{n-1})} - \frac{1}{2}|\}$ *converges, then the continuous spectrum is* $[-1, 1]$ *and the distribution function* ψ *satisfies*

$$ \limsup_n \{\psi'(x)\sqrt{1-x^2}\, p_n^2(x)/\lambda_n\} = \frac{2}{\pi} \quad \text{a.e. on } [-1, 1]. \tag{6.35} $$

Let $d_\psi = \Psi' \, dx$ on $\lfloor -1, 1 \rfloor$. We now go back to (3.7) and derive

$$ F_n(\cos\theta) \sim \frac{(\alpha/\xi;q)_\infty(\alpha/\zeta;q)_\infty}{(\alpha/\beta;q)_\infty(q;q)_\infty}\alpha^n + \text{conjugate}, \qquad 0 < \theta < \pi, \tag{6.36} $$

since $\alpha = 1/\beta = e^{i\theta}$. The recursion (1.29), and (1.6) imply

$$ \lambda_n = (1 - U\Delta)(\Delta^2; q)_n/(q;q)_n. \tag{6.37} $$

Theorem 6.4 enables us to evaluate $\Psi'(x)$. It is given by

$$ \Psi'(\cos\theta) = \frac{C}{\sin\theta}\left|\frac{(e^{2i\theta};q)_\infty}{(e^{i\theta}/\xi;q)_\infty(e^{i\theta}/\zeta;q)_\infty}\right|^2, \tag{6.38} $$

with

$$ C = \frac{1}{2\pi}(1 - U\Delta)(\Delta^2;q)_\infty(q;q)_\infty. \tag{6.39} $$

Finally, one can show that the discrete spectrum is empty when

$$ q, U, \Delta \in \lfloor 0, 1) \quad \text{and} \quad 1 - U^2 \pm 2V > 0. \tag{6.40} $$

Acknowledgments. Richard Askey read a preliminary version of this paper and made several helpful suggestions. The referees caught several minor errors. We gratefully acknowledge the help of Askey and the referees. We also thank Erna Penta and Marlene Salvato for the patience and care they took in typing this manuscript.

REFERENCES

[1] W. AL-SALAM, W. ALLAWAY AND R. ASKEY, *Sieved ultraspherical polynomials*, Trans. Amer. Math. Soc., 284 (1984), pp. 39-55.

[2] R. Askey and M. E. H. Ismail, *The Rogers q-ultraspherical polynomials*, in Approximation Theory, III, E. W. Cheney, ed., Academic Press, New York, 1980, pp. 175-182.

[3] ———, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, 1983, pp. 55-78.

[4] ———, *Recurrence relations, continued fractions and orthogonal polynomials*, Mem. Amer. Math. Soc., 300 (1984), 108 pp.

[5] R. Askey and D. P. Shukla, *Sieved Jacobi polynomials*, to appear.

[6] R. Askey and J. Wilson, *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1984), 58 pp.

[7] E. Bank and M. E. H. Ismail, *The attractive Coulomb potential polynomials*, Constr. Approx., 1 (1985), pp. 103-119.

[8] J. Charris and M. E. H. Ismail, *On sieved ultraspherical polynomials*, II: *Random walk polynomials*, Canad. J. Math., 38 (1986), pp. 397-415.

[9] T. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[10] G. Freud, *Orthogonal Polynomials*, English translation, Pergamon Press, Oxford, 1971.

[11] J. Geronimus, *Polynomials Orthogonal on a Circle and Interval*, English translation, Pergamon Press, Oxford, 1960.

[12] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[13] J. Hadamard, *Le Problème de Cauchy et les Equations aux Dérivées Partielles Linéaires*, Hermann, Paris, 1932.

[14] M. E. H. Ismail, *On sieved orthogonal polynomials*, I: *Symmetric Pollaczek analogues*, this Journal, 16 (1985), pp. 1093-1113.

[15] ———, *On sieved orthogonal polynomials*, III: *orthogonality on several intervals*, Trans. Amer. Math. Soc., 294 (1986), pp. 89-111.

[16] ———, *On sieved orthogonal polynomials*, IV: *Generating functions*, J. Approx. Theory, 46 (1986), pp. 284-296.

[17] N. N. Lebedev, *Special Functions and Their Applications*, English translation, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[18] A. Maté, P. Nevai and V. Totik, *Orthogonal polynomials and absolutely continuous measures*, in Approximation Theory, IV, C. K. Chui et al., eds., Academic Press, New York, 1983, pp. 611-617.

[19] ———, *Asymptotics for orthogonal polynomials defined by a recurrence relation*, to appear.

[20] P. Nevai, *Orthogonal polynomials*, Mem. Amer. Math. Soc., 213 (1979), 182 pp.

[21] ———, *Orthogonal polynomials defined by a recurrence relation*, Trans. Amer. Math. Soc., 250 (1979), pp. 369-384.

[22] A. Novikoff, *On a special system of orthogonal polynomials*, Doctoral dissertation, Stanford University, Stanford, CA, 1954.

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

[24] F. Pollaczek, *Sur une généralisation des polynômes de Legendre*, Comptes Rendus de l'Academie des Sciences, Paris, 228 (1949), pp. 1363-1365.

[25] ———, *Sur une généralisation des polynômes de Jacobi*, Mémorial des Sciences Mathématiques, 131 (1956).

[26] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[27] J. J. Shohat and J. D. Tamarkin, *The Problem of Moments, Mathematical Surveys, Vol.* 1, Amer. Math. Soc., Providence, RI, 1963.

[28] L. J. Slater, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

[29] C. Szegö, *On certain special sets of orthogonal polynomials*, Proc. Amer. Math. Soc., 1 (1950), pp. 731-737; reprinted in Collected Papers, Vol. I, Birkhaüser, Boston, 1982, pp. 795-805.

[30] ———, *Orthogonal Polynomials, fourth ed.*, Colloquium Publications, Vol. 23, Amer. Math. Soc., Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 18, No. 4, July 1987

020

CORRIGENDA: THE SUMMABILITY SIGNIFICANCE OF ROOTS OF BESSEL FUNCTIONS*

E. C. OBI†

On page 1490, the third line in the last displayed equation should read:

$$\pi_m(\nu)=4^m \prod_{s=1}^{m} (\nu+s)^{[m/s]},$$

On page 1492, the fourth line should read:

Thus the first Silverman–Toeplitz condition holds for $(\mathcal{C}, m)$ with $\lambda=1$. Next, since

* This Journal, 17 (1986), pp. 1489-1494.

† Department of Mathematics, University of Nigeria, Nsukka, Nigeria.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

001

A NOTE ON WILSON POLYNOMIALS*

WILLARD MILLER, JR.†

Abstract. Local symmetry (recurrence relation) techniques are a powerful tool for the efficient derivation of properties associated with families of hypergeometric and basic hypergeometric functions. Here these ideas are applied to the Wilson polynomials, a generalization of the classical orthogonal polynomials, to obtain the orthogonality relations and an elementary evaluation of the norm.

Key words. Wilson polynomials, orthogonal polynomials, q-series, basic hypergeometric functions

AMS(MOS) subject classifications. 33A65, 33A75, 39A10

1. Introduction. In [1] Wilson introduced a family of hypergeometric orthogonal polynomials that included as special or limiting cases the classical polynomials and the 6-j symbols of angular momentum. In the Memoir [2] Askey and Wilson introduced a still more general class of basic hypergeometric orthogonal polynomials, the most extensive generalization of classical orthogonal polynomials known. The orthogonality proofs in these papers, while not unmotivated, are quite technical and rely on Mellin-Barnes contour integrals and several hypergeometric summation formulas that are unfamiliar to most mathematicians. The Askey-Wilson and Wilson polynomials are important and useful; they deserve to be more widely known. Furthermore, the appropriate algebraic and group theoretic setting for these general families is as yet unclear. The elementary algebraic treatment of Wilson polynomials presented here is offered in the hope that it will help to increase the "audience" for the polynomials as well as to shed some light on their structure.

In [3] the author, with Agarwal and Kalnins, introduced symmetry techniques for the study of families of basic hypergeometric functions, in analogy with the local Lie theory techniques for ordinary hypergeometric functions. The fundamental objects in this study are the recurrence relations obeyed by the families, expressed in terms of difference or q-difference equations. Generating functions and identities for each family are characterized in terms of the recurrence relations. These ideas were applied in [4] to obtain a strikingly simple derivation of the orthogonality relations for the Askey-Wilson q-polyomials. The treatment of the Wilson (ordinary hypergeometric) polynomials presented here is very similar to that in [4]. However several minor complications arise, due to the fact that whereas the first order q-difference equation $f(qz)=f(z)$ has only the solution $f(z)\equiv$ constant, the first order difference equation $f(z+1)=f(z)$ is satisfied by any periodic function f with period 1. Thus the treatment presented here is not entirely algebraic: a few simple facts about Fourier series and the gamma function are required.

2. The results. The (unnormalized) Wilson polynomials are:

$$\Phi_n^{(a,b,c,d)}(z^2)={}_4F_3\left(\begin{matrix}-n,\, n+a+b+c+d-1,\, a+z,\, a-z\\ a+b,\, a+c,\, a+d\end{matrix};1\right) \tag{2.1}$$

where $n=0,1,2,\cdots$ and $a, b, c, d>0$. The hypergeometric function ${}_4F_3$ is given by the series

$${}_4F_3\left(\begin{matrix}a_1, a_2, a_3, a_4\\ b_1, b_2, b_3\end{matrix};t\right)=\sum_{m=0}^{\infty}\frac{(a_1)_m(a_2)_m(a_3)_m(a_4)_m}{(b_1)_m(b_2)_m(b_3)_m}\frac{t^m}{m!}$$

* Received by the editors September 22, 1986; accepted for publication December 15, 1986. This work was supported in part by the National Science Foundation under grant DMS-86-00372.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

where

$$(a)_m=\begin{cases}1 & \text{if } m=0,\\ a(a+1)\cdots(a+m-1) & \text{if } m=1,2,\cdots.\end{cases}$$

The functions (2.1) are polynomials of order n in z^2. (In [1] the parameters a, b, c, d are also permitted to become complex, an important extension; but we shall not consider that case here.)

Two fundamental recurrence relations for the Wilson polynomials are:

$$\tau^{(a,b,c,d)}\Phi_n^{(a,b,c,d)}=\frac{n(n+a+b+c+d-1)}{(a+b)(a+c)(a+d)}\Phi_{n-1}^{(a+1/2,b+1/2,c+1/2,d+1/2)}, \tag{2.2a}$$

$$\mu^{(a,b,c,d)}\Phi_n^{(a,b,c,d)}=-(a+b-1)\Phi_n^{(a-1/2,b-1/2,c+1/2,d+1/2)} \tag{2.2b}$$

where

$$\begin{aligned}\tau^{(a,b,c,d)}&=\frac{1}{2z}(E_z^{1/2}-E_z^{-1/2}),\\ \mu^{(a,b,c,d)}&=\frac{1}{2z}\left[-\left(a+z-\frac{1}{2}\right)\left(b+z-\frac{1}{2}\right)E_z^{1/2}+\left(a-z-\frac{1}{2}\right)\left(b-z-\frac{1}{2}\right)E_z^{-1/2}\right]\end{aligned} \tag{2.3}$$

and $E_z^\alpha f(z)=f(z+\alpha)$. Here (2.2a) follows from

$$\tau(a+z)_k(a-z)_k=-k(a+\tfrac{1}{2}+z)_{k-1}(a+\tfrac{1}{2}-z)_{k-1}$$

and (2.2b) follows from

$$\mu(a+z)_k(a-z)_k=-(a+b-k-1)(a-\tfrac{1}{2}+z)_k(a-\tfrac{1}{2}-z)_k.$$

The first relation was discussed by Askey and Wilson [2]; I have not found (2.2b) in the literature.

Relation (2.2b) suggests the existence of an operator μ^* mapping $\Phi_n^{(a-1/2,b-1/2,c+1/2,d+1/2)}$ to $\Phi_n^{(a,b,c,d)}$. We find that

$$\mu^{(c+1/2,d+1/2,a-1/2,b-1/2)}\Phi_n^{(a-1/2,b-1/2,c+1/2,d+1/2)}=\frac{-(n+c+d)(n+a+b-1)}{(a+b-1)}\Phi_n^{(a,b,c,d)}, \tag{2.4}$$

which follows from

$$\begin{aligned}&\mu^{(c+1/2,d+1/2,a-1/2,b-1/2)}(a-\tfrac{1}{2}+z)_k(a-\tfrac{1}{2}-z)_k\\ &\quad=-(k+c+d)(a+z)_k(a-z)_k+k(a+c+k-1)(a+d+k-1)(a+z)_{k-1}(a-z)_{k-1}.\end{aligned}$$

We try to introduce a pre-Hilbert space structure such that $\mu^*\equiv\mu^{(c+1/2,d+1/2,a-1/2,b-1/2)}$ is the adjoint operator to $\mu\equiv\mu^{(a,b,c,d)}$. Let $w_{a,b,c,d}(z)$ be analytic as a function of the complex variable z in a neighborhood of the imaginary axis $z=iy$, $-\infty<y<\infty$, real analytic in the variables a, b, c, d, of exponential decrease as $|y|\to\infty$ and such that $w_{a,b,c,d}(iy)\geqq 0$. Define an inner product:

$$(g_1,g_2)_{a,b,c,d}=\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}g_1(z^2)g_2(z^2)w_{a,b,c,d}(z)\,dz$$

where the contour is a deformation of the imaginary axis and g_1, g_2 are real polynomials in z^2. Let $S_{a,b,c,d}$ be the space of such polynomials with this inner product. We have

$$\mu : S_{a,b,c,d} \to S_{a-1/2,b-1/2,c+1/2,d+1/2},$$

$$\mu^* : S_{a-1/2,b-1/2,c+1/2,d+1/2} \to S_{a,b,c,d}$$

and seek a weight function $w_{a,b,c,d}$ such that

$$(2.5) \qquad (f, \mu g)_{a-1/2,b-1/2,c+1/2,d+1/2} = (\mu^* f, g)_{a,b,c,d}$$

for all polynomials $f \in S_{a-1/2,b-1/2,c+1/2,d+1/2}$ and $g \in S_{a,b,c,d}$. A straightforward computation yields the necessary and sufficient condition

$$\frac{w_{a,b,c,d}(z+1)}{w_{a,b,c,d}(z)} = \frac{(z+1)(a+z)(b+z)(c+z)(d+z)}{z(a-z-1)(b-z-1)(c-z-1)(d-z-1)}$$

with general solution

$$\begin{aligned} w_{a,b,c,d}(z) &= \frac{\Gamma(a+z)\Gamma(a-z)\Gamma(b+z)\Gamma(b-z)\Gamma(c+z)\Gamma(c-z)}{\Gamma(2z)\Gamma(-2z)} \\ &\quad \cdot \Gamma(d+z)\Gamma(d-z)h(a, b, c, d, z) \\ &= \hat{w}_{a,b,c,d}(z)h(a, b, c, d, z) \end{aligned}$$

where h satisfies the periodicity properties

$$\begin{aligned} h(a-\tfrac{1}{2}, b-\tfrac{1}{2}, c+\tfrac{1}{2}, d+\tfrac{1}{2}, z+\tfrac{1}{2}) &= h(a-\tfrac{1}{2}, b-\tfrac{1}{2}, c+\tfrac{1}{2}, d+\tfrac{1}{2}, z-\tfrac{1}{2}) \\ &= h(a, b, c, d, z). \end{aligned}$$

Here $\Gamma(z)$ is the gamma function [5, Chap. XII]. From Stirling's series for the gamma function, $\hat{w}_{a,b,c,d}(z) = (z)^{2(a+b+c+d)-3}O(e^{-2\pi|y|})$ as $|y| \to \infty$, where $z = x+iy$. Thus we must require that $h(z) = o(e^{2\pi|y|})$ as $|y| \to \infty$ in order that $\hat{w}h$ be a suitable weight function. Furthermore, the "integration by parts" formula (2.5) will not be valid unless $h(z)$ is analytic in an open set containing the strip $-\frac{1}{2} \leqq x \leqq \frac{1}{2}$. Since $h(z) = h(z+1)$ it follows that h can be analytically continued to an entire periodic function of z:

$$h(z) = \sum_{m=-\infty}^{\infty} c_m(y)\, e^{2\pi imx}, \qquad c_m(y) = \frac{1}{2\pi i} \int_0^{2\pi} h(z)\, e^{-2\pi inx}\, dx.$$

Using the Cauchy-Riemann conditions for analytic functions we find that $c_m(y) = a_m e^{-2\pi my}$ where a_m is independent of z. Since $h(iy) = o(e^{2\pi|y|})$ we have that $|a_m e^{-2\pi my}| = o(e^{2\pi|y|})$ as $|y| \to \infty$, so $a_m = 0$ for $m \neq 0$. Thus h is independent of z and, without loss of generality, we can set $h \equiv 1$:

$$(2.6) \quad w_{a,b,c,d}(z) = \frac{\Gamma(a+z)\Gamma(a-z)\Gamma(b+z)\Gamma(b-z)\Gamma(c+z)\Gamma(c-z)\Gamma(d+z)\Gamma(d-z)}{\Gamma(2z)\Gamma(-2z)}.$$

Since (2.5) holds, $\mu^*\mu$ is formally selfadjoint:

$$(2.7) \qquad (\mu^*\mu g_1, g_2)_{a,b,c,d} = (g_1, \mu^*\mu g_2)_{a,b,c,d}.$$

From recurrence relations (2.2b) and (2.4) it follows that the Wilson polynomials are eigenfunctions of $\mu^*\mu$:

$$(2.8) \qquad \mu^*\mu\Phi_n^{(a,b,c,d)} = \lambda_n \Phi_n^{(a,b,c,d)}, \qquad \lambda_n = (a+b+n-1)(c+d+n).$$

Note that $\lambda_n = \lambda_m$ iff $n = m$. Since the eigenfunctions corresponding to distinct eigenvalues are orthogonal, we have

$$(\Phi_n^{(a,b,c,d)}, \Phi_m^{(a,b,c,d)})_{a,b,c,d} = 0 \quad \text{if } n \neq m.$$

The operator $\mu^*\mu$ and the weight function are symmetric with respect to the interchange $a \leftrightarrow b$. Thus the polynomials $\{\Phi_n^{(b,a,c,d)}(z^2)\}$ are also orthogonal in $S_{a,b,c,d}$ and are eigenfunctions of $\mu^*\mu$. This means that there exists a constant K_n such that

$$\Phi_n^{(b,a,c,d)}(z^2) = K_n \Phi_n^{(a,b,c,d)}(z^2).$$

Equating coefficients of z^{2n} on both sides of this expression to obtain K_n, we find that

$$\begin{aligned}(2.9)\quad & {}_4F_3\left(\begin{matrix}-n, n+b+a+c+d-1, b+z, b-z \\ b+a, b+c, b+d\end{matrix}; 1\right) \\ & = \frac{(a+c)_n(a+d)_n}{(b+c)_n(b+d)_n} {}_4F_3\left(\begin{matrix}-n, n+a+b+c+d-1, a+z, a-z \\ a+b, a+c, a+d\end{matrix}; 1\right).\end{aligned}$$

This is a transformation formula due to Bailey [6, p. 56] and, as Wilson pointed out [1], it essentially contains the symmetries of the 6-j symbols. It follows from this result that the renormalized polynomials

$$(a+b)_n(a+c)_n(a+d)_n \Phi_n^{(a,b,c,d)}(z^2)$$

are symmetric in all four parameters a, b, c, d.

Setting $f(z^2) = g(z^2) = 1$ in (2.5) and using (2.2b) and (2.1), we obtain the following relationship between the norms on $S_{a-1/2,b-1/2,c+1/2,d+1/2}$ and $S_{a,b,c,d}$:

$$(2.10)\qquad \|1\|^2_{a-1/2,b-1/2,c+1/2,d+1/2} = \frac{c+d}{a+b-1}\|1\|^2_{a,b,c,d}.$$

The symmetry of the weight function in a, b, c, d yields 5 more such relations.

Now consider the recurrence (2.2a):

$$\tau^{(a,b,c,d)}: S_{a,b,c,d} \to S_{a+1/2,b+1/2,c+1/2,d+1/2}.$$

We seek the adjoint τ^* to $\tau \equiv \tau^{(a,b,c,d)}$:

$$(2.11)\qquad (f, \tau g)_{a+1/2,b+1/2,c+1/2,d+1/2} = (\tau^* f, g)_{a,b,c,d}$$

for all $f \in S_{a+1/2,\cdots,d+1/2}$, $g \in S_{a,\cdots,d}$. A simple computation using (2.11) yields

$$\begin{aligned}(2.12)\quad \tau^* & \equiv \tau^{*(a+1/2,b+1/2,c+1/2,d+1/2)} \\ & = \frac{1}{2z}[(a+z)(b+z)(c+z)(d+z)E_z^{1/2} - (a-z)(b-z)(c-z)(d-z)E_x^{-1/2}].\end{aligned}$$

From (2.11) and the orthogonality relations it follows that

$$(2.13)\qquad \tau^*\Phi_{n-1}^{(a+1/2,b+1/2,c+1/2,d+1/2)} = H_n \Phi_n^{(a,b,c,d)}.$$

Comparing coefficients of z^{2n} on both sides of this expression we find that

$$H_n = (a+b)(a+c)(a+d).$$

Thus $\tau^*\tau$ is selfadjoint on $S_{a,b,c,d}$ and the eigenvalue equation is:

$$(2.14)\qquad \tau^*\tau\Phi_n^{(a,b,c,d)} = n(n+a+b+c+d-1)\Phi_n^{(a,b,c,d)}.$$

We also have the Rodrigues formula

$$\Phi_n^{(a,b,c,d)} = J_n \tau^{*(a+1/2,\cdots,d+1/2)}\tau^{*(a+1,\cdots,d+1)}\cdots\tau^{*(a+n/2,\cdots,d+n/2)}(1),$$
$$J_n=(a+b)_n(a+c)_n(a+d)_n. \tag{2.15}$$

Substituting $f=\Phi_{n-1}^{(a+1/2,\cdots,d+1/2)}$, $g=\Phi_n^{(a,\cdots,d)}$ in (2.11), we obtain the recurrence

$$\|\Phi_n^{(a,\cdots,d)}\|_{a,\cdots,d}^2=\frac{n(n+a+b+c+d-1)}{(a+b)^2(a+c)^2(a+d)^2}\|\Phi_{n-1}^{(a+1/2,\cdots,d+1/2)}\|_{a+1/2,\cdots,d+1/2}^2 \tag{2.16}$$

which enables us to compute the norm of any Wilson polynomial once the norm $\|1\|_{a,\cdots,d}^2$ is determined for all $a,\cdots,d>0$. We now turn to this last task.

From the orthogonality relation $(\Phi_1^{(a,\cdots,d)}, \Phi_0^{(a,\cdots,d)})_{a,\cdots,d}=0$ and the explicit expression (2.1) for Wilson polynomials we find that

$$\|1\|_{a,\cdots,d}^2=\frac{(a+b+c+d)}{(a+b)(a+c)(a+d)}\|1\|_{a+1,b,c,d}^2. \tag{2.17}$$

Here we have used the evident relation

$$(g_m^a, 1)_{a,\cdots,d}=\|1\|_{a+m,b,c,d}^2, \qquad g_m^a(z^2)=(a+z)_m(a-z)_m.$$

From (2.10) and (2.17) and the obvious invariance of $\|1\|_{a,\cdots,d}$ with respect to a permutation of a, b, c, d we find:

$$\|1\|_{a,b,c,d}^2=\frac{\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)}M(a,b,c,d) \tag{2.18}$$

where M satisfies the periodicity properties

$$M(a,b,c,d)=M(a+\tfrac{1}{2}, b+\tfrac{1}{2}, c+\tfrac{1}{2}, d+\tfrac{1}{2})=M(a+1,b,c,d)$$

and is invariant under any permutation of $a,\cdots,d$. Now replace a by $a+k$, k a positive integer, in (2.18) and write this expression in the following form:

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{\Gamma(a+k+iy)\Gamma(a+k-iy)\Gamma(a+k+b+c+d)}{\Gamma(a+k+b)\Gamma(a+k+c)\Gamma(a+k+d)}\right)\\&\qquad\cdot\left|\frac{\Gamma(b+iy)\Gamma(c+iy)\Gamma(d+iy)}{\Gamma(2iy)}\right|^2dy\\&=\Gamma(b+c)\Gamma(b\ +d)\Gamma(c+d)M(a,b,c,d).\end{aligned} \tag{2.19}$$

From Stirling's series

$$\Gamma(z+k)=\sqrt{2\pi}(k)^{z+k-1/2}e^{-k}\left(1+O\left(\frac{1}{k}\right)\right)$$

as $k\to+\infty$, so

$$\lim_{k\to+\infty}\frac{\Gamma(a+k+iy)\Gamma(a+k-iy)\Gamma(a+k+b+c+d)}{\Gamma(a+k+b)\Gamma(a+k+c)\Gamma(a+k+d)}=1$$

and

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\frac{\Gamma(b+iy)\Gamma(c+iy)\Gamma(d+iy)}{\Gamma(2iy)}\right|^2dy=\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)M(a,b,c,d). \tag{2.20}$$

(The passage to the limit under the integral sign is easily justified since $|\Gamma(a+iy)||\leqq\Gamma(a)$ for $a>0$.) It is evident from (2.20) that M is independent of a. By symmetry, M is constant. To evaluate the constant we set $b=0$, $c=d=\frac{1}{2}$ in (2.20). Using the multiplication and reflection theorems for the gamma function, we reduce (2.20) to

$$2\int_{-\infty}^{\infty}\frac{dy}{\cosh(\pi y)}=M$$

or $M=2$. Thus

$$\begin{aligned}\|1\|^2_{a,b,c,d}&=\frac{1}{\pi}\int_0^{\infty}\left|\frac{\Gamma(a+iy)\Gamma(b+iy)\Gamma(c+iy)\Gamma(d+iy)}{\Gamma(2iy)}\right|^2dy\\&=\frac{2\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)}.\end{aligned}\tag{2.21}$$

Note that this integral, and special cases of it, were originally derived by contour integration, evaluation of the residues at the poles of the integrand in the right half plane and use of known summation theorems to sum the resulting infinite series. Wilson [1] used Bailey's Theorem [6, p. 27]

$$\begin{aligned}&{}_5F_4\left(\begin{matrix}2a,\ a+1,\ a+b,\ a+c,\ a+d\\ a,\ a-b+1,\ a-c+1,\ a-d+1\end{matrix};1\right)\\&=\frac{\Gamma(a-b+1)F(a-c+1)\Gamma(a-d+1)\Gamma(-a-b-c-d+1)}{\Gamma(2a+1)\Gamma(-b-c+1)\Gamma(-b-d+1)\Gamma(-c-d+1)}\end{aligned}$$

to compute (2.21). Since we have independently obtained the value of this integral we can consider the usual contour integral technique as a derivation of Bailey's ${}_5F_4$ summation.

For the Racah polynomials (discrete orthogonality), [1], the recurrence relation methods of this paper yield a purely algebraic derivation of the orthogonality, including as a byproduct the terminating version of Bailey's Theorem: $a+b=-N$.

Note added in proof. Recurrence techniques similar to those used in this paper have been employed by Nikiforov, Suslov and Uvarov [7] and Nikiforov and Suslov [8], but these authors have apparently not applied them to the computation of contour integrals and summation formulas.

Acknowledgments. The author thanks Dick Askey, Mourad Ismail and Dennis Stanton for several helpful conversations.

REFERENCES

[1] J. A. WILSON, *Some hypergeometric orthogonal polynomials*, this Journal, 11 (1980), pp. 690-701.

[2] R. A. ASKEY AND J. A. WILSON, *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. American Mathematical Society, 319, Providence, RI, 1985.

[3] A. K. AGARWAL, E. G. KALNINS AND W. MILLER, *Canonical equations and symmetry techniques for q-series*, this Journal, 18 (1987), to appear.

[4] E. G. KALNINS AND W. MILLER, *Symmetry techniques for q-series: Askey-Wilson polynomials*, Proc. Constructive Function Theory, Univ. Alberta, Edmonton, Alberta, Canada, 1986, to appear.

[5] E. T. WHITTAKER AND G. N. WATSON, *A Course in Modern Analysis*, Cambridge University Press, London, 1958.

[6] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London, 1935.

[7] A. F. NIKIFOROV, S. K. SUSLOV AND V. B. UVAROV, *Classical Orthogonal Polynomials of a Discrete Variable*, Nauka, Moscow, 1985. (In Russian.)

[8] A. F. NIKIFOROV AND S. K. SUSLOV, *Classical orthogonal polynomials of a discrete variable on nonuniform lattices*, Lett. Math. Phys., 11 (1986), pp. 27-34.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

THE BEHAVIOR AT UNIT ARGUMENT OF THE HYPERGEOMETRIC FUNCTION ${}_3F_2$*

WOLFGANG BÜHRING†

Abstract. The behavior at $z=1$ of the generalized hypergeometric function ${}_3F_2(a, b, c; e, f; z)$ is investigated. First the analytic continuation near $z=1$ is obtained for the general case when $s=e+f-a-b-c$ is not equal to an integer. The corresponding continuation formulas for the special cases when s is equal to an integer are then derived by appropriate limiting processes. When $f=c$ or $e=c$, the formulas immediately reduce to the well-known continuation formulas of the Gaussian hypergeometric function.

Key words. special functions, hypergeometric series, hypergeometric functions, continuation formulas, hypergeometric differential equations

AMS(MOS) subject classifications. 33A30, 34A20, 34A30, 30B40

1. Introduction.

The behavior near $z=1$ of the Gaussian hypergeometric function or series

$$(1.1)\qquad {}_2F_1(a, b; e; z)=\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(e)_n n!}z^n$$

is given by a well-known continuation formula which may be written

$$(1.2)\qquad \begin{aligned}\frac{\Gamma(a)\Gamma(b)}{\Gamma(e)}{}_2F_1(a, b; e; z)&=\frac{\Gamma(a)\Gamma(b)\Gamma(s)}{\Gamma(a+s)\Gamma(b+s)}{}_2F_1(a, b; 1-s; 1-z)\\&\quad+\Gamma(-s)(1-z)^s{}_2F_1(a+s, b+s; 1+s; 1-z)\\&\qquad\qquad(|\arg(1-z)|<\pi),\end{aligned}$$

$$(1.3)\qquad s=e-a-b.$$

As it stands, (1.2) is valid if s is not equal to an integer, otherwise it is the starting point from which the relevant formulas may be derived by a limiting process. It is the aim of the present work to obtain a corresponding formula for the generalized hypergeometric function ${}_3F_2(a, b, c; e, f; z)$, including the formulas for the exceptional cases when

$$(1.4)\qquad s=e+f-a-b-c$$

is equal to an integer.

Like the ${}_2F_1$, the ${}_3F_2$ is also a particular solution of a certain linear differential equation, now of the third order, with three regular singular points at $z=0, 1, \infty$. The characteristic exponents 0, 1, s and the local power series solutions of this differential equation relative to the point $z=1$ are known from the detailed investigation by Nørlund [8]. The analytic structure near $z=1$ of the general solution is therefore known and, if s is not equal to an integer, is given by

$$(1.5)\qquad A+B(1-z)+O(\{1-z\}^2)+C(1-z)^s\{1+O(1-z)\}.$$

It is more difficult to obtain the connecting constants A, B, C for the function ${}_3F_2(a, b, c; e, f; z)$. While the value of C (and of K in (1.6) below) can be inferred from

* Received by the editors October 21, 1985; accepted for publication (in revised form) September 17, 1986.

† Physikalisches Institut, Universität Heidelberg, D-6900 Heidelberg, West Germany.

[8], A and B are not known in general. This matter has recently received attention in the special case when $s=0$. Then, as $z\to 1$ in the sector $|\arg(1-z)|<\pi$, we have

$$\begin{aligned}\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ e,f\end{matrix}\middle|\, z\right)\\ =L+M(1-z)+O(\{1-z\}^2)+K\ln(1-z)\{1+O(1-z)\}\end{aligned}\tag{1.6}$$

where

$$K=-1\tag{1.7}$$

and, if $\operatorname{Re}(c)>0$,

$$L=2\psi(1)-\psi(a)-\psi(b)+\sum_{m=1}^{\infty}\frac{(e-c)_m(f-c)_m}{m(a)_m(b)_m}.\tag{1.8}$$

Here $\psi(x)$ denotes the logarithmic derivative $\Gamma'(x)/\Gamma(x)$ of the gamma function. The expression (1.8) for L, contained without proof in Ramanujan's note books [9], has recently been proved by Evans and Stanton [3], but, as they state themselves, "because of the inductive nature" of their proofs, their paper "unfortunately sheds little light on how Ramanujan might have made this remarkable discovery." A new proof along different lines is therefore of interest. Also, the third connecting constant M is not known. Its value will be given in (4.9) below.

In the present paper we first consider the general case when s is not equal to an integer and determine the connecting constants A, B, C of (1.5) for the function ${}_3F_2(a,b,c;e,f;z)$. The results for the exceptional case when s is equal to zero or, more generally, equal to any integer are then obtained by a limiting process.

A lemma is provided in § 2. The main theorem for the general case is derived in § 3. The results for the exceptional cases are supplied in § 4.

2. A lemma. In order to avoid inconvenient and unnecessary restrictions of the denominator parameters, we follow Olver [6] and consider $\{\Gamma(e)\Gamma(f)\}^{-1}{}_3F_2(a,b,c;e,f;z)$ rather than ${}_3F_2$ itself, but without explicitly introducing a new symbol for this quantity. As a function of z, it has a finite value at $z=1$ if $\operatorname{Re}(e+f-a-b-c)>0$. Following Wimp [11], [12] we may consider this value as a function of the five parameters. This suggests the following.

Definition 1. For those points of the parameter space which satisfy $\operatorname{Re}(e+f-a-b-c)>0$ let the symbol ${}_3\mathrm{F}_2(a,b,c;e,f)$ be defined by

$$\frac{1}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ e,f\end{matrix}\right)=\frac{1}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ e,f\end{matrix}\middle|\, 1\right)\tag{2.1}$$

and let it then be defined for the other points by analytic continuation.

Using this definition we supply for later application the following.

Lemma 1. *With $s=e+f-a-b-c$ there holds*

$$\frac{1}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ e,f\end{matrix}\right)=\frac{\Gamma(s)}{\Gamma(c)\Gamma(a+s)\Gamma(b+s)}\,{}_3F_2\left(\begin{matrix}e-c,f-c,s\\ a+s,b+s\end{matrix}\right).\tag{2.2}$$

Proof. If $\operatorname{Re}(s)>0$ and $\operatorname{Re}(c)>0$ we have from [2] or [5]

$$\frac{1}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ e,f\end{matrix}\middle|\, 1\right)=\frac{\Gamma(s)}{\Gamma(c)\Gamma(a+s)\Gamma(b+s)}\,{}_3F_2\left(\begin{matrix}e-c,f-c,s\\ a+s,b+s\end{matrix}\middle|\, 1\right).\tag{2.3}$$

Here the conditions on s and c are needed to ensure the convergence of the hypergeometric series of unit argument on the left- and right-hand side, respectively. By means of Definition 1 and analytic continuation with respect to the parameters we get rid of these restrictions, which completes the proof.

3. Derivation of the main theorem. The generalized hypergeometric function ${}_3F_2(a, b, c; e, f; z)$ is a particular solution of the third order linear differential equation

$$\{D(zD+e-1)(zD+f-1)-(zD+a)(zD+b)(zD+c)\}w(z)=0 \tag{3.1}$$

where $D=d/dz$. Relative to the regular singular point $z=1$ of the differential equation, the Frobenius ansatz

$$w(z)=\sum_{n=0}^{\infty} g_n(r)(1-z)^{r+n}, \tag{3.2}$$

written with powers of $1-z$ rather than $z-1$ for later convenience, yields for the coefficients g_n the recurrence relation

$$\begin{aligned}(r+n)(r+n-1)(r+n+a+b+c-e-f)g_n(r)\\ =(r+n-1)\{(r+n-2)(2r+2n-1-e-f+2a+2b+2c)\\ -ef+ab+bc+ca+a+b+c+1\}g_{n-1}(r)\\ -(r+n-2+a)(r+n-2+b)(r+n-2+c)g_{n-2}(r),\end{aligned} \tag{3.3}$$

valid for $n=0, 1, 2, \cdots$, provided that we have defined $g_{-2}(r)=g_{-1}(r)=0$ while $g_0(0)\neq 0$. The equation for $n=0$ then leads to the indicial equation which determines the characteristic exponents

$$r\in\{0, 1, s\}, \quad s \text{ as in } (1.4). \tag{3.4}$$

We now assume in this section that s is not equal to an integer, so that the exponent $r=s$ yields a solution of the differential equation. Of the two exponents which are integers the larger one always gives a solution, but for the smaller one we have to check if (3.3) is consistent for all the n. Since the difference of the exponents is 1, the equation for $n=1$ is crucial. With (1.4) we have from (3.3)

$$\begin{aligned}(r+1)r(r+1-s)g_1(r)=r\{(r-1)(2r+1-e-f+2a+2b+2c)\\ -ef+ab+bc+ca+a+b+c+1\}g_0(r).\end{aligned} \tag{3.5}$$

For $r=0$ not only the left-hand side vanishes but the right-hand side happens to vanish also, and so (3.5) is consistent irrespective of the value of $g_1(0)$. Therefore $g_1(0)$, besides $g_0(0)$, may be assigned arbitrarily. The exponent $r=0$ in this way yields a solution containing two constants of integration, where changing $g_1(0)$ is equivalent to adding a constant multiple of the solution with the exponent $r=1$ mentioned above. Thus we have shown that

$$\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a, b, c\\ e, f\end{matrix}\,\middle|\, z\right)=\sum_{n=0}^{\infty} g_n(0)(1-z)^n+(1-z)^s\sum_{n=0}^{\infty} g_n(s)(1-z)^n \tag{3.6}$$

where the connecting constants $g_0(0)$, $g_1(0)$, $g_0(s)$ still have to be determined while the other coefficients $g_n(r)$ are then given by the recurrence relation

$$g_n(r)=\frac{1}{(r-s+n)(r+n)(r+n-1)}$$

$$\begin{aligned}\cdot((r+n-1)\{(r+n-2)(2r-s+2n-1+a+b+c)\\ -ef+ab+bc+ca+a+b+c+1\}g_{n-1}(r)\\ -(r+n-2+a)(r+n-2+b)(r+n-2+c)g_{n-2}(r)),\end{aligned}\tag{3.7}$$

to be used for $r=0$ with $n=2,3,4,\cdots$, or for $r=s$ with $n=1,2,3,\cdots$, and $g_{-1}(s)=0$.

The connecting constant $g_0(s)$, which multiplies the (at $z=1$) singular contribution, can most conveniently be determined by Darboux's method [6] as follows. The left-hand side of (3.6), when written in the form $\sum u_n z^n$, has coefficients

$$u_n=\frac{\Gamma(a+n)\Gamma(b+n)\Gamma(c+n)}{\Gamma(e+n)\Gamma(f+n)\Gamma(1+n)}\tag{3.8}$$

with the asymptotic behavior

$$u_n\sim n^{a+b+c-e-f-1}\left\{1+O\left(\frac{1}{n}\right)\right\}\tag{3.9}$$

as $n\to\infty$, which follows by means of the formula

$$\frac{\Gamma(x+n)}{\Gamma(y+n)}\sim n^{x-y}\left\{1+O\left(\frac{1}{n}\right)\right\}.\tag{3.10}$$

The leading singular term on the right-hand side of (3.6) is $g_0(s)(1-z)^s$, which, when expanded in powers of z by means of the formula

$$(1-z)^s=\sum_{n=0}^{\infty}\frac{(-s)_n}{n!}z^n\tag{3.11}$$

and written in the form $\sum v_n z^n$, has coefficients

$$v_n=\frac{g_0(s)}{\Gamma(-s)}\frac{\Gamma(-s+n)}{\Gamma(1+n)}\tag{3.12}$$

with the asymptotic behavior

$$v_n\sim\frac{g_0(s)}{\Gamma(-s)}n^{-s-1}\left\{1+O\left(\frac{1}{n}\right)\right\}\tag{3.13}$$

as $n\to\infty$. Thus u_n and v_n have, in view of the definition of s according to (1.4), the same n-dependence of the leading asymptotic term, as expected, and comparison of the constant factors in (3.9) and (3.13) yields

$$g_0(s)=\Gamma(-s).\tag{3.14}$$

This result has been obtained in a different way by Nørlund [8]. More recently such connection problems have been treated by Naundorf [7] and by Schäfke and Schmidt [10].

The connecting constants $g_0(0)$ and $g_1(0)$, which appear in the (at $z=1$) regular term, will now be determined. If $\operatorname{Re}(s)>0$, the singular contribution vanishes and (3.6) evaluated at $z=1$ yields

$$g_0(0)=\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix}a,b,c\\e,f\end{matrix}\middle|\,1\right). \tag{3.15}$$

In a similar way, if $\operatorname{Re}(s)>1$, the derivative with respect to z of (3.6) gives

$$g_1(0)=-\frac{\Gamma(a+1)\Gamma(b+1)\Gamma(c+1)}{\Gamma(e+1)\Gamma(f+1)}\,{}_3F_2\left(\begin{matrix}a+1,b+1,c+1\\e+1,f+1\end{matrix}\middle|\,1\right). \tag{3.16}$$

By analytic continuation with respect to the parameters, Definition 1 and Lemma 1, we then finally obtain

$$g_0(0)=\frac{\Gamma(a)\Gamma(b)\Gamma(s)}{\Gamma(a+s)\Gamma(b+s)}\,{}_3F_2\left(\begin{matrix}e-c,f-c,s\\a+s,b+s\end{matrix}\right), \tag{3.17}$$

$$g_1(0)=-\frac{\Gamma(a+1)\Gamma(b+1)\Gamma(s-1)}{\Gamma(a+s)\Gamma(b+s)}\,{}_3F_2\left(\begin{matrix}e-c,f-c,s-1\\a+s,b+s\end{matrix}\right). \tag{3.18}$$

All the other coefficients $g_n(r)$ appearing in (3.6) are now uniquely defined by the recurrence relation (3.7). On the computer they may efficiently be evaluated in this way, but for analytical work, in particular the investigation of the so far excluded exceptional cases when s is equal to an integer, an explicit representation of all the $g_n(r)$ is desirable.

When $r=0$ we may consider the nth derivative with respect to z of (3.6) and proceed essentially in the same way as above with $g_1(0)$. As a generalization of (3.17), (3.18) we then immediately obtain

$$g_n(0)=(-1)^n\frac{\Gamma(a+n)\Gamma(b+n)\Gamma(s-n)}{\Gamma(a+s)\Gamma(b+s)n!}\,{}_3F_2\left(\begin{matrix}e-c,f-c,s-n\\a+s,b+s\end{matrix}\right). \tag{3.19}$$

When $r=s$ we suspect that the $g_n(s)$ can be represented in a similar form. From (3.7) and (3.14) we have, after rearrangement of the terms,

$$g_1(s)=-\Gamma(-s-1)\{s^2+(a+b)s-ef+ab+c(a+b+s)\}. \tag{3.20}$$

By the definition of s the factor of c is equal to $e+f-c$. Making use of this fact we lose the symmetry with respect to the numerator parameters and obtain

$$\begin{aligned}g_1(s)&=-\Gamma(-s-1)\{(a+s)(b+s)-(e-c)(f-c)\}\\&=-\Gamma(-s-1)(a+s)(b+s)\,{}_3F_2\left(\begin{matrix}e-c,f-c,-1\\a+s,b+s\end{matrix}\middle|\,1\right).\end{aligned} \tag{3.21}$$

To proceed further, it is convenient to get rid of some factors and to consider $y_n(r)$ defined by

$$n!g_n(r)=(-1)^n\Gamma(s-2r-n)y_n(r) \tag{3.22}$$

rather than $g_n(r)$ itself. In view of $r\in\{0,s\}$, the recurrence relation for $y_n(r)$ then reads

$$\begin{aligned}y_n(r)=\{&(r+n-2)(2n-1+2r-s+a+b+c)\\&-ef+ab+bc+ca+a+b+c+1\}y_{n-1}(r)\\&+(s-r-n+1)(a+r+n-2)(b+r+n-2)(c+r+n-2)y_{n-2}(r),\end{aligned} \tag{3.23}$$

and we are interested in its special solution which, according to (3.14), (3.17), (3.18), (3.21), has the starting values

$$y_0(r)=\frac{\Gamma(a+r)\Gamma(b+r)}{\Gamma(a+s)\Gamma(b+s)}{}_3F_2\left(\begin{matrix}e-c, f-c, s-r\\ a+s, b+s\end{matrix}\right), \tag{3.24}$$

$$y_1(r)=\frac{\Gamma(a+r+1)\Gamma(b+r+1)}{\Gamma(a+s)\Gamma(b+s)}{}_3F_2\left(\begin{matrix}e-c, f-c, s-r-1\\ a+s, b+s\end{matrix}\right). \tag{3.25}$$

The factor of $y_{n-1}(r)$ in (3.23) may be rewritten, using the definition of s, as

$$\begin{aligned}&(n-2)(2n-1+4r-s+a+b+c)-(e+r)(f+r)+(a+r)(b+r)\\ &\quad+(b+r)(c+r)+(c+r)(a+r)+a+b+c+3r+1.\end{aligned} \tag{3.26}$$

We then may see that the recurrence relation for $y_n(r)$ can be obtained from that for $y_n(0)$ by the simultaneous substitutions

$$a\to a+r,\quad b\to b+r,\quad c\to c+r,\quad e\to e+r,\quad f\to f+r, \tag{3.27}$$

and as a consequence,

$$s\to s-r. \tag{3.28}$$

The same is true for the starting values (3.24), (3.25). Since we already know from (3.19) that

$$y_n(0)=\frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(a+s)\Gamma(b+s)}{}_3F_2\left(\begin{matrix}e-c, f-c, s-n\\ a+s, b+s\end{matrix}\right), \tag{3.29}$$

it follows by the substitutions (3.27)-(3.28) that

$$y_n(r)=\frac{\Gamma(a+r+n)\Gamma(b+r+n)}{\Gamma(a+s)\Gamma(b+s)}{}_3F_2\left(\begin{matrix}e-c, f-c, s-r-n\\ a+s, b+s\end{matrix}\right). \tag{3.30}$$

By means of (3.22) we obtain the required explicit representation for $g_n(r)$ which is valid for both $r=0$ and $r=s$ and so we have proved the following.

THEOREM 1. *If $s=e+f-a-b-c$ is not equal to an integer, then the analytic continuation of the ${}_3F_2$-series near $z=1$ in the sector $|\arg(1-z)|<\pi$ is given by the formula*

$$\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}{}_3F_2\left(\begin{matrix}a, b, c\\ e, f\end{matrix}\middle| z\right)=\sum_{n=0}^{\infty}g_n(0)(1-z)^n+(1-z)^s\sum_{n=0}^{\infty}g_n(s)(1-z)^n \tag{3.31}$$

where the series on the right-hand side converge inside the circle $|z-1|=1$ and the coefficients $g_n(r)$, with $r\in\{0, s\}$, are

$$g_n(r)=(-1)^n\frac{\Gamma(a+r+n)\Gamma(b+r+n)\Gamma(s-2r-n)}{\Gamma(a+s)\Gamma(b+s)n!}{}_3F_2\left(\begin{matrix}e-c, f-c, s-r-n\\ a+s, b+s\end{matrix}\right). \tag{3.32}$$

We may observe that when $f=c$ or $e=c$ then the ${}_3F_2$ on the right-hand side of (3.32) becomes equal to 1 and the Theorem 1 reduces to the continuation formula for the ${}_2F_1$.

The recurrence (3.7) for the quantities (3.32) can also be deduced from the work of Lewanowicz [4] and is, essentially, a special case of the recursion relation for the Wilson polynomials.

4. The exceptional cases. When s is an integer, the hypergeometric series is called s-balanced. In such a case the continuation formula may be derived from Theorem 1 by an approximate limiting process. For this purpose it is convenient to make the simultaneous substitutions $a \to a-\varepsilon$, $b \to b-\varepsilon$, $c \to c-\varepsilon$, $e \to e-\varepsilon$, $f \to f-\varepsilon$, and as a consequence, $s \to s+\varepsilon$ and then to consider the resulting equations for integral s, separately for $s \geqq 0$ or $s \leqq 0$, respectively. Since the series representation of the ${}_3F_2$ in (3.32) is needed, a restriction of the parameter c now comes in to ensure the convergence. By performing the limit $\varepsilon \to 0$, which is a somewhat lengthy but standard [5] procedure, we may derive the following results.

COROLLARY 1. *If $s = e+f-a-b-c$ is equal to an integer $t \geqq 0$, then the analytic continuation of the ${}_3F_2$-series near $z = 1$ in the sector $|\arg(1-z)| < \pi$ is given by the formula*

$$\tag{4.1} \begin{aligned} &\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix} a, b, c \\ e, f \end{matrix}\,\middle|\, z\right) \\ &\qquad = \sum_{n=0}^{t-1} k_n(1-z)^n + (1-z)^t \sum_{n=0}^{\infty} \{p_n + q_n \ln(1-z)\}(1-z)^n \end{aligned}$$

where the series on the right-hand side converges inside the circle $|z-1| = 1$ and the coefficients are

$$\tag{4.2} k_n = (-1)^n \frac{\Gamma(a+n)\Gamma(b+n)(t-n-1)!}{\Gamma(a+t)\Gamma(b+t)n!}\,{}_3F_2\left(\begin{matrix} e-c, f-c, t-n \\ a+t, b+t \end{matrix}\right),$$

$$\tag{4.3} \begin{aligned} p_n = \frac{(a+t)_n(b+t)_n}{(t+n)!n!}&\left((-1)^t \sum_{m=0}^{n} \frac{(e-c)_m(f-c)_m(-n)_m}{(a+t)_m(b+t)_m m!}\right. \\ &\cdot\{\psi(1+n-m)+\psi(1+t+n)-\psi(a+t+n)-\psi(b+t+n)\} \\ &\left. +(-1)^{t+n} n! \sum_{m=n+1}^{\infty} \frac{(e-c)_m(f-c)_m(m-n-1)!}{(a+t)_m(b+t)_m m!}\right) \\ &\textit{if } \operatorname{Re}(c) > -t-n, \end{aligned}$$

$$\tag{4.4} q_n = -(-1)^t \frac{(a+t)_n(b+t)_n}{(t+n)!n!}\,{}_3F_2\left(\begin{matrix} e-c, f-c, -n \\ a+t, b+t \end{matrix}\,\middle|\, 1\right).$$

COROLLARY 2. *If $s = e+f-a-b-c$ is equal to an integer $-t \leqq 0$, then the analytic continuation of the ${}_3F_2$-series near $z = 1$ in the sector $|\arg(1-z)| < \pi$ is given by the formula*

$$\tag{4.5} \begin{aligned} &\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(e)\Gamma(f)}\,{}_3F_2\left(\begin{matrix} a, b, c \\ e, f \end{matrix}\,\middle|\, z\right) \\ &\qquad = (1-z)^{-t} \sum_{n=0}^{t-1} h_n(1-z)^n + \sum_{n=0}^{\infty} \{u_n + v_n \ln(1-z)\}(1-z)^n \end{aligned}$$

where the series on the right-hand side converges inside the circle $|z-1| = 1$ and the coefficients are

$$\tag{4.6} h_n = (-1)^n \frac{(a-t)_n(b-t)_n(t-n-1)!}{n!}\,{}_3F_2\left(\begin{matrix} e-c, f-c, -n \\ a-t, b-t \end{matrix}\,\middle|\, 1\right),$$

$$u_n = \frac{(a-t)_{t+n}(b-t)_{t+n}}{n!(t+n)!}$$
$$\tag{4.7} \cdot\Bigg((-1)^t \sum_{m=0}^{t+n} \frac{(e-c)_m(f-c)_m(-t-n)_m}{(a-t)_m(b-t)_m m!} \cdot\{\psi(1+t+n-m)+\psi(1+n)-\psi(a+n)-\psi(b+n)\} + (-1)^n(t+n)! \sum_{m=t+n+1}^{\infty} \frac{(e-c)_m(f-c)_m(m-t-n-1)!}{(a-t)_m(b-t)_m m!}\Bigg)$$

if Re $(c) > -n$,

$$\tag{4.8} v_n = -(-1)^t \frac{(a-t)_{t+n}(b-t)_{t+n}}{(t+n)!n!} {}_3F_2\left(\begin{matrix} e-c, f-c, -t-n \\ a-t, b-t \end{matrix}\middle| 1\right).$$

When $t=0$, the empty sums in (4.1) or (4.5) have to be interpreted as 0.

If $f=c$ or $e=c$, the formulas in Corollaries 1 and 2 immediately reduce to the corresponding formulas for the Gaussian hypergeometric function, which appear, apart from some minor differences of presentation, as (15.3.10)-(15.3.12) in [1], for instance.

Taking the leading terms of (4.1) or (4.5) with $t=0$ we may obtain (1.6) with the constants K and L as expected and the third connecting constant

$$\tag{4.9} M = \{ab-(e-c)(f-c)\}\{2+2\psi(1)-\psi(a+1)-\psi(b+1)\} + (e-c)(f-c) - ab \sum_{m=2}^{\infty} \frac{(e-c)_m(f-c)_m}{(m-1)m(a)_m(b)_m} \quad (\text{Re}\,(c) > -1).$$

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] W. N. BAILEY, *Generalized Hypergeometric Series*, Stechert-Hafner, New York, 1964.

[3] R. J. EVANS AND D. STANTON, *Asymptotic formulas for zero-balanced hypergeometric series*, this Journal, 15 (1984), pp. 1010-1020.

[4] S. LEWANOWICZ, *Recurrence relations for hypergeometric functions of unit argument*, Math. Comp., 45 (1985), pp. 521-535.

[5] Y. L. LUKE, *The Special Functions and Their Approximations, Vol.* 1, Academic Press, New York, 1969.

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

[7] F. NAUNDORF, *Ein Verfahren zur Lösung des Zusammenhangproblems bei linearen Differentialgleichungen zweiter Ordnung mit mehreren singulären Stellen*, Z. Angew. Math. Mech., 59 (1979), pp. 273-275.

[8] N. NØRLUND, *Hypergeometric functions*, Acta Math., 94 (1955), pp. 289-349.

[9] S. RAMANUJAN, *Notebooks*, 2 volumes, Tata Institute of Fundamental Research, Bombay, India, 1957.

[10] R. SCHÄFKE AND D. SCHMIDT, *The connection problem for general linear ordinary differential equations at two regular singular points with applications in the theory of special functions*, this Journal, 11 (1980), pp. 848-862.

[11] J. WIMP, *The computation of* ${}_3F_2(1)$, Internat. J. Comput. Math., 10 (1981), pp. 55-62.

[12] ———, *Computation with Recurrence Relations*, Pitman, Boston, 1984.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

FABER EXPANSIONS OF RATIONAL AND ENTIRE FUNCTIONS*

ELGIN H. JOHNSTON†

Abstract. The Faber polynomials for a bounded, simply connected domain D can, under certain circumstances, be used to give series expansions of functions analytic on D [5] and to give good polynomial approximations to such functions [3]. We provide explicit, easily computable formulae for the coefficients of the Faber polynomials and formulae for the Faber coefficients of certain types of rational and analytic functions. Finally, since the study of the properties of an analytic function may require us to look at its derivative, we also provide formulae for the Faber expansions of the derivatives of Faber polynomials.

Key words. Faber polynomials, Faber expansion, rational function, entire function

AMS(MOS) subject classifications. Primary 30B50; secondary 30C99

1. Introduction.

Let

$$g(\xi)=\xi+b_0+\frac{b_1}{\xi}+\frac{b_2}{\xi^2}+\cdots \tag{1}$$

be analytic and univalent in $\Delta=\{\xi: |\xi|>1\}$. Then $E=\mathbf{C}\backslash g(\Delta)$ is a continuum of capacity 1. The Faber polynomials $\{F_n(w)\}_{n=0}^{\infty}$ associated with g (or E) are defined by the generating function relation [4, p. 57]

$$\log\left(\frac{g(\xi)-w}{\xi}\right)=-\sum_{n=1}^{\infty}\frac{1}{n}F_n(w)\xi^{-n}. \tag{2}$$

We define $F_0(w)\equiv 1$. As well as providing a valuable tool in studying properties of univalent functions, the Faber polynomials also provide a means for expressing functions analytic on E or Int E. In fact, if $f(z)$ is such a function then under suitable conditions,

$$f(z)=\sum_{n=0}^{\infty}a_nF_n(z) \qquad (z\in \operatorname{Int} E) \tag{3}$$

where

$$a_n=\frac{1}{2\pi i}\int_{|\xi|=r}\frac{f(g(\xi))}{\xi^{n+1}}\,d\xi. \tag{4}$$

In (4), r is a positive number, usually close to 1, and of course $f(g(\xi))$ must be defined on $|\xi|=r$; this gets back to the "suitable conditions" alluded to above and will be made more precise in later sections. In spite of (4), very little seems to have been done concerning methods for actually obtaining Faber expansions. In this paper we address this and other problems concerning Faber polynomials.

In § 2 we find explicit formulae for the coefficients of the Faber polynomials. Various expressions for the Faber coefficients have been obtained by Todorov [7].

* Received by the editors January 20, 1986; accepted for publication September 17, 1986.

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

However we consider the coefficients for the expansion about b_0,

$$F_n(w) = \sum_{n=0}^{n} c_k^{(n)}(w-b_0)^n.$$

Such an expression seems natural in view of the recursion relation [4, p. 57]

$$F_{n+1}(w) = (w-b_0)F_n(w) - \sum_{k=1}^{n-1} b_{n-k}F_k(w) - (n+1)b_n. \tag{5}$$

The resulting expressions for the $c_k^{(n)}$, being independent of b_0, seem somewhat simpler than Todorov's expressions.

In § 3 we find explicit formulae for the Faber expansions of the polynomials $(w-b_0)^k$ $(k=0, 1, \cdots)$ and use these to obtain Faber expansions of entire and some analytic functions. As examples, we present some expansions of certain entire functions in terms of Chebyshev polynomials.

In § 4 we find explicit formulae for the Faber expansions of derivatives of the Faber polynomials. Such expressions may prove useful for working with Faber expansions of functions, and also lead to explicit formulae for the Faber expansions of rational functions.

In the remainder of the paper we will encounter many summations which, in certain cases, may be summing over empty sets. Such sums are taken to be 0.

2. Coefficients of Faber polynomials.

THEOREM 1. *For* $n=0, 1, 2, \cdots,$

$$F_n(w) = (w-b_0)^n + \sum_{k=0}^{n-2} c_k^{(n)}(w-b_0)^k \tag{6}$$

where, for $1 \leqq k \leqq n-2$,

$$c_k^{(n)} = \frac{n}{k}\sum (-1)^m \binom{m+k-1}{k-1} b_{r_1-1}\cdots b_{r_m-1} \tag{7}$$

and

$$c_0^{(n)} = \sum_{j=2}^{n-2} jb_{j-1}\left(\sum (-1)^{m+1} b_{r_1-1}\cdots b_{r_m-1}\right) - nb_{n-1}. \tag{8}$$

In (7), *the sum is over all ordered m-tuples* $(r_1, \cdots, r_m)$ $(m=1, 2, \cdots)$ *of integers* $\geqq 2$ *with* $r_1+r_2+\cdots+r_m = n-k$. *In* (8), *the inner sum is over all ordered m-tuples* $(r_1, r_2, \cdots, r_m)$ $(m=1, 2, \cdots,)$ *of integers* $\geqq 2$ *with* $r_1+r_2+\cdots+r_m = n-j$. *In* (6), (7) *and* (8) *all empty sums will be taken to be* 0. *For convenience, we will let* $c_n^{(n)}=1$ *and* $c_{n-1}^{(n)}=0$.

Remark. Todorov [7] found expressions for the coefficients of the Faber polynomials when expanded in powers of w (rather than $w-b_0$). The formulae in [7] involve homogeneous isobaric polynomials [6] and seem to be more complicated than those presented in Theorem 1. However, we have sacrificed, for this simplicity, the convenience of having the polynomials in powers of w.

For the proof of Theorem 1 it is convenient to first find an expression for the reciprocal of a power series. Such expressions have been found previously [2], but we need an alternate form that does not appear to be in a convenient place in the literature.

LEMMA 2. *Let $p(z)=\alpha_0+\alpha_1 z+\alpha_2 z^2+\cdots$ be a formal power series with $\alpha_0\neq 0$. Then the formal power series for $1/p(z)$ is $\beta_0+\beta_1 z+\beta_2 z^2+\cdots$ where $\beta_0=\alpha_0^{-1}$ and*

$$\beta_k=\sum\frac{(-1)^m}{\alpha_0^{m+1}}\alpha_{r_1}\alpha_{r_2}\cdots\alpha_{r_m}\qquad (k=1,2,\cdots). \tag{9}$$

The sum in (9) *is over all m-tuples $(r_1,\cdots,r_m)$ of positive integers with $r_1+r_2+\cdots+r_m=k$.*

Proof. That $\beta_0=\alpha_0^{-1}$ is clear. Assume (9) correct for indices $j\leqq k-1$ ($k\geqq 1$). Since

$$\left(\sum_0^\infty \alpha_k z^k\right)\left(\sum_0^\infty \beta_k z^k\right)\equiv 1,$$

we have $\sum_{j=0}^k \beta_j\alpha_{k-j}=0$. Thus

$$\begin{aligned}\beta_k&=-\frac{1}{\alpha_0}\sum_{j=0}^{k-1}\beta_j\alpha_{k-j}\\ &=-\frac{1}{\alpha_0}\sum_{j=1}^{k-1}\alpha_{k-j}\left(\sum\frac{(-1)^m}{\alpha_0^{m+1}}\alpha_{r_1}\cdots\alpha_{r_m}\right)-\frac{\alpha_k}{\alpha_0}.\end{aligned}\tag{10}$$

The inner sum is over all m-tuples $(r_1,\cdots,r_m)$ of positive integers with $r_1+r_2+\cdots+r_m=j$. The inner sum, combined with the first sum, then gives a sum over all ordered m-tuples ($m\geqq 2$) $(r_1,\cdots,r_m)$ with $r_1+r_2+\cdots+r_m=k$. (The double sum in (10) simply counts all such m-tuples with $k-1$ in the last position, then those with $k-2$ in the last position, etc.). The $-\alpha_k/\alpha_0$ term takes care of the "1-tuple" (k) appearing in (9). Bringing the factor $-1/\alpha_0$ inside the expression (10) thus gives (9) and completes the proof by induction. □

Before using (9) to prove Theorem 1, we first rearrange (9) to group like powers of α_1. We observe that if l 1's appear in the m-tuple $(r_1,r_2,\cdots,r_m)$, then after the 1's are deleted, the remaining $(m-l)$-tuple $(r_1',\cdots,r_{m-l}')$ has coordinates $\geqq 2$ and summing to $k-l$. Now starting from $(r_1',\cdots,r_{m-l}')$ we may insert l 1's in this $(m-l)$-tuple and fill it out to an m-tuple in $\binom{m}{l}$ ways.

Thus

$$\beta_k=\sum_{l=0}^{k-2}\left(\sum\binom{m+l}{l}a_0^{(-1)^{m+l}/(m+l+1)}\alpha_{r_1}\alpha_{r_2}\cdots\alpha_{r_m}\right)\alpha_1^l+\frac{(-1)^k}{\alpha_0}\frac{\alpha_1^k}{\alpha_0}\tag{11}$$

where the inner sum is over all m-tuples $(r_1,\cdots,r_m)$ of integers $\geqq 2$ with $r_1+r_2+\cdots+r_m=k-l$ and empty sums are taken to be 0. We then note that (11) holds for all $k\geqq 0$.

Proof of Theorem 1. Differentiating (2) with respect to w and then multiplying by ξ we get

$$\frac{\xi}{g(\xi)-w}=\sum_{n=0}^\infty\frac{1}{n+1}F_{n+1}'(w)\xi^{-n}.\tag{12}$$

We apply Lemma 2 to $p(z)=[g(1/z)-w]z$, noting that $p(z)=\sum_{k=0}^\infty \alpha_k z^k$ with

$$\alpha_0=1,\quad \alpha_1=b_0-w\quad\text{and}\quad \alpha_k=\beta_{k-1}\quad (k\geqq 2).\tag{13}$$

Replacing z by $1/\xi$ we find that

$$\frac{\xi}{g(\xi)-w}=\sum_{n=0}^\infty\beta_n\xi^{-n}\tag{14}$$

where we may take the β_n's as given by (11). Equating coefficients in (12) and (14), then using (11) and (13) we have, for $n \geqq 1$,

$$\begin{aligned}\frac{1}{n}F'_n(w)&=\beta_{n-1}\\&=(w-b_0)^{n-1}+\sum_{l=0}^{n-3}\left(\sum\binom{m+l}{l}(-1)^m b_{r_1-1}b_{r_2-1}\cdots b_{r_m-1}\right)(w-b_0)^l\end{aligned}\tag{15}$$

where the inner sum in (15) is over the same values as the one in (11) for β_{n-1}. When we integrate both sides of (15) and multiply by n we have that

$$\begin{aligned}F_n(w)&=(w-b_0)^n+n\sum_{l=0}^{n-3}\left(\sum(-1)^m\binom{m+l}{l}b_{r_1-1}\cdots b_{r_m-1}\right)\frac{(w-b_0)^{l+1}}{l+1}+F_n(b_0)\\&=(w-b_0)^n+\sum_{l=1}^{n-2}\frac{n}{l}\left(\sum(-1)^m\binom{m+l-1}{l-1}b_{r_1-1}\cdots b_{r_m-1}\right)(w-b_0)^l+F_n(b_0).\end{aligned}\tag{16}$$

The inner sum in (16) is over all m-tuples $(r_1,\cdots,r_m)$ of integers $\geqq 2$ with $r_1+r_2+\cdots+r_m=n-l$. Thus we have, as defined in the statement of Theorem 1,

$$F_n(w)=(w-b_0)^n+\sum_{k=1}^{n-2}c_k^{(n)}(w-b_0)^k+F_n(b_0).$$

We need only prove that $c_0^{(n)}=F_n(b_0)$ is as claimed.

If we differentiate (5) and let $w=b_0$ we have

$$F_n(b_0)=F'_{n+1}(b_0)+\sum_{k=1}^{n-1}b_{n-k}F'_k(b_0).$$

Substituting (15), with $w=b_0$, into this last expression and noting $F'_1\equiv 1$, we have

$$\begin{aligned}F_n(b_0)&=(n+1)\sum_{\textcircled{1}}(-1)^m b_{r_1-1}\cdots b_{r_m-1}\\&\quad+\sum_{k=2}^{n-1}kb_{n-k}\sum_{\textcircled{2}}(-1)^m b_{r_1-1}\cdots b_{r_m-1}+b_{n-1}\end{aligned}\tag{17}$$

where sum ① (resp. ②) is over all m-tuples $(r_1,\cdots,r_m)$ of integers $\geqq 2$ with $r_1+r_2+\cdots+r_m=n$ (resp. $k-1$). We rearrange the first sum in (17) to group together all of those m-tuples whose last entry is $n-k$ $(0\leqq k\leqq n-2)$. Noting that when $k=2$, $\sum_{\textcircled{2}}$ is 0 we find that

$$\begin{aligned}F_n(b_0)&=(n+1)\sum_{k=2}^{n-2}b_{n-k-1}\left(\sum(-1)^{m+1}b_{r_1-1}\cdots b_{r_m-1}\right)\\&\quad-\sum_{k=2}^{n-2}(k+1)b_{n-k-1}\left(\sum(-1)^{m+1}b_{r_1-1}\cdots b_{r_m-1}\right)-nb_{n-1}\\&=\sum_{k=2}^{n-2}(n-k)b_{n-k-1}\left(\sum(-1)^{m+1}b_{r_1-1}\cdots b_{r_m-1}\right)-nb_{n-1}\end{aligned}\tag{18}$$

where the inner sums in (18) are over m-tuples $(r_1,\cdots,r_m)$ of integers $\geqq 2$ with $r_1+\cdots+r_m=k$. With a change of variable in the first sum $(k\leftrightarrow n-k)$ this is equivalent to (8). □

To conclude this section, we admit that using expressions (7) and (8) to compute a large number of coefficients can be tedious. In this event we should point out that (5) gives rise to an easily proven recursion relation.

LEMMA 3. *With the same notation as Theorem* 1, *and taking* $c_n^{(n)}=1$ $(n=0,1,2,\cdots)$ *and* $c_{n-1}^{(n)}=0$ $(n=1,2,\cdots)$ *we have that*

$$c_k^{(n+1)}=c_{k-1}^{(n)}-\sum_{l=k}^{n-1} b_{n-l}c_k^{(l)} \qquad (n\geqq 0,\ 1\leqq k\leqq n-1)$$

and

$$c_0^{(n+1)}=-\sum_{l=1}^{n-1} b_{n-l}c_0^{(l)}-(n+1)b_n \qquad (n\geqq 1).$$

3. Expansions of analytic functions. In this section we will derive an explicit formula for the Faber expansion of functions that are entire, or at least analytic on an appropriate disk containing E or Int (E). It is known [5, p. 42] that if $f(z)$ is analytic on E or if ∂E is analytic and $f(z)$ is analytic on Int E, then (3) and (4) hold. In the first case, the r in (4) will be >1 while in the second case, will be <1. In either case the series (3) converges uniformly on compact subsets of Int (E). Throughout this section we will assume $f(z)$ and/or E are as discussed above, though more general circumstances under which (3) and (4) hold are known [3].

THEOREM 4. *Let* $k\geqq 0$ *be an integer. Then*

$$(w-b_0)^k=F_k(w)+\sum_{l=0}^{k-1}\gamma_l^{(k)}F_l(w) \tag{19}$$

where

$$\gamma_l^{(k)}=\sum_{s=[(k+l+1)/2]}^{k-1}\binom{k}{s}\sum b_{r_1}\cdots b_{r_{k-s}} \qquad (0\leqq l\leqq k-1) \tag{20}$$

and the inner sum in (20) *is over all* $(k-s)$*-tuples* $(r_1,\cdots,r_{k-s})$ *of positive integers with* $r_1+\cdots+r_{k-s}=s-l$. *We take* $\gamma_k^{(k)}=1$ $(k=0,1,\cdots)$. *In* (20), [ ] *is the greatest integer function.*

Proof. It is clear that the coefficient of F_k is 1. Since $(w-b_0)^k$ is entire, we may use (4) to obtain

$$\begin{aligned}\gamma_l^{(k)}&=\frac{1}{2\pi i}\int_{|\xi|=r}\frac{(g(\xi)-b_0)^k}{\xi^{l+1}}\,d\xi \qquad (0\leqq l\leqq k-1)\\ &=\text{coefficient of } \xi^l \text{ in } (g(\xi)-b_0)^k\\ &=\sum b_{r_1}b_{r_2}\cdots b_{r_k}\end{aligned} \tag{21}$$

where the sum in (21) is over all ordered k-tuples $(r_1, r_2,\cdots,r_k)$ with $r_j\in\{-1\}\cup\mathbf{Z}^+$ $(1\leqq j\leqq k)$, $r_1+r_2+\cdots+r_k=-l$ and where $b_{-1}=1$. If we rearrange (21) to first sum over those r_j's that may be -1, we find that to achieve the sum $-l$, at least $[(k+l+1)/2]$ of the r_j's must be -1. Since $l\leqq k-1$ in (21) and the sum is over k-tuples, we see at most $k-1$ of the r_j's may be -1. Thus

$$\gamma_l^{(k)}=\sum_{s=[(k+l+1)/2]}^{k-1}\binom{k}{s}\sum b_{r_1}\cdots b_{r_{k-s}}$$

where the inner sum is as described in the statement of the theorem. □

We may note that in (20) the sum is empty when $l=k-1$. Thus $\gamma_{k-1}^{(k)}\equiv 0$ $(k=1, 2,\cdots)$. As was the case for the coefficients $c_k^{(n)}$ of Theorem 1, we can also generate the $\gamma_l^{(k)}$'s through a recursion relation.

LEMMA 5. *We have* $\gamma_k^{(k)}=1$ $(k=0, 1, \cdots)$ *and* $\gamma_{k-1}^{(k)}=0$ $(k=1, 2, \cdots)$. *For* $k \geqq 1$,

$$\gamma_l^{(k+1)}=\gamma_{l-1}^{(k)}+\sum_{j=l+1}^{k} \gamma_j^{(k)} b_{j-l} \qquad (1 \leqq l \leqq k-1) \tag{22}$$

and

$$\gamma_0^{(k+1)}=\sum_{j=0}^{k}(j+1)\gamma_j^{(k)} b_j. \tag{23}$$

Proof. By (5) and (19) we have, for $k \geqq 1$,

$$\sum_{l=0}^{k+1} \gamma_l^{(k+1)} F_l(w)=(w-b_0)^{k+1}=(w-b_0)(w-b_0)^k \tag{24}$$

$$=\sum_{j=0}^{k} \gamma_j^{(k)}(w-b_0)F_j(w)$$

$$=\sum_{j=0}^{k} \gamma_j^{(k)}\left[F_{j+1}(w)+\sum_{l=1}^{j-1} b_{j-l}F_l(w)+(j+1)b_j\right]$$

$$=\sum_{j=1}^{k+1} \gamma_{j-1}^{(k)} F_j+\sum_{l=1}^{k-1}\left\{\sum_{j=l+1}^{k} \gamma_j^{(k)} b_{j-l}\right\}F_l+\sum_{j=0}^{k}(j+1)\gamma_j^{(k)} b_j$$

$$=F_{k+1}+\sum_{l=1}^{k-1}\left\{\gamma_{l-1}^{(k)}+\sum_{j=l+1}^{k} \gamma_j^{(k)} b_{j-l}\right\}F_l(w) \tag{25}$$

$$+\sum_{j=0}^{k}(j+1)\gamma_j^{(k)} b_j F_0(w).$$

Equating coefficients of $F_l(w)$ in (24) and (25) we get (22) and (23). □

With Theorem 4 we easily obtain the Faber expansions of entire functions.

THEOREM 6. *Suppose that either*

(i) ∂E *is analytic and* $f(z)$ *is analytic on some disk* $\Delta(b_0, r) \supset \operatorname{Int} E$,

or

(ii) $f(z)$ *is analytic on some disk* $\Delta(b_0, r) \supset E$. *Then if*

$$f(z)=\sum_{j=0}^{\infty} a_j(z-b_0)^j \qquad (|z-b_0|<r),$$

then

$$f(z)=\sum_{l=0}^{\infty} \alpha_l F_l(w) \tag{26}$$

where, for $l \geqq 0$,

$$\alpha_l=\sum_{k=l}^{\infty} a_k \gamma_l^{(k)} \tag{27}$$

and the $\gamma_l^{(k)}$*'s are defined in Theorem* 4. *In case* (i) *series* (26) *converges uniformly on compact subsets of* Int E *while in case* (ii) *series* (26) *converges uniformly on* E.

Proof. Let $S_n(z)=\sum_{j=0}^n a_j(z-b_0)^j$. By Theorem 4,

$$S_n(z)=\sum_{j=0}^n a_j\left\{F_j(z)+\sum_{l=0}^{j-1}\gamma_l^{(j)}F_l(z)\right\}$$

$$=\sum_{l=0}^n\left\{a_l+\sum_{j=l+1}^{n-1}a_j\gamma_l^{(j)}\right\}F_l(z).$$

In case (i), $\{S_n(z)\}_1^\infty$ converges to $f(z)$ uniformly on compact subsets of Int E. In case (ii), $\{S_n(z)\}_1^\infty$ converges to $f(z)$ uniformly on some disk $\Delta(b_0, r')$ where $\Delta(b_0, r)\supset\Delta(b_0, r')\supset E$. In either case, the α_l's are given by (3) and (4) for appropriate p.

$$\alpha_l=\int_{|\xi|=p}\frac{f(g(\xi))}{\xi^{l+1}}\,d\xi$$

$$=\lim_{n\to\infty}\int_{|\xi|=p}\frac{S_n(g(\xi))}{\xi^{l+1}}\,d\xi$$

$$=\lim_{n\to\infty}\left\{a_l+\sum_{j=l+1}^{n-1}a_j\gamma_l^{(j)}\right\}$$

$$=\sum_{j=l}^{\infty}a_j\gamma_l^{(j)}.$$

The uniform convergence of the Faber series follows from [5, p. 42]. □

As an example, we consider the case in which $g(z)=z+b_0+e^{i\theta}/z$ (θ real) and $f(z)=e^z$. In this case g maps Δ onto the complement of the segment

$$\left\{b_0+2e^{i\theta/2}\cos\left(t-\frac{\theta}{2}\right): 0\leqq t\leqq 2\pi\right\}.$$

For this choice of g, the Faber polynomials are translations of the familiar Chebyshev polynomials. In fact, we have that

$$F_n(w)=\frac{1}{2^n}\{[(w-b_0)+\sqrt{(w-b_0)^2-4e^{i\theta}}]^n+[(w-b_0)-\sqrt{(w-b_0)^2-4e^{i\theta}}]^n\}.$$

We may apply Theorem 1 to write $F_n(w)$ in more conventional form. Since $b_1=e^{i\theta}$ and $b_k=0$ ($k\geqq 2$), (7) shows that for $1\leqq k\leqq n-2$

$$c_k^{(n)}=\begin{cases}0 & (n-k\text{ odd}),\\ \dfrac{n}{k}(-1)^{(n-k)/2}\dbinom{(n+k-2)/2}{k-1}e^{((n-k)/2)i\theta} & (n-k\text{ even}),\end{cases}$$

while (8) gives

$$c_0^{(n)}=\begin{cases}0 & (n\text{ odd}),\\ 2(-1)^{n/2}e^{in\theta/2} & (n\text{ even}).\end{cases}$$

Thus

$$F_n(w) = (w-b_0)^n + \frac{(-1)^{n/2} n e^{in\theta/2}}{2} \sum_{k=1}^{(n/2-1)} \frac{(-1)^k}{k} \binom{\frac{1}{2}n+k-1}{2k-1} e^{-ik\theta}(w-b_0)^{2k}$$

$$+2(-1)^{n/2}\, e^{in\theta/2} \quad (n \text{ even})$$

and

$$F_n(w) = (w-b_0)^n + (-1)^{(n+1)/2} n\, e^{((n+1)/2)i\theta}$$

$$\cdot \sum_{k=1}^{(n-1)/2} \frac{(-1)^k}{2k-1} \binom{(n-1)/2+k-1}{2k-2} e^{-ik\theta}(w-b_0)^{2k-1} \quad (n \text{ odd}).$$

If we apply Theorem 4 to this particular $g(z)$, we find that

$$(w-b_0)^k = F_k(w) + \sum_{l=0}^{k-1} \gamma_l^{(k)} F_l(w)$$

where

$$(28) \qquad \gamma_l^{(k)} = \begin{cases} 0 & (l+k \text{ odd}), \\ \dbinom{k}{(l+k)/2} e^{i((k-l)/2)\theta} & (l+k \text{ even}). \end{cases}$$

Thus we find that

$$(w-b_0)^k = F_k(w) + \sum_{l=0}^{k/2-1} \binom{k}{k/2+l} e^{i(k/2-l)\theta} F_{2l}(w) \quad (k \text{ even})$$

and

$$(w-b_0)^k = F_k(w) + \sum_{l=0}^{(k-3)/2} \binom{k}{(k+1)/2+l} e^{(i((k-1)/2)-l)\theta} F_{2l+1}(w) \quad (k \text{ odd}),$$

giving expansions for $(w-b_0)^k$ in terms of Chebyshev polynomials. Finally, taking

$$f(z) = e^z = e^{b_0} \sum_{j=0}^{\infty} \frac{1}{j!}(z-b_0)^j,$$

Theorem 6 says, for $g(z) = z + b_0 + e^{i\theta}/z$,

$$e^z = \sum_{l=0}^{\infty} \alpha_l F_l(z)$$

where

$$\alpha_l = e^b \sum_{k=l}^{\infty} \frac{1}{k!} \gamma_l^{(k)}$$

and the $\gamma_l^{(k)}$'s are, in this case, given by (28). Hence, if l is even,

$$\begin{aligned}\alpha_l &= e^b \sum_{k=l/2}^{\infty} \frac{1}{(2k)!}\binom{2k}{k+l/2} e^{i(k-l/2)\theta} \\ &= e^b \sum_{j=0}^{\infty} \frac{e^{ij\theta}}{j!(j+l)!} \qquad (29)\\ &= e^{b-il\theta/2} J_l(2ie^{i\theta/2}).\end{aligned}$$

If l is odd

$$\begin{aligned}\alpha_l &= e^b \sum_{k=(l+1)/2}^{\infty} \frac{1}{(2k-1)!}\binom{2k-1}{k+(l-1)/2} e^{i(k-((l+1)/2))\theta} \\ &= e^b \sum_{j=0}^{\infty} \frac{e^{ij\theta}}{j!(j+l)!} \qquad (30)\\ &= e^{b-(il\theta/2)} J_l(2ie^{i\theta/2}).\end{aligned}$$

In (29) and (30) $J_l(t)$ is Bessel's function of order l of the first kind [1, p. 175]. Thus, in terms of the Chebyshev polynomials $\{F_n(w)\}$ generated by $z+b_0+e^{i\theta}/z$, we find that

$$e^z = e^b \sum_{l=0}^{\infty} e^{-il\theta/2} J_l(2ie^{i\theta/2}) F_l(z).$$

4. Rational functions and derivatives of Faber polynomials. Let $R(\mathbf{z})$ be a rational function. We may write

$$R(\mathbf{z}) = P(\mathbf{z}) + \sum_{j=1}^{n} \sum_{k=1}^{m_j} \frac{a_{j,k}}{(\mathbf{z}-\rho_j)^k}$$

where $P(\mathbf{z})$ is a polynomial ($P(\mathbf{z}) \equiv 0$ is possible) and $\rho_1, \rho_2, \cdots, \rho_n$ are the finite poles of R. We will assume that R has at least one finite pole and that for $1 \leqq j \leqq n$, R has a pole of order m_j at ρ_j. If none of the poles of R lies in E, then R is analytic on E and (3) and (4) hold with $f(\mathbf{z}) = R(\mathbf{z})$. Thus we may ask for the Faber expansion of R with respect to E.

Since the Faber expansion of a polynomial can be obtained using Theorems 4 and 6, it suffices to consider the Faber expansion for

$$f(\mathbf{z}) = \frac{1}{(\mathbf{z}-\rho)^k} \qquad (k=1,2,\cdots,\rho \in C\backslash E). \tag{31}$$

For future reference, we observe that if ρ is sufficiently large, then the techniques of § 3 can be used to obtain the expansion for (31). In fact, the following result is an easy application of Theorem 6.

THEOREM 7. *Let $\rho \in \mathbf{C}\backslash E$ with $E \subseteq \Delta(b_0, |\rho-b_0|)$. Then for $\mathbf{z} \in E$*

$$\frac{1}{(\mathbf{z}-\rho)^k} = \sum_{l=0}^{\infty} \beta_l F_l(\mathbf{z}) \tag{32}$$

where

$$\beta_l = \frac{1}{(b_0-\rho)^k} \sum_{r=l}^{\infty} \binom{k+r-1}{r} \frac{\gamma_l^{(r)}}{(\rho-b_0)^r} \tag{33}$$

and the $\gamma_l^{(r)}$'s are defined by (20).

Since (32) and (33) are valid whenever $E \subseteq \overline{\Delta(b_0, |\rho-b_0|)}$ and since $E \subseteq \overline{\Delta(b_0, 2)}$ [4, p. 19] it follows that (33) must converge whenever $|\rho-b_0|>2$. This leads to the following estimate on the growth of the $\gamma_l^{(r)}$'s.

COROLLARY 8. *For the coefficients $\gamma_l^{(r)}$ defined in* (20) *we have*

$$\limsup_{r\to\infty} |\gamma_l^{(r)}|^{1/r} \leqq 2. \tag{34}$$

This estimate is best possible in that the constant 2 *in* (34) *cannot be replaced by a smaller number.*

Proof. Applying the root test to (33), we see that for those ρ for which the series converges

$$1 \geqq \limsup_{r\to\infty}\left[\binom{k+r-1}{r}\frac{|\gamma_l^{(r)}|}{|\rho-b_0|^2}\right]^{1/2}$$

$$=\frac{1}{|\rho-b_0|}\limsup_{r\to\infty}|\gamma_l^{(r)}|^{1/r}.$$

Thus $|\rho-b_0|\geqq \limsup_{r\to\infty}|\gamma_l^{(r)}|^{1/r}$ for each ρ for which (33) converges. Since (33) converges for $|\rho-b_0|>2$, the desired inequality follows.

Taking $g(\mathbf{z})=\mathbf{z}+b_0+e^{i\theta}/\mathbf{z}$, the resulting $\gamma_l^{(r)}$'s given in (28) show the estimate is best possible. □

If $\rho\in\mathbf{C}\backslash E$ but $E\not\subseteq\Delta(b_0,|\rho-b_0|)$, then Theorem 7 does not apply. Nonetheless, we can still find the Faber expansion for (31) if we assume $g^{-1}(\rho)=\mathbf{z}_0\in\Delta$ is known. We will need the following lemma.

Lemma 9. *Let n, k be positive integers with* $k\leqq n$. *Then*

$$F_n^{(k)}(w)=k!\binom{n}{k}F_{n-k}(w)+k!\sum_{l=0}^{n-k-2}\eta_l^{(n,k)}F_l(w) \tag{35}$$

where

$$\eta_l^{(n,k)}=\left[\gamma_l^{(n-k)}+\binom{l+k}{k}c_{l+k}^{(n)}+\sum_{t=l}^{n-k-2}\binom{t+k}{k}c_{t+k}^{(n)}\gamma_l^{(t)}\right] \tag{36}$$

$(0\leqq l\leqq n-k-2)$ *and the c's and* γ*'s are defined in* (7), (8) *and* (20).

Proof. By Theorem 1

$$F_n(w)=(w-b_0)^n+\sum_{r=0}^{n-2}c_r^{(n)}(w-b_0)^r$$

where the $c_r^{(n)}$'s are given in (7) and (8). Thus

$$F_n^{(k)}(w)=k!\left\{\binom{n}{k}(w-b_0)^{n-k}+\sum_{t=0}^{n-k-2}c_{t+k}^{(n)}\binom{t+k}{k}(w-b_0)^t\right\}.$$

Where we apply Theorem 4 the last expression becomes

$$F_n^{(k)}(w)=k!\left\{\binom{n}{k}\left[F_{n-k}(w)+\sum_{l=0}^{n-k-1}\gamma_l^{(n-k)}F_l(w)\right]\right.$$
$$\left.+\sum_{t=0}^{n-k-2}c_{t+k}^{(n)}\binom{t+k}{k}\left[F_t(w)+\sum_{l=0}^{t-1}\gamma_l^{(t)}F_l(w)\right]\right\}.$$

Using $\gamma_{s-1}^{(s)}=c_{s-1}^{(s)}=0$ and interchanging the order of summation, we have

$$F_n^{(k)}(w)=k!\left\{\binom{n}{k}F_{n-k}(w)+\sum_{l=0}^{n-k-2}\left[\gamma_l^{(n-k)}+C_{l+k}\binom{l+k}{k}\right.\right.$$
$$\left.\left.+\sum_{t=l}^{n-k-2}C_{t-k}^{(n)}\gamma_l^{(t)}\binom{t+k}{k}\right]F_l(w)\right\}.$$

This completes the proof of the lemma. □

Theorem 10. *Let* $\mathbf{z}_0\in\Delta$ *and set* $\rho=g(\mathbf{z}_0)$. *Then for* $k=1, 2, \cdots,$

$$\frac{1}{(z-\rho)^k}=\sum_{l=0}^{\infty}\beta_lF_l(w) \tag{37}$$

where

$$\beta_l = (-1)^k k \sum_{n=l+k}^{\infty} \frac{\eta_l^{(n,k)}}{n\mathbf{z}_0^n} \tag{38}$$

and the $\eta_l^{(n,k)}$*'s are defined by* (36).

Proof. Let $A_{r_0} = \mathbf{C}\backslash\{g(\xi): |\xi| \geqq |z_0|\}$ $(r_0 = |z_0|)$. Then the series

$$\log\left(\frac{g(\mathbf{z}_0) - w}{\mathbf{z}_0}\right) = -\sum_{n=1}^{\infty} \frac{1}{n} F_n(w)\mathbf{z}_0^{-n}$$

is an analytic function of $w \in A_{r_0}$ and the series on the right converges uniformly (as a function of w) on compact subsets of A_{r_0}. Thus if we differentiate with respect to w, we find

$$\frac{1}{g(\mathbf{z}_0) - w} = \sum_{n=1}^{\infty} \frac{1}{n} F_n'(w)\mathbf{z}_0^{-n} \tag{39}$$

uniformly on compact subsets of A_{r_0}. Letting $g(\mathbf{z}_0) = \rho$ and differentiating (39) $k-1$ times with respect to w, we find

$$\begin{aligned} \frac{1}{(w-p)^k} &= \frac{(-1)^k}{(k-1)!} \frac{d^{(k-1)}}{dw^{(k-1)}}\left(\frac{1}{g(z) - w}\right) \\ &= \frac{(-1)^k}{(k-1)!} \sum_{n=k}^{\infty} \frac{1}{n} F_n^{(k)}(w)\mathbf{z}_0^{-n}. \end{aligned} \tag{40}$$

For $N \geqq k$, let

$$S_N(w) = \frac{(-1)^k}{(k-1)!} \sum_{n=k}^{N} \frac{1}{nz_0^n} F_n^{(k)}(w).$$

Then the sequence $\{S_N(w)\}_{N=k}^{\infty}$ converges to $1/(w-\rho)^k$ uniformly on compact subsets of A_{r_0}. The same is true if we rewrite $S_N(w)$ after substituting expression (35), with $\eta_{n-k}^{(n,k)} = \binom{n}{k}$ and $\eta_{n-k-1}^{(n,k)} = 0$:

$$\begin{aligned} S_N(w) &= (-1)^k k\left(\sum_{n=k}^{N} \frac{1}{n\mathbf{z}_0^n}\left(\sum_{l=0}^{n-k} \eta_l^{(n,k)} F_l(w)\right)\right) \\ &= (-1)^k k \sum_{l=0}^{N-k}\left(\sum_{n=l+k}^{N} \frac{\eta_l^{(n,k)}}{n\mathbf{z}_0^n}\right) F_l(w). \end{aligned}$$

If $r > 1$ is chosen so that $|z_0| > r$, then $1/(w-\rho)^k$ is analytic on $\mathbf{C}\backslash\{g(\xi): |\xi| \geqq r\}$. Thus (4) holds, that is, $f(w) = 1/(w-\rho)^k = \sum_{l=0}^{\infty} \beta_l F_l(w)$ with

$$\begin{aligned} \beta_l &= \frac{1}{2\pi i} \int_{|\xi|=r} \frac{f(g(\xi))}{\xi^{l+1}}\, d\xi \\ &= \lim_{N\to\infty} \frac{1}{2\pi i} \int_{|\xi|=r} \frac{S_N(g(\xi))}{\xi^{l+1}}\, d\xi \end{aligned}$$

$$= \lim_{N\to\infty} (-1)^k k \sum_{n=l+k}^{N} \frac{\eta_l(n,k)}{nz_0^n}$$

$$= (-1)^k k \sum_{n=l+k}^{\infty} \frac{\eta_l^{(n,k)}}{nz_0^n}. \qquad \square$$

5. Remarks and examples.

Remark. Unfortunately, the formula for β_l in Theorem 10 is rather complicated. However this seems unavoidable, even if we disregard the fact that working with coefficients of $(g(\xi)-w)^{-k}$ must lead to complications. In fact, we first note that the formula

$$\beta_l = (-1)^k k \sum_{n=l+k}^{\infty} \frac{\eta_l^{(n,k)}}{n\mathbf{z}_0^n} = (-1)^k k \sum_{n=l+k}^{\infty} \frac{\eta_l^{(n,k)}}{n(g^{-1}(\rho))^n} \tag{41}$$

must be valid for all $|\mathbf{z}_0|>1$ or $\rho \in \mathbf{C}\backslash E$. On the other hand, Theorem 7 shows that if $|\rho - b_0| > 2$, then

$$\beta_l = \frac{1}{(b_0-\rho)^k} \sum_{r=l}^{\infty} \binom{k+r-1}{r} \frac{\gamma_l(r)}{(\rho-b_0)^r}. \tag{42}$$

Since the Faber expansion coefficients for a function are unique, we have

$$\frac{1}{(b_0-\rho)^k} \sum_{r=l}^{\infty} \binom{k+r-1}{r} \frac{\gamma_l(r)}{(\rho-b_0)^r} = (-1)^k k \sum_{n=l+k}^{\infty} \frac{\eta_l^{(n,k)}}{n(g^{-1}(\rho))^n}$$

for $|\rho-b_0|>2$. Thus (41) is an analytic continuation of (42) to $\mathbf{C}\backslash E$. The fact that ∂E may be rather complicated means that this analytic continuation may be rather complicated also.

In spite of the complexity of the $\eta_l^{(n,k)}$'s we can, however, say that since (38) must converge for all $|\mathbf{z}_0|>1$,

$$\limsup_{n\to\infty} |\eta_l^{(n,k)}|^{1/n} \leqq 1. \tag{43}$$

Remark. Other methods for finding the $\eta_l^{(n,k)}$'s and the β_l's are possible, but also lead to complicated expressions. For example, we may first find the series expansion for $1/(g(z_0)-w)$ using Lemma 2, and then the coefficients for $(1/(g(\mathbf{z}_0)-w))^k$ as in the proof of Theorem 4. We chose the method presented in the proof of Lemma 9 and Theorem 10 because it gave access to and used expressions for the derivatives of the Faber polynomials. Such expressions can be valuable in working with Faber expansions.

Example. As an example of the use of formulas (35) and (36), again consider $g(\xi)=\xi+b_0+e^{i\theta}/\xi$. In § 3 we saw that for this choice of g we have that

$$e^z = e^b \sum_{l=0}^{\infty} e^{-il\theta/2} J_l(2ie^{i\theta/2}) F_l(z).$$

Equating e^z to its own derivative and substituting (35) with $k=1$ gives the following:

$$\sum_{l=0}^{\infty} e^{-il\theta/2} J_l(2ie^{i\theta/2}) F_l(z) \tag{44}$$

$$= \sum_{l=1}^{\infty} e^{-il\theta/2} J_l(2ie^{i\theta/2}) \left[lF_{l-1}(z) + \sum_{j=0}^{l-3} \eta_j^{(l,1)} F_j(w) \right]$$

$$(45)\qquad = \sum_{j=0}^{\infty} \left[\sum_{l=j+1}^{\infty} e^{-il\theta/2} \eta_j^{(l,1)} J_l(2ie^{i\theta/2}) \right] F_j(w),$$

where $\eta_j^{(j+1,1)} = j+1$ and $\eta_j^{(j+2,1)} = 0$. Equating coefficients of $F_j(w)$ in (44) and (45) we see that

$$(46)\qquad J_n(2ie^{i\theta/2}) = e^{in\theta/2} \sum_{l=n+1}^{\infty} e^{-il\theta/2} \eta_n^{(l,1)} J_l(2ie^{i\theta/2}).$$

The $\eta_n^{(l,1)}$'s are somewhat complicated even in this simple case. Nonetheless (46) exhibits an interesting relation between Bessel functions of different orders.

REFERENCES

[1] E. A. Coddington, *An Introduction to Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1961.

[2] J. Hagen, *On division of series*, Amer. J. Math., 5 (1883), pp. 236-237.

[3] T. Kovari and Ch. Pommerenke, *On Faber polynomials and Faber expansions*, Math. Z., 99 (1967), pp. 193-206.

[4] Ch. Pommerenke, *Univalent Functions*, VandenHoeck and Ruprecht, Göttingen, 1975.

[5] G. Schoeber, *Univalent Functions-Selected Topics*, Springer-Verlag, Berlin-New York, 1975.

[6] P. Todorov, *New explicit formulas for the coefficients of p-symmetric functions*, Proc. Amer. Math. Soc., 77 (1979), pp. 81-86.

[7] ———, *Explicit formulas for the coefficients of Faber polynomials with respect to univalent functions of the class* $\sum$, Proc. Amer. Math. Soc., 82 (1981), pp. 431-438.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

004

A TOPOLOGICAL APPROACH TO NEVANLINNA-PICK INTERPOLATION*

TRYPHON T. GEORGIOU†

Abstract. We study the set of rational solutions of an $(N+1)$-point Nevanlinna-Pick problem, that has degree bounded by N. Based on the invariance of the topological degree of a certain mapping under deformation, we establish that when the $(N+1)$-point Nevanlinna-Pick problem is solvable, then for any dissipation polynomial of degree N or less, there corresponds an interpolating function with dimension at most N. Our results provide a novel topological proof for the sufficiency of Pick's criterion for the solvability of the Nevanlinna-Pick problem, and also give a solution to an extended interpolation problem.

Key words. Nevanlinna-Pick interpolation, bounded degree solution, topological degree theory

AMS(MOS) subject classifications. Primary 46; secondary 30, 55, 93

1. Introduction. The Nevanlinna-Pick interpolation theory has a long history in mathematics. Its origin can be traced back to the beginning of the century in the work of Pick [20] and Nevanlinna [19] and it has reached a high degree of achievement in the recent work of Adamjan, Arov and Krein [1], Sarason [21], Sz.-Nagy and Foias [25] and Ball and Helton [3].

In engineering, it was in a circuit theoretical context where interpolation theory found the first applications (see Belevitch [4] and Wohlers [28]). In recent years, renewal of interest in the Nevanlinna-Pick interpolation problem has been motivated by a multitude of applications to system theoretic problems. These have been in the area of robust control, circuit theory, approximation theory, filtering, and stochastic processes (see Zames and Francis [30], Khargonekar and Tannenbaum [15], Helton [11], Genin and Kung [7], Dewilde, Vieira and Kailath [6] and Delsarte, Genin and Kamp [5]).

This paper addresses certain questions that carry a significant interest from an engineering standpoint.

It is known that whenever an $(N+1)$-point Nevanlinna-Pick problem is solvable, there exist rational solutions of degree at most N. Generically, the solution is nonunique. In this paper we present a study of the solutions of the $(N+1)$-point Nevanlinna-Pick problem that are at most of degree N. We show in Theorem 5.3 that for any dissipation polynomial (for a definition, see § 4) of degree at most N there exists a corresponding solution of degree at most N. This provides a description of the set of degree N solutions. We must point out that the degree of the interpolating function is related to the dimension of a controller in a feedback system, to the dimension of a modeling filter of a stochastic process, or to the McMillan degree of a certain transfer function in a circuit theoretic context. We show the above by exploiting the invariance of the topological degree of a certain mapping under deformation.

This approach also provides an independent topological proof of the sufficiency of Pick's criterion for the solvability of the Nevanlinna-Pick problem.

Our results are applied to tackle the solvability of an extended interpolation problem (see § 5) where, in addition to the $N+1$ interpolating conditions of the

* Received by the editors September 23, 1985; accepted for publication (in revised form) August 6, 1986. This research was supported in part by the National Science Foundation under grant ECS-8451519, and in part by grants from Honeywell, 3M, and the MEIS Center at the University of Minnesota, Minneapolis, Minnesota 55455.

† Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011.

standard problem, we require that the real part of the function satisfy extra N interpolating conditions on the boundary of the "stability" region. These N interpolating conditions are interpreted as attenuation zeros of an associated transmittance function.

This work follows the lines of an investigation on the Carathéodory problem [8], and a preliminary version was reported in [9].

A variety of different terminologies has appeared in connection with the Nevanlinna-Pick problem. For instance, the reflectance of a passive system is known also as a bounded real function or as a Schur function, etc. We have chosen to use a rather mathematical terminology as it appears in the classical references (e.g. Akhiezer [2]), although occasionally we indicate the "translation" of the various terms in the circuit theoretic or stochastic terminology.

2. Notation and terminology.

$\mathbb{C} = \{\text{complex numbers}\}$.

$\mathbb{R} = \{\text{real numbers}\}$.

$D =$ open unit disc
$= \{z \in \mathbb{C}: |z| < 1\}$.

X^c, X^0, ∂X indicate the closure, the interior and the boundary of a set X, respectively.

$H(D) = \{\text{functions holomorphic in } D\}$.

$C =$ class C (for Carathéodory)
$= \{f(z) \in H(D): \operatorname{Re}\{f(z)\} \geqq 0 \text{ for all } z \text{ in } D\}$.

$S =$ class S (for Schur)
$= \{f(z) \in H(D): |f(z)| \leqq 1 \text{ for all } z \text{ in } D\}$.

$z^* =$ complex conjugate of $z \in \mathbb{C}$.

If $a(z) = a_0 + a_1 z + \cdots + a_n z^n + \cdots \in H(D)$, then
$a(z)_* = a^*(z^{-1})$
$= a_0^* + a_1^* z^{-1} + \cdots + a_n^* z^{-n} + \cdots$ is analytic in $\mathbb{C} - D^c$.

L^2: the space of squarely integrable functions on ∂D.

H^2: the space of L^2-functions that have analytic continuation in D.

3. Nevanlinna-Pick interpolation. Consider two sets of $N+1$ points in $\mathbb{C}$,

$$z = \{z_\kappa: z_\kappa \in D \text{ for } \kappa = 0, 1, \cdots, N\} \quad \text{and} \quad w = \{w_\kappa: w_\kappa \in \mathbb{C} \text{ for } \kappa = 0, 1, \cdots, N\}.$$

For simplicity we will always assume that the points z_κ are all distinct. The Nevanlinna-Pick problem can be stated as follows.

PROBLEM NP (z, w). Construct, if possible, a function $f(z) \in C$ that satisfies the interpolation conditions

$$f(z_\kappa) = w_\kappa \quad \text{for } \kappa = 0, 1, \cdots, N. \tag{3.1}$$

In particular,

(NP$_1$) find necessary and sufficient conditions on the data (z, w) for the existence of a solution $f(z)$, and

(NP$_2$) give a complete description of *the set $C(z, w)$ of all C-functions satisfying* (3.1).

The solvability criterion was derived by Pick and a constructive algorithm was provided by Nevanlinna—we now outline these. For a more detailed exposition see Walsh [27].

Pick criterion. There exists a function $f(z) \in C$ that satisfies (3.1) if and only if the Pick matrix

$$P(z, w) := \left[\frac{w_k + w_l^*}{1 - z_\kappa z_l^*}\right]_{\kappa, l=0}^{N}$$

is nonnegative definite.

A similar interpolation problem can be stated in terms of functions of class S instead of C. Both formulations are equivalent. However, the Nevanlinna recursive scheme is simpler to describe in terms of S-functions. Define

$$\zeta_\kappa(z) := \frac{z - z_\kappa}{1 - z z_\kappa^*} \quad \text{for } \kappa = 0, 1, \cdots, N.$$

Nevanlinna recursive scheme. A function $f(z) \in C$ satisfies the interpolation conditions (3.1) if and only if

$$\text{(3.2a)} \qquad \operatorname{Re} w_0 \geqq 0$$

and

$$\text{(3.2b)} \qquad s_1(z) := \zeta_0(z)^{-1} \frac{f(z) - w_0}{f(z) + w_0^*}$$

belongs to the class S and satisfies the interpolation conditions

$$\text{(3.3)} \qquad s_1(z_\kappa) = v_{1,\kappa} := \zeta_0(z_\kappa)^{-1} \frac{f(z_\kappa) - w_0}{f(z_\kappa) + w_0^*} \quad \text{for } \kappa = 1, 2, \cdots, N.$$

Furthermore, a function $s_l(z) \in S$ such that

$$s_l(z_\kappa) = v_{l,\kappa} \quad \text{for } \kappa = l, l+1, \cdots, N$$

exists if and only if either

$$\text{(3.4a)} \qquad |v_{l,l}| < 1 \quad \text{and} \quad s_{l+1}(z) = \zeta_l(z)^{-1} \frac{s_l(z) - v_{l,l}}{1 - v_{l,l}^* s_l(z)} \text{ belongs to } S,$$

or,

$$\text{(3.4b)} \qquad |v_{l,l}| = 1 \quad \text{and} \quad s_l(z) = v_{l,l} = v_{l,l+1} = \cdots = v_{l,N}.$$

In case (3.4a) holds, $s_{l+1}(z)$ satisfies

$$s_{l+1}(z_\kappa) = v_{l+1,\kappa} := \zeta_\kappa(z_\kappa)^{-1} \frac{v_{l,\kappa}(z_\kappa) - v_{l,l}}{1 - v_{l,l}^* v_{l,\kappa}(z_\kappa)} \quad \text{for } \kappa = l+1, \cdots, N.$$

(For a proof see Walsh [27].)

Notice that at each step of this procedure the number of interpolation conditions is reduced. Thus, it leads to a recursive solution of NP (z, w). This is summarized below.

PROPOSITION 3.5. *The NP problem is solvable if and only if* $\operatorname{Re} w_0 > 0$ *and either*

$$\text{(3.5a)} \qquad |v_{\kappa,\kappa}| < 1 \quad \textit{for } \kappa = 1, \cdots, N$$

or

$$\text{(3.5b)} \qquad |v_{\kappa,\kappa}| < 1 \quad \textit{for } \kappa = 1, \cdots, m-1 \quad \textit{and} \quad |v_{m,m}| = 1, \quad v_{m,m} = \cdots = v_{m,N}.$$

In the later case the solution is unique, whereas in the former case the general solution is obtained using

$$f(z)=j\ \mathrm{Im}\ w_0+\mathrm{Re}\ w_0\frac{1+\zeta_0(z)s_1(z)}{1-\zeta_0(z)s_1(z)} \tag{3.6a}$$

and

$$s_l(z)=\frac{v_{l,l}+\zeta_l(z)s_{l+1}(z)}{1+v_{l,l}^*\zeta_l(z)s_{l+1}(z)} \quad \textit{for } l=n, n-1, \cdots, 1 \tag{3.6b}$$

from $v_{l,l}$, $l=1, \cdots, n$ *and an arbitrary* $s_{l+1}(z)\in S$.

We are interested in the "indeterminate" case when there is more than one solution. A necessary and sufficient condition for the problem to be indeterminate is (3.5a). This condition is equivalent to the positive definiteness of the associated Pick matrix. Hence, from now on, we will assume that *P is positive definite.*

In the indeterminate case, one particular solution of the NP problem is obtained by setting $s_{n+1}(z)=0$ in (3.6). We state some related facts in the following proposition.

PROPOSITION 3.7. *Let the Pick matrix P be positive definite and let* $f_0(z)$ *denote the solution of the NP problem that is obtained by setting* $s_{n+1}(z)\equiv 0$. *Then,* (a) $f_0(z)$ *is a rational function, and* (b) *if* $f_0(z)=\pi_0(z)/\chi_0(z)$ *with* $\pi_0(z)$ *and* $\chi_0(z)$ *coprime polynomials in z, then* $\max\{\deg \pi_0(z), \deg \chi_0(z)\}\leqq n$ *and* $\chi_0(z)\neq 0$ *for all* $z\in D^c$.

Proof of Proposition 3.7. From (3.6) it is easy to see that $f_0(z)$ is a rational function of degree less than or equal to *n*. (The *degree of a rational function* $\pi_0(z)/\chi_0(z)$ is defined to be the maximum of $\{\deg \pi_0(z), \deg \chi_0(z)\}$ where $\pi_0(z)$, $\chi_0(z)$ are polynomials in *z*.) Also using (3.6), one can derive that

$$\pi_0(z)\chi_0(z)_*+\chi_0(z)\pi_0(z)_*=k\prod_{\kappa=0}^{n-1}(z-z_\kappa)(z^{-1}-z_\kappa^*)$$

for some scalar $k>0$. Now, since $|z_\kappa|<1$ for all κ, $\chi_0(z)$ (and for that matter $\pi_0(z)$ also) cannot have a root on ∂D, otherwise, $\chi_{0*}(z)$ would have a root at the same point. This cannot happen because the right-hand side of the above has no root on ∂D. Finally, that $\chi_0(z)$ has no root outside D^c is a consequence of the fact that $f_0(z)$ is a *C*-function (Proposition 3.5). Q.E.D.

4. Rational *C*-functions.

A well-known characterization of rational *C*-functions is given below (see Siljak [24]).

PROPOSITION 4.1. *Let* $\pi(z)$, $\chi(z)$ *be coprime polynomials in z. The rational function* $\pi(z)/\chi(z)$ *belongs to C if and only if*

$$\pi(z)+\chi(z)\neq 0 \quad \textit{for all } z\in D^c \tag{4.1a}$$

and

$$d(z, z^{-1}):=\pi(z)\chi(z)_*+\chi(z)\pi(z)_*\geqq 0 \quad \textit{for all } z\in\partial D. \tag{4.1b}$$

A polynomial $d(z, z^{-1})\in\mathbb{C}[z, z^{-1}]$ that satisfies (4.1b) will be called a *dissipation polynomial* (following Kalman [12]). The degree of the highest power of *z* will be called the *degree of* $d(z, z^{-1})$. (A necessary condition for $d(z, z^{-1})$ to be a dissipation polynomial is that $d(z, z^{-1})_*=d(z, z^{-1})$. Hence $\deg d(z, z^{-1})$ is also equal to the highest power of z^{-1}.)

Allowing the polynomials $\pi(z)$, $\chi(z)$ to have common factors, condition (4.1a) can be somewhat relaxed. The following modification of (4.1) will be utilized in the sequel.

PROPOSITION 4.2. *Let $\pi(z)$, $\chi(z)$ be polynomials in z (not necessarily coprime). If*

(4.2a) $$\pi(z)+\chi(z)\neq 0 \quad \textit{for all } z\in D \ (\textit{and not necessarily in } D^c)$$

and

(4.2b) $$d(z, z^{-1})\geqq 0 \quad \textit{for all } z\in \partial D,$$

then $\pi(z)/\chi(z)$ is a C-function.

Proof. From

$$|\pi(z)+\chi(z)|^2-|\pi(z)-\chi(z)|^2=2d(z, z^{-1})\geqq 0 \quad \text{for } z\in\partial D,$$

it follows that any root of $\pi(z)+\chi(z)$ on ∂D is also a root of $\pi(z)-\chi(z)$ and hence, of both $\pi(z)$ and $\chi(z)$. After extracting all common roots of $\pi(z)$ and $\chi(z)$ that lie on ∂D, we are left with a pair of polynomials $(\tilde{\pi}(z), \tilde{\chi}(z))$ that satisfy (4.1a), (4.1b) and also $\pi(z)/\chi(z)=\tilde{\pi}(z)/\tilde{\chi}(z)$. Therefore, $\pi(z)/\chi(z)$ is in C. Q.E.D.

5. Interpolation with rational *C*-functions of degree *N*: Main results.

PROBLEM 5.1: $I(z, w, \xi)$. Let (z, w) be a set of $N+1$ interpolating conditions, and let $\xi=\{\xi_\kappa: \xi_\kappa\in\partial D, \kappa=1, 2, \cdots, N\}$. Find necessary and sufficient conditions for the existence of a rational function $f(z)=\pi(z)/\chi(z)\in C$ that satisfies the interpolation conditions

(5.1a) $$f(z_\kappa)=w_\kappa \quad \text{for } \kappa=0, 1, \cdots, N,$$

and also

(5.1b) $$\pi(\xi_\kappa)\chi(\xi_\kappa)_*+\chi(\xi_\kappa)\pi(\xi_\kappa)_*=0 \quad \text{for } \kappa=1, \cdots, N.$$

Note that in case $\pi(z)$, $\chi(z)$ have no common factor, conditions (5.1b) can be written as

(5.1c) $$\operatorname{Re} f(\xi_\kappa)=0 \quad \text{for } \kappa=1, \cdots, N$$

(which represent Löwner-type interpolation conditions). In a circuit theoretic context the points $\xi_\kappa\in\partial D$ correspond to attenuation zeros for the corresponding Schur-bounded real transmittance function $s(z)$. That is, if $s(z)=1/z[f(z)-f(0)]/\ [f(z)+f(0)^*]$, then (5.1c) implies that $|s(\xi_\kappa)|=1$ and hence the attenuation $\log|s(\xi_\kappa)|=0$ for $\kappa=1, 2, \cdots, N$.

Although we have two sets of interpolation conditions it turns out that the solvability depends again on the positive definiteness of the Pick matrix.

THEOREM 5.2. *Problem $I(z, w, \xi)$ is solvable if and only if the Pick matrix associated with (z, w) is positive definite. Moreover, in this case, there always exists a solution of degree less than or equal to N.*

Theorem 5.2 is a direct corollary of the following more general one.

MAIN THEOREM 5.3. *Let (z, w) be a set of $N+1$ interpolating conditions such that the associated Pick matrix is positive definite, and also let $d(z, z^{-1})$ be an arbitrary dissipation polynomial of degree at most N. Then, there exists a pair of polynomials $(\pi(z), \chi(z))$ such that*

(5.3a) $$f(z)=\frac{\pi(z)}{\chi(z)}\in C \textit{ and satisfies } f(z_\kappa)=w_\kappa \textit{ for } \kappa=0, 1, \cdots, N,$$

(5.3b) $$\pi(z)\chi(z)_*+\chi(z)\pi(z)_*=k\,d(z, z^{-1}) \quad \textit{for some } k>0,$$

(5.3c) $$\deg f(z)\leqq N.$$

The proof of the above makes use of Topological Degree Theory (see § 6) and thus provides a novel approach to establish the sufficiency of Pick's criterion for the solvability of the NP problem which follows.

COROLLARY. *If the Pick matrix $P(z, w)$ is positive definite, then the* NP (z, w) *problem is solvable.*

Proof. This is a direct consequence of Theorem 5.3. Q.E.D.

Naturally, one would like to know whether NP (z, w) (or $I(z, w, \xi)$) has a solution of degree strictly less than N and for that matter, to determine the minimal degree (see Youla and Saito [29] and Kalman [13]). Unfortunately, this question seems to be tractable only by methods of decision theory. In fact, the problem of finding the minimal degree of a rational solution $f(z)$ when NP (z, w) is solvable, is a decidable one. The reason for that is that both the set of interpolation conditions and the conditions that guarantee that $f(z)$ belongs to C (see Proposition 4.1 and also Siljak [24]) can be phrased in terms of the solvability of a finite set of equations that depend polynomially on the coefficients of $f(z)$. For the existence of a solution the theory of Tarski [26] and Seidenberg [23] can be used. However, using the tools developed for the proof of Theorem 5.2 we obtain the following.

PROPOSITION 5.4. *The set of $N+1$-pairs (z, w) for which NP (z, w) is solvable but has no solution of degree strictly less than N, is open and nonempty for all N.*

Below we demonstrate the implications of our results to a particular case.

Example 5.5. Consider the problem NP (z, w) where $z=\{0, \frac{1}{2}\}$ and $w=\{1, 2\}$. The associated Pick matrix

$$P=\begin{pmatrix} 2 & 3 \\ 3 & 16/3 \end{pmatrix}$$

is positive definite. Consequently, the NP (z, w) is solvable. The general solution is given by

$$f(z)=\frac{1+zs_1(z)}{1-zs_1(z)}$$

where $s_1(z)$ is an S-function that satisfies

$$s_1\left(\frac{1}{2}\right)=-\frac{2}{3}$$

and a general expression for it is given by (3.6b).

Let us restrict our attention to $f(z)$ of degree 1 or less. A rational function $f(z)$ that meets the interpolation data (z, w) is given by

$$f(z)=\frac{1+\alpha z}{1+\beta z} \quad \text{where } \alpha=2(1+\beta). \tag{5.6}$$

In order for $f(z)$ to belong to C it is necessary and sufficient that

$$|\alpha+\beta|+|\alpha-\beta|\leqq 2. \tag{5.7}$$

This follows easily from Proposition 4.1.

Let us now consider solutions of $I(z, w, \xi)$ for various points $\xi=\exp\{j\theta\}\in\partial D$. It can easily be verified that for all $\xi\neq+1$ there exists a degree 1 function $f(z)\in C$ as above, such that

$$\operatorname{Re} f(\xi)=0. \tag{5.8}$$

For instance, if $\xi=-1$, then the required $f(z)$ is

$$f(z)=\frac{1+z}{1-\frac{1}{2}z}.$$

(For all other choices of $\xi, f(z)$ has complex coefficients.) It is straightforward to verify (5.6)-(5.8). However, for $\xi=+1$, the function $f(z)$ sought in Theorem 5.1 is

$$f(z)=\frac{1}{1-z}.$$

This satisfies (5.6), (5.7) and also (5.1b). But $f(z)$ has a pole at $\xi=+1$ and, in this case,

$$\lim_{z\to\xi}\operatorname{Re} f(z)=\tfrac{1}{2}\neq 0.$$

In the general case, similar situations occur with probability zero. In other words, with generic data (z, w, ξ), the solutions to $I(z, w, \xi)$ have no poles on ∂D and in this case (5.1c) is equivalent to (5.1b).

6. Proof of the main results. First we recall certain tools of the geometric-functional theoretic approach of Sarason [21] to the interpolation problem.

Let $B(z)$ denote the finite Blaschke product (all-pass function) with simple zeros at z_κ, $\kappa=0, 1, \cdots, N$; i.e.,

$$B(z)=\prod_{\kappa=0}^{N}\frac{z-z_\kappa}{1-z_\kappa^* z}\cdot\frac{|z_\kappa|}{z_\kappa}$$

(where $|z_\kappa|/z_\kappa$ is replaced by 1 when $z_\kappa=0$). Let K denote the subspace of H^2

$$K:=H^2\ominus B(z)H^2.$$

The orthogonal projection in L^2 with range K is denoted by $[\ ]_K$, whereas $\langle\,,\rangle$ denotes the inner product in L^2.

K is an $(N+1)$-dimensional vector space and a commonly used basis for K is

$$B=\left\{g_\kappa(z)=\frac{1}{1-z_\kappa^* z},\ \kappa=0, 1, \cdots, N\right\}.$$

Note that for all $q(z)$ in H^2, $\langle q(z), g_\kappa(z)\rangle=q(z_\kappa)$.

Any element $q(z)$ in K can be represented as the ratio

$$q(z)=\frac{\chi(z)}{r(z)},$$

where $\chi(z)$ is a polynomial of degree N and

$$r(z):=\prod_{\kappa=0}^{N}(1-z_\kappa^* z).$$

Let $C(z, w)$ *be the set of solutions of* NP (z, w). The Pick matrix is assumed to be positive definite. Consequently, by Proposition 3.7, there exists a solution $f(z)$ in $C(z, w)$ that also belongs to $H(D^c)$. Define the linear operator

$$T: K\to K: q(z)\to[f(z)q(z)]_K.$$

It turns out that T depends only on the interpolation data (z, w) and not on the particular solution $f(z)$ in $C(z, w)$. Moreover, as it will become clear below, T can be defined directly on the basis of (z, w) (and does not require the solvability of NP (z, w)).

LEMMA 6.1. *Let $p(z)=[f(z)q(z)]_K$, where $q(z)\in K$, and $f(z)\in C(z,w)\cap H(D^c)$ as above. Then $p(z)$ is independent of the particular $f(z)$, it depends only on (z,w) and it satisfies*

$$\frac{p(z_\kappa)}{q(z_\kappa)}=w_\kappa,\qquad \kappa=0,1,\cdots,N.$$

Proof. Let $q(z)=\sum_{\kappa=0}^N b_\kappa g_\kappa(z)$ and $p(z)=\sum_{\kappa=0}^N a_\kappa g_\kappa(z)$. Then

$$\begin{aligned}p(z_\kappa)&=\langle p(z),g_\kappa(z)\rangle=\langle [f(z)q(z)]_K,g_\kappa(z)\rangle\\&=w_\kappa q(z_\kappa).\end{aligned}$$

This is a set of $N+1$ equations in the $N+1$ coefficients of $p(z)$:

$$G\begin{bmatrix}a_0\\a_1\\\vdots\\a_N\end{bmatrix}=\begin{bmatrix}w_0q(z_0)\\w_1q(z_1)\\\vdots\\w_Nq(z_N)\end{bmatrix}=\begin{bmatrix}w_0&&&\\&w_1&&\\&&\ddots&\\&&&w_N\end{bmatrix}G\begin{bmatrix}b_0\\b_1\\\vdots\\b_N\end{bmatrix}$$

where G is the Gram matrix

$$G=\lfloor\langle g_k(z),g_l(z)\rangle\rfloor_{\kappa,l=D}^N=\left[\frac{1}{1-z_\kappa z_l^*}\right]_{\kappa,l=0}^N.$$

Since the $g_\kappa(z)$, $\kappa=0,1,\cdots,N$ are linearly independent, G is nonsingular. (This can also be shown directly by computing the determinant of G. G is related in a simple way to the so-called Hilbert matrix and a formula for the determinant of a Hilbert matrix can be found in Knuth [16, p. 36].) Thus, T is the linear transformation specified by $p(z)=Tq(z)$ where

$$\begin{bmatrix}a_0\\a_1\\\vdots\\a_N\end{bmatrix}=G^{-1}\begin{bmatrix}w_0&&&\\&w_1&&\\&&\ddots&\\&&&w_N\end{bmatrix}G\begin{bmatrix}b_0\\b_1\\\vdots\\b_N\end{bmatrix}.\tag{6.1a}$$

Q.E.D.

PROPOSITION 6.2 (Sarason [21]). *The Pick matrix P is the real part of T.*

Proof. Let $q(z)=\sum_{\kappa=0}^N b_\kappa g_\kappa(z)$. Then,

$$\begin{aligned}2\,\mathrm{Re}\,\langle[f(z)q(z)]_K,q(z)\rangle&=\langle[f(z)q(z)]_K,q(z)\rangle+\langle q(z),[f(z)q(z)]_K\rangle\\&=\sum_{\kappa,l=0}^N b_l\frac{w_\kappa+w_l}{1-z_\kappa z_l^*}b_\kappa.\end{aligned}\tag{6.1b}$$

Q.E.D.

Let $f(z)\in C(z,w)\cap H(D^c)$ as before. Define the following linear map:

$$\Psi:K\to K:q(z)\to u(z)=[(1+f(z))q(z)]_K.$$

Since $f(z)\in C$, then $1+f(z)$ has an inverse in $H(D^c)$ and Ψ is invertible. Define

$$\psi:=\Psi^{-1}:K\to K:u(z)\to q(z)=[(1+f(z))^{-1}q(z)]_K.$$

Note that ψ (and also Ψ) depends only on (z,w) and not on our choice of $f(z)\in C(z,w)$. This can be readily established as in Lemma 6.1 and, in point of fact, if $u(z)=\Sigma u_\kappa g_\kappa$, then $\psi[u(z)]=q(z)=\Sigma b_\kappa q_\kappa$ where

$$\begin{bmatrix}b_0\\b_1\\\vdots\\b_N\end{bmatrix}=G^{-1}\begin{bmatrix}1/(1+w_0)&&&\\&1/(1+w_1)&&\\&&\ddots&\\&&&1/(1+w_N)\end{bmatrix}G\begin{bmatrix}u_0\\u_1\\\vdots\\u_N\end{bmatrix}.\tag{6.2a}$$

We also note that if $p(z)=T[q(z)]$ then $u(z)=p(z)+q(z)$. Using these facts we proceed to the proof of Theorem 5.3.

Proof of Theorem 5.3. Given $u(z)\in K$ we can readily obtain a pair $(p(z), q(z))\in K\times K$, by $q(z)=\psi[u(z)]$ and $p(z)=T[q(z)]$ using (6.1a) and (6.2a), such that

$$\frac{p(z)}{q(z)} \text{ is a rational function of degree at most } n,$$

$$\frac{p(z_i)}{q(z_i)}=w_i \qquad \text{(Lemma 6.1)}, \tag{6.2b}$$

$$p(z)+q(z)=u(z).$$

Let $q(z)=\chi(z)/r(z)$ and $p(z)=\pi(z)/r(z)$ where $\pi(z)$, $\chi(z)$ are polynomials of degree n and $r(z)$ is as earlier. Consider the function

$$\begin{aligned} e(z) &:= p(z)q(z)_* + q(z)p(z)_* \\ &= \frac{\pi(z)\chi(z)_* + \chi(z)\pi(z)_*}{r(z)r(z)_*} \\ &= d(z, z^{-1})\rho(z) \quad \text{with } z\in\partial D \end{aligned}$$

where $\rho(z)=(r(z), r(z)_*)^{-1}$ ($e(z)$ can be considered as an L^1-function).

In order for $p(z)/q(z)=\pi(z)/\chi(z)$ to be a C-function it is sufficient (by Proposition 4.2) that

$$u(z)\neq 0 \quad \text{for all } z\in D^c \tag{6.3a}$$

and

$$d(z, z^{-1})\geqq 0 \quad \text{for all } z\in\partial D. \tag{6.3b}$$

We will establish the theorem by studying the correspondence

$$u(z)\to d(z, z^{-1}),$$

and showing that the image of $\{u(z)\colon u(z)\in K$, such that (6.3a) holds$\}$ contains the set of all polynomials $d(z, z^{-1})$ of degree at most n that satisfy (6.3b) (and are properly normalized). We now proceed to consider a normalization of $u(z)$ and $d(z, z^{-1})$ so that their correspondence becomes a continuous map between smooth manifolds.

Both $e(z)$ and $\rho(z)$ can be easily seen to be in L^2. Let e_k (resp., ρ_k) with $\kappa\in Z$ denote the Fourier coefficients of $e(z)$ (resp., $\rho(z)$). The polynomial $d(z, z^{-1})$ is of degree N (in both z and z^{-1}) and it holds that

$$e_0=\sum_{\kappa=-N}^{N} d_\kappa \rho_{-\kappa}.$$

On the other hand, using Proposition 6.2, we have that if $q(z)=\sum_{\kappa=0}^{N} b_\kappa g_\kappa(z)$ then

$$e_0=(b_0^* \cdots b_N^*)P\begin{bmatrix} b_0 \\ \vdots \\ b_N \end{bmatrix}.$$

Since P is positive definite, $e_0\neq 0$ (unless $q(z)\equiv 0$).

We now define the sets:

$$Y := \left\{ \tilde{d}(z, z^{-1}) = \sum_{\kappa=-N}^{N} \tilde{d}_\kappa z^\kappa \text{ such that } \sum_{k=-N}^{N} \tilde{d}_\kappa \rho_{-\kappa} = 1 \right\},$$

$$Y_+ := \{\tilde{d}(z, z^{-1}) \in Y \text{ such that } \tilde{d}(z, z^{-1}) \geqq 0 \text{ for all } z \in \partial D\},$$

$$X := \{u(z) \in K \text{ such that } u(0) = 1\},$$

$$X_+ := \{u(z) \in X \text{ such that } u(z) \neq 0 \text{ for all } z \in D^c\},$$

$$W := \{\text{rational functions } \pi(z)/\chi(z) \text{ of dimension at most } N \text{ that satisfy } \pi(z_k)/\chi(z_\kappa) = w_\kappa \text{ for } \kappa = 0, 1, \cdots, N\}.$$

Also we consider the following mappings:

$$\theta : K - \{0\} \to Y: q(z) \to \tilde{d}(z, z^{-1}) = \frac{1}{e_0} d(z, z^{-1}),$$

where e_0, $d(z, z^{-1})$ and $p(z)$ are computed from $q(z)$ as before, and

$$\omega := \theta \circ \psi|_X : X \to Y : u(z) \to \tilde{d}(z, z^{-1}).$$

Both mappings are completely specified by the interpolation data, and since $0 \in \chi(X)$, it is easy to see that ω is a continuous map. Now, X and Y are (smooth) linear manifolds of real dimension $2N$, and X_1 and Y_1 are compact subsets of X and Y respectively. On the other hand the mapping

$$X \to W : u(z) \to \frac{p(z)}{q(z)} = \frac{\pi(z)}{\chi(z)}$$

where $q(z) = \psi(u(z))$ and $p(z) = T(q(z))$, is clearly surjective. Hence, in order to establish the theorem we only need to show that

$$\omega(X_+^c) \supseteq Y_+.$$

To show this we will exploit the dependence of ω on the interpolation data w.

Consider z being fixed and define the set of w that render $P(z, w)$ positive definite:

$$B := \{w \in \mathbb{C}^{N+1} \text{ such that } P(z, w) \text{ is positive definite}\}.$$

We want to establish that B is a pathwise connected set. This follows immediately from the continuous dependence of w on the parameter $v_{\kappa\kappa}$, $\kappa = 1, 2, \cdots, N$ and the fact that the positive definiteness of P is equivalent to the conditions

$$\operatorname{Re} w_0 > 0 \quad \text{and} \quad |v_{\kappa\kappa}| < 1 \quad \text{for } \kappa = 1, 2, \cdots, N.$$

Now, provided the Pick matrix is positive definite (nonsingular would suffice), ω is a continuous map. Also, ω depends continuously on the parameters w. We shall indicate this by writing ω_w.

Since B is pathwise connected, we can construct a (continuous) homotopy H from $\omega_{w_{in}}$ to ω_w; by following a continuous path from an initial $w_{in} := \{w_\kappa = 1, \kappa = 0, 1, \cdots, N\}$ to any other point w in B, i.e.,

$$H : X \times [0, 1] \to Y,$$

such that $H(u(z), 0) = \omega_{w_{in}}(u(z))$ and $H(u(z), 1) = \omega_w(u(z))$.

We now proceed as follows: we first show that

$$H(X_+^c, t) \supseteq Y_+,$$

for $t=0$, and then that the same property holds for all $t\in[0,1]$; i.e., it remains invariant under the homotopy.

We first note that $f_0(z)\equiv 1\in C(z, w_{in})$. Then, the map $H(\cdot, 0)=\omega_{W_{in}}(\cdot)$ assumes a simple form where

$$q(z)=\tfrac{1}{2}u(z)=p(z).$$

If $q(z)=\chi(z)/r(z)$ as before,

$$\omega_{w_{in}}: X\to Y: \frac{2\chi(z)}{r(z)}\to k\chi(z)\chi(z)_*,$$

where k is a positive scalar making $k\chi(z)\chi(z)_*$ an element of Y. By the Riesz–Fejer Theorem ([10, p. 21]) any element of Y_+ assumes a unique representation

$$\tilde{d}(z, z^{-1})=k\chi(z)\chi(z)_*,$$

with $\chi(z)$ a polynomial in z, devoid of zeros in D.

From the above it readily follows that

$$\omega_{w_{in}}(X_+^c)=Y_+.$$

Moreover, the correspondence

$$\omega_{w_{in}}|_{X_+^c}: X_+^c\to Y_c \quad \text{is bijective.} \tag{6.4}$$

Now let $d(X_+, \omega, \tilde{d})$ denote the *topological degree of the map ω at $\tilde{d}$ relative to the set X_+*. The topological degree is a "measure" of the number of preimages in X_+ of the point $\tilde{d}$ under the mapping ω. In particular (6.4) implies that

$$d(X_+, H(\cdot, 0), \tilde{d})=1 \quad \text{for all } \tilde{d}\in Y_+^0, \tag{6.5}$$

where Y_+^0 indicates the interior of Y_+. For a comprehensive exposition of various aspects of degree theory see Lloyd [17] and Milnor [18].

We now show that

$$d(X_+, H(\cdot, t), \tilde{d})=1 \quad \text{for all } \tilde{d}\in Y_+^0 \text{ and } t\in[0,1]. \tag{6.6}$$

This follows from a very powerful theorem on the invariance of the degree under homotopy (Lloyd [17, p. 23]) after we prove that the image of the boundary of X_+ never intersects the interior of Y_+:

$$H(\partial X_+, t)\cap Y_+^0=\varnothing \quad \text{for all } t\in[0,1] \tag{6.7}$$

(see also Lloyd [17, p. 32], Milnor [18] and Schwartz [22] for the case of continuous deformations of maps between smooth manifolds).

We now prove (6.7). Assume that the above intersection was not empty and let $\tilde{d}(z, z^{-1})=H(u(z), t)\in Y_+^0$, where $u(z)\in\partial X_+$, and $t\in[0,1]$. Hence, $u(a)=p(a)+q(a)=0$ for some value $z=a\in\partial D$. Also

$$\begin{aligned}\tilde{d}(z, z^{-1}) &= k[p(z)q(z)_* + q(z)p(z)_*]\\ &= \frac{k}{2}[|p(z)+q(z)|^2-|p(z)-q(z)|^2]\geqq 0,\end{aligned}$$

for all $z\in\partial D$, while k is a positive scalar. Hence,

$$d(a, a^{-1})=0,$$

and $\tilde{d}(z, z^{-1})$ is not in the interior of Y_+. This is a contradiction.

Consequently, (6.7) is valid. Then, (6.6) follows from (6.5) and (6.7). Finally, (6.6) implies that

$$Y_+^0 \subseteq H(X_+, 1) = \omega_w(X_+),$$

which in turn, due to the compactness of X_+^c and the continuity of $\omega_w(\cdot)$, implies that

$$Y_+ \subseteq \omega_w(X_+^c).$$

This establishes the theorem. Q.E.D.

Proof of Theorem 5.2. From the Pick criterion it follows that $P(z, w)$ being nonnegative definite is a necessary condition for $I(z, w, \xi)$ to be solvable. To show the sufficiency part, let

$$d(z, z^{-1}) = \prod_{\kappa=1}^{N} (z-\xi_\kappa)(z^{-1}-\xi_\kappa^*)$$

and apply Theorem 5.3. Q.E.D.

Proof of Proposition 5.4. Consider the interpolation data (z, w) where

$$w_0 = 1 + a, \quad a \in \mathbb{C} \quad \text{and} \quad w_\kappa = 1 \quad \text{for } \kappa = 1, 2, \cdots, N.$$

The Pick matrix depends continuously on the parameters w and, consequently, it also depends continuously on the parameter a. For $a = 0$ the Pick matrix is positive definite. Hence, for $a \neq 0$ but is sufficiently small, the Pick matrix is still positive definite and NP (z, w) admits a solution. However, there is no rational function of degree *strictly* less than N that interpolates (z, w) unless $a = 0$. To see this, assume that such a function $f(z)$ exists, and let $f(z) = \pi(z)/\chi(z)$, where $\pi(z)$, $\chi(z)$ are polynomials of degree less than N. Then,

$$f(z) - 1 = \frac{\pi(z) - \chi(z)}{\chi(z)} = 0$$

for N different values of z. Therefore, $\pi(z) - \chi(z)$, being of degree at most $N-1$, is identically zero. Hence, $f(z) \equiv 1$, which is a contradiction. This establishes the proposition. Q.E.D.

7. Concluding remarks. Theorems 5.2 and 5.3 provide existence-type results to an inherently nonlinear problem. However, the homotopy used in the derivation can be used to provide an algorithmic way to find the sought solutions. For a study of homotopy methods as they relate to deriving numerical algorithms see the work of Kellogg, Li and Yorke [14].

In this paper we have presented a description of the set of interpolating functions of degree N to the $(N+1)$-point NP problem. However, it is not known at the moment whether the correspondence in Theorem 5.3 represents in fact a parametrization of this set; i.e., whether the correspondence between interpolating functions of degree N and dissipation polynomials as in Theorem 5.3 is in fact bijective.

Finally, a simple criterion to determine whether there exists an $f(z)$ in $C(z, w)$ of degree strictly less than N is still lacking. Such a criterion seems necessary for a thorough understanding of minimal degree solutions to NP (z, w) as considered in Youla and Saito [29] and Kalman [13].

REFERENCES

[1] V. M. Adamjan, D. Z. Arov and M. G. Krein, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb. (1971), pp. 31-73.

[2] N. I. AKHIEZER, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner, 1965.

[3] J. A. BALL AND J. W. HELTON, *A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107-142.

[4] V. BELEVITCH, *Classical Network Theory*, Holden-Day, San Francisco, 1968.

[5] P. DELSARTE, Y. GENIN AND Y. KAMP, *On the role of the Nevanlinna-Pick problem in circuit and system theory*, Circuit Theory and Appl., 9 (1981), pp. 177-187.

[6] P. DEWILDE, A. VIEIRA AND T. KAILATH, *On a generalized Szego-Levinson realization algorithm for optimal predictors based on a network synthesis approach*, IEEE Trans. Circuits and Systems, CA-25 (1978), pp. 663-675.

[7] Y. GENIN AND S. Y. KUNG, *A two-variable approach to the model reduction problem with Hankel norm criterion*, IEEE Trans. Circuits and Systems, CAS-28 (1981), pp. 912-924.

[8] T. GEORGIOU, *Realization of power spectra from partial covariance sequences*, submitted for publication, and *Topological aspects of the caratheodory problem*, Proc. Internat. Conf. on ASSP, IEEE, San Diego, CA, 1984.

[9] ———, *A topological view of the Nevanlinna-Pick problem*, Proc. Internat. Conf. on Decision and Control, IEEE, Las Vegas, New York, 1984.

[10] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Los Angeles, CA, 1958.

[11] J. W. HELTON, *Noneuclidean functional analysis and electronics*, Bull. Amer. Math. Soc., 1 (1982), pp. 1-64.

[12] R. E. KALMAN, unpublished report.

[13] R. E. KALMAN, *Realization of covariance sequences*, in Toeplitz Centennial, I. C. Gohberg, ed., Operator Theory: Advances and Applications, 4, 1981, pp. 331-342.

[14] R. B. KELLOGG, T. Y. LI AND J. YORKE, *A constructive proof of the Brouwer fixed-point theorem and computational results*, SIAM J. Numer. Anal., 13 (1976), pp. 473-483.

[15] P. KHARGONEKAR AND A. TANNENBAUM, *Noneuclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-30(10) (1985), pp. 1005-1013.

[16] D. E. KNUTH, *The Art of Computer Programming—Vol. 1, Fundamental Algorithms*, Addison-Wesley, Reading, MA, 1969.

[17] N. G. LLOYD, *Degree Theory*, Cambridge University Press, Cambridge, 1978.

[18] J. W. MILNOR, *Topology from the Differentiable Viewpoint*, The University Press of Virginia, Charlottesville, VA, 1965.

[19] R. NEVANLINNA, *Über beschränkte Functionen, die in gegebenen Punkten vorgeschriebene Werte annehmen*, Ann. Acad. Sci. Fenn. Ser. A, no. 1, 13 (1919).

[20] G. PICK, *Über die Beschränkungen analytischer Functionen, welchedurch vorgegebenen Functionwerte bewirkt sind*, Math. Ann., 77 (1916), pp. 7-23.

[21] D. SARASON, *Generalized interpolation in H^∞*, Amer. Math. Soc. Trans., 127 (1967), pp. 179-203.

[22] J. T. SCHWARTZ, *Nonlinear Functional Analysis*, Courant Institute of Mathematical Sciences, New York University, New York, 1965.

[23] A. SEIDENBERG, *A new decision method for elementary geometry*, Ann. Math., 60 (1954), pp. 365-374.

[24] D. D. SILJAK, *Algebraic criteria for positive realness relative to the unit circle*, J. Franklin Inst., 296 (1973), pp. 115-122.

[25] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[26] A. TARSKI, *A Decision Method for Elementary Algebra and Geometry*, University of California Press, Berkeley, CA, 1951.

[27] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ., Providence, RI, 20, 1956.

[28] R. WOHLERS, *Lumped and Distributed Passive Networks*, Academic Press, New York, 1969.

[29] D. C. YOULA AND M. SAITO, *Interpolation with positive-real functions*, J. Franklin Inst., 284 (1967), pp. 77-108.

[30] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity and optimal robustness*, IEEE Trans. Automat. Control, AC-28(5) (1983), pp. 585-600.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

005

THE CONSTRUCTION AND SMOOTHNESS OF INVARIANT MANIFOLDS BY THE DEFORMATION METHOD*

JERROLD MARSDEN† AND JÜRGEN SCHEURLE‡

Abstract. This paper proves optimal results for the invariant manifold theorems near a fixed point for a mapping (or a differential equation) by using the deformation, or Lie transform, method from singularity theory. The method was inspired by the difficulties encountered by the implicit function theorem technique in the case of the center manifold. The idea here is simply to deform the given system into its linearization and to track this deformation using the flow of a time-dependent vector field. Corresponding to the difficulties with the center manifold encountered by other techniques, we run into a "derivative loss" in this case as well, which is overcome by utilizing estimates on the differentiated equation. A survey of the other methods used in the literature is also presented.

Key words. invariant manifold, deformation method, center manifold, Lie transform

AMS(MOS) subject classification. 58F

1. Introduction. The theory of invariant manifolds is one of the fundamental ingredients in the study of dynamical systems. In this theory one looks for submanifolds of the phase space which are invariant under the flow, i.e., trajectories which start on such a manifold at some time, stay on it.

This problem is not only of interest from a qualitative point of view, but can lead to quantitative results. In fact, by restriction to an invariant manifold, an original system is reduced to a lower-dimensional one which might be relatively simple. In particular, this is the case when the phase space of the original system is infinite-dimensional and one considers finite-dimensional invariant submanifolds. An important example for applications is the center manifold which contains all bounded solutions near a fixed point [3], [15], [16].

The well-known invariant manifold theorems refer to the flow generated by a nonlinear vector field or diffeomorphism defined in a neighborhood of a fixed point [8], [10], [11], [14]. They give sufficient conditions for the existence of an invariant submanifold which contains this fixed point. For example, each component of the spectral decomposition of the phase space corresponding to a linear operator is an invariant subspace for the flow generated by this linear operator. In the general nonlinear theory one begins with such an invariant subspace of the linearized system and shows its persistence as an invariant submanifold for the full system (at least locally) and then one determines the smoothness of the resulting nonlinear manifold ([6]; cf. also [9]).

To construct such invariant manifolds, two different approaches have been used in the literature so far. First, the invariance property of the manifold has been used to derive an equation for a representing map [10], [11], [14]. The manifold is sought as a graph and an iteration scheme is used on the graphs. For diffeomorphisms, this "graph transform method" developed in [11] yields optimal results and even holds for "Lipeomorphisms" (also [18] and [20]). Second, asymptotic properties of the flow on the manifold have been used to derive an equation for the corresponding trajectories [2], [5], [7], [8], [12]. All these trajectories together span the desired manifold and

* Received by the editors May 9, 1986; accepted for publication (in revised form) August 14, 1986.

† Department of Mathematics, University of California, Berkeley, California 94720. The research of this author was partially supported by National Science Foundation contract DMS 84-04506.

‡ Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523, and Mathematisches Institut A der Universität Stuttgart, Pfaffenwaldring 57, D-7000 Stuttgart 80, West Germany.

invariance is a consequence. Alternatively, this second approach may be phrased as a fixed point problem for a map representing the manifold by considering the initial values of the trajectories parameterized over the invariant subspace of the linearized system [4], [19]. In both cases in the second approach, the resulting equation may be solved iteratively. For stable and unstable invariant manifolds the equation for the trajectories also has been solved using the classical implicit function theorem [12]. This yields optimal smoothness for C^k vector fields and diffeomorphisms, and even in the analytic case.

Unfortunately, it is not obvious how to apply the classical implicit function theorem for general invariant manifolds, e.g., for center manifolds. In general, the operator underlying the equation for the trajectories is not continuously differentiable in a space of functions which have the right asymptotic behavior (exponential growth). This difficulty always occurs for the equation of a representing map in a space of maps with a certain smoothness. Sacker [17] uses a smoothing technique to overcome this difficulty, but he still loses one order of smoothness for the solution. For unsuccessful attempts to apply the implicit function theorem in the case of center manifolds, see [4] and [13].

In the present paper we solve the equation for a representing map using a different approach, namely the "Lie transform" method of integrating a differential equation which is based on a deformation principle. This method has been used for the Darboux theorem, the Frobenius theorem and the Poincaré lemma [1] and is a common tool in singularity theory. The idea is to consider a one-parameter family of systems connecting the given system with its linearization. Differentiation with respect to the parameter yields a linear equation for a vector field which eventually has to be integrated in order to get the desired map. An initial condition is known from the invariant manifold of the linearized system.

We consider only diffeomorphisms here, although a similar approach for vector fields is possible. Our approach applies for general invariant manifolds; although we shall concentrate on the harder case of center manifolds, we indicate how results about other invariant manifolds can be obtained. Our smoothness results are optimal. We note at the outset that the diffeomorphisms which we are going to consider have to be of class C^3_{Lip} at least. This is the price we pay for our more sophisticated method.

The plan of the paper is as follows. In § 2 we state our main results. Theorem 2.1 is an existence and uniqueness result for a global center stable or center unstable invariant manifold of a C^3_{Lip} map in a Banach space. Corollary 2.2 contains the corresponding smoothness result for C^k ($k \geqq 4$) and C^k_{Lip} ($k \geqq 3$) maps. In Remark 2.3 we list certain modifications and generalizations of these results. Finally, §§ 3 and 4 contain the proofs of Theorem 2.1 and Corollary 2.2.

2. Formulation of the problem and results.

Let X and Y be Banach spaces. The product space is denoted by $X \times Y$ and equipped with the sup-norm. The Banach space of k-linear continuous maps from X to Y equipped with the usual norm induced by the norms of X and Y is denoted by $\mathcal{L}^k(X, Y)$, and we let $\mathcal{L}^k(X, X) = \mathcal{L}^k(X)$.

Also we introduce the Banach space $C^k(V, Y)$ of k-times continuously differentiable maps f from an open subset $V \subset X$ into Y, equipped with the norm

$$\|f\|_k = \sup_{\substack{x \in V \\ 0 \leqq i \leqq k}} \|D^i f(x)\|$$

where $D^i f$ denotes the ith derivative of f. Similar to the above, we set $C^k(X, X) = C^k(X)$. The linear subspace of those elements of $C^k(V, Y)$ for which the kth derivative is Lipschitz continuous in V is denoted by $C^k_{\mathrm{Lip}}(V, Y)$. Furthermore, we introduce the notation $C^k_L(V, Y)$ for the closed subset of elements of $C^k(V, Y)$ which satisfy a Lipschitz condition in V with a particular Lipschitz constant L.

Open balls in Banach spaces are denoted by $B_r(\cdot)$, where r is the radius and the point stands for the center of the ball. The closure of a set V is written as $\mathrm{cl}\,(V)$.

Let us consider a map

$$T: X\times Y\to X\times Y, \qquad (x,y)\mapsto(\phi_1(x,y), \phi_2(x,y))$$

given by

$$\phi_1(x,y)=Ax+f(x,y), \qquad \phi_2(x,y)=By+g(x,y),$$

where $A\in\mathscr{L}(X)$, $B\in\mathscr{L}(Y)$, and f and g are (nonlinear) perturbations. We consider the following hypotheses, for $\delta\in\mathbb{R}^+$ and k an integer:

$(\mathrm{L}1)_k$ $\|B^{-1}\|\,\|A^j\|<1$ for $0\leqq j\leqq k$.

$(\mathrm{L}2)_k$ $\|B\|\,\|A^{-j}\|<1$ for $0\leqq j\leqq k$.

$(\mathrm{N}1)_\delta$ $f\in C^3_{\mathrm{Lip}}(X\times U, X)$, and $g\in C^3_{\mathrm{Lip}}(X\times U, Y)$ where U is some neighborhood of 0 in Y;
$\|f\|_1<\delta$ and $\|g\|_1<\delta$.

(N2) $f(0,0)=0, \qquad g(0,0)=0.$

(N3) $D_xg(0,0)=0.$

Note that $(0,0)$ is a fixed point of T when (N2) holds.

We shall prove the following theorem about a so-called center stable or center unstable manifold.

THEOREM 2.1. *Let the assumption* $(\mathrm{L}1)_4$ *or* $(\mathrm{L}2)_4$, *and* $(\mathrm{N}1)_\delta$ *hold, where* $\delta>0$ *is sufficiently small. Then there is a map* $h\in C^3_{\mathrm{Lip}}(X, Y)$ *with* $\|h\|_1=O(\delta)$ *as* $\delta\to0$, *such that the manifold*

$$M=\{(x,y)\in X\times Y \mid y=h(x)\}$$

is invariant under the iteration of the map T, *i.e.*, $(x,y)\in M$ *implies* $T(x,y)\in M$; *the map* h *is unique in* $C_L(X, Y)$, *where* $L=O(1/\delta)$ *as* $\delta\to0$.

If in addition (N2) (*resp.* (N2) *and* (N3)) *hold, then* $h(0)=0$ (*resp.* $h(0)=0$ *and* $D_xh(0)=0$).

COROLLARY 2.2. *Assume that f and g are of class* C^{k-1}_{Lip} (*resp.* C^k) *for some* $k\geqq4$. *Furthermore, let* $(\mathrm{L}1)_k$ *or* $(\mathrm{L}2)_k$, *and* $(\mathrm{N}1)_\delta$ *hold. Then h is of class* C^{k-1}_{Lip} (*resp.* C^k) *provided that* δ *is sufficiently small.* (*In general* δ *depends on k for given A and B.*)

In the following remark we state some generalizations and modifications of the above results, which are obvious from the proofs in the next sections.

Remarks 2.3. (i) If B decomposes into two parts B_1 and B_2 such that B_1 satisfies $(\mathrm{L}1)_k$ and B_2 satisfies $(\mathrm{L}2)_k$ for some $k\geqq4$, then the above assertions remain true. In this case M is called a center manifold.

(ii) If $\|B^{-1}\|<1$ and $\|A\|<1$ (resp. $\|B\|<1$ and $\|A^{-1}\|<1$), then M is called the stable (resp. unstable) invariant manifold. In this case M is a C^∞ manifold if f and g are of class C^∞. (Here δ does not depend on k.) Moreover, in this case M is even analytic for analytic maps f and g. For the stable manifold this follows by using spaces of complex analytic functions instead of C^1_{Lip} functions in the existence proof. The unstable manifold case is reduced to the stable one just by considering the map T^{-1} instead of T, provided that it exists.

(iii) If the assumptions are only fulfilled when x is restricted to some neighborhood of 0 in X, then one can use a cut-off function $\chi: X \to \mathbb{R}$ to extend f and g to the domain $X \times U$. This is a C^∞ function with the property $\chi(x) = 1$ for $\|x\| \leqq \frac{1}{2}$ and $\chi(x) = 0$ for $\|x\| \geqq 1$. Such a function always exists if X is finite-dimensional. The extensions are given by $\tilde{f}(x, y) = f(\chi(\mu x)x, y)$ and $\tilde{g}(x, y) = g(\chi(\mu x)x, y)$ with an appropriate constant $\mu > 0$. Applying our results for $\tilde{f}$ and $\tilde{g}$ then yields a local invariant manifold for the original map T by restricting h to the ball $\|x\| < \mu^{-1}/2$.

This cut-off procedure destroys uniqueness and analyticity for the local case. On the other hand, we do not need the cut-off procedure for the local theory when $\|A\| < 1$ (or $\|A^{-1}\| < 1$). In that case we can directly work with spaces of maps which are defined only in some ball around $x = 0$. This yields local results for general spaces X and, in particular, analyticity. Hence, local stable (unstable) invariant manifolds are analytic if f and g are analytic. Furthermore, under the additional hypothesis that f and g together with all partial derivatives of g with respect to x up through order $l-1$ vanish at $(0, 0)$, the local results still hold when $\|A\| < 1$ (or $\|A^{-1}\| < 1$) and the inequalities in $(\mathrm{L1})_k$ ($(\mathrm{L2})_k$) only hold for $l \leqq j \leqq k$ for some $l \geqq 1$. In this case one has to work with functions h and $H = D_x h$ which have the properties $\|h(x)\| \leqq C_1\|x\|^l$, $\|Dh(x)\| \leqq C_2\|x\|^{l-1}$, $\|H(x)\| \leqq C_3\|x\|^{l-1}$ and $\|DH(x)\| \leqq C_4\|x\|^{l-2}$ in some ball around $x = 0$ with certain constants C_j. It finally follows that $\|D^j h(x)\| \leqq C_j\|x\|^{l-j}$ for $0 \leqq j \leqq l-1$. Note that strong, stable (or unstable) invariant manifolds, where $l = 1$, and also certain weak stable (or unstable) invariant manifolds are included in this local theory.

(iv) To obtain a smoothness result for M with respect to a parameter $\lambda \in \Lambda$, where Λ is some Banach space, we can consider λ as a component of x by adding the trivial component $\lambda \mapsto \lambda$ to the original map T (cf., [15]).

(v) Theorem 2.1 and Corollary 2.2 remain true if, in the definition of T, the terms Ax and By are replaced by any maps $A(x): X \to X$ and $B(x)y: X \times Y \to Y$ which are as smooth as f and g and satisfy the following assumptions:

$$(\tilde{\mathrm{L}}1)_k \qquad \|B(x)^{-1}\| \, \|DA(x)^j\| < 1 \quad \text{for } 0 \leqq j \leqq k,$$

$$(\tilde{\mathrm{L}}2)_k \qquad \|B(x)\| \, \|DA^{-1}(x)^j\| < 1 \quad \text{for } 0 \leqq j \leqq k.$$

For example, this generalization is relevant when one deals with a suspension of a nonautonomous system in the extended phase space which is the product of the (discrete) time axis and the original phase space.

(vi) Finally, we remark that it suffices to require $\|D_x f\| < \delta$ instead of $\|Df\| < \delta$ to prove the above results.

3. Proof of Theorem 2.1. We begin with the existence part. First we outline the basic ideas of our proof in a more or less formal way. Afterwards we shall justify each step by means of a series of lemmas.

We consider the following one-parameter family of maps:

$$T_\varepsilon: X \times Y \to X \times Y, \qquad (x, y) \mapsto (\phi_1(\varepsilon, x, y), \phi_2(\varepsilon, x, y))$$

given by

$$(3.1) \qquad \phi_1(\varepsilon, x, y) = Ax + \varepsilon f(x, y), \qquad \phi_2(\varepsilon, x, y) = By + \varepsilon g(x, y),$$

for ε a real number. Obviously T_ε defines a homotopy between the linear map

$$T_0: (x, y) \mapsto (Ax, By) \quad \text{and} \quad T_1 = T.$$

For each T_ε we are looking for an invariant manifold M_ε of the form $y = h_\varepsilon(x)$, where the map $h_\varepsilon: X \to Y$ depends smoothly on ε. The invariance property leads to the equation

$$h_\varepsilon(\phi_1(\varepsilon, x, h_\varepsilon(x))) = \phi_2(\varepsilon, x, h_\varepsilon(x)). \tag{3.2}$$

Moreover, we require

$$h_0(x) = 0 \quad \text{for } x \in X, \tag{3.3}$$

since $y = 0$ is an invariant manifold of T_0. Thus, we aim to solve the system of equations (3.2) and (3.3) for $h_\varepsilon(x)$.

The main idea now is to derive a first order differential equation for the function $\varepsilon \mapsto h_\varepsilon$ and to integrate this in the interval $0 \leqq \varepsilon \leqq 1$ with (3.3) as an initial condition. Actually, we shall consider a differential system for the function $\varepsilon \mapsto (h_\varepsilon, H_\varepsilon)$, where

$$H_\varepsilon(x) = D_x h_\varepsilon(x) \in \mathscr{L}(X, Y). \tag{3.4}$$

Thus we get a linear equation for the corresponding vector field which can be solved explicitly. Subsequently the arguments of ϕ_1, ϕ_2 and all their derivatives are $(\varepsilon, \cdot, h_\varepsilon(\cdot))$, if not indicated otherwise. A dot above a symbol for a map denotes the partial derivative with respect to ε.

First we differentiate equation (3.2) with respect to ε, which yields

$$\dot{h}_\varepsilon(\phi_1) - \mathscr{B}\dot{h}_\varepsilon = \mathscr{F}_1, \tag{3.5}$$

where

$$\mathscr{B} = \mathscr{B}(\varepsilon, h_\varepsilon, H_\varepsilon) = D_y\phi_2 - H_\varepsilon(\phi_1)D_y\phi_1$$

is a map from X to $\mathscr{L}(Y)$ and

$$\mathscr{F}_1 = \mathscr{F}_1(\varepsilon, h_\varepsilon, H_\varepsilon) = \dot{\phi}_2 - H_\varepsilon(\phi_1)\dot{\phi}_1$$

is a map from X to Y. An equation for H_ε is obtained by differentiating (3.2) with respect to x. Thus we obtain

$$H_\varepsilon(\phi_1)\mathscr{A} - D_y\phi_2 H_\varepsilon = D_x\phi_2 \tag{3.6}$$

where

$$\mathscr{A} = \mathscr{A}(\varepsilon, h_\varepsilon, H_\varepsilon) = D_x\phi_1 + D_y\phi_1 H_\varepsilon$$

is a map from X to $\mathscr{L}(X, Y)$. Since this equation is still nonlinear, we again differentiate it with respect to ε. Setting

$$G_\varepsilon(x) = D_x H_\varepsilon(x) \in \mathscr{L}^2(X, Y), \tag{3.7}$$

we thus get

$$\dot{H}_\varepsilon(\phi_1)\mathscr{A} - \mathscr{B}\dot{H}_\varepsilon + \mathscr{C}\dot{h}_\varepsilon = \mathscr{F}_2 \tag{3.8}$$

where

$$\begin{aligned} \mathscr{C}\dot{h}_\varepsilon &= \mathscr{C}(\varepsilon, h_\varepsilon, H_\varepsilon, G_\varepsilon)\dot{h}_\varepsilon \\ &= G_\varepsilon(\phi_1)(\mathscr{A}, D_y\phi_1\dot{h}_\varepsilon) + H_\varepsilon(\phi_1)(D^2_{xy}\phi_1\dot{h}_\varepsilon + D^2_{yy}\phi_1(H_\varepsilon, \dot{h}_\varepsilon)) \\ &\quad - D^2_{xy}\phi_2(\cdot, \dot{h}_\varepsilon) - D^2_{yy}\phi_2(H_\varepsilon, \dot{h}_\varepsilon), \end{aligned}$$

$$\begin{aligned}\mathcal{F}_2 &= \mathcal{F}_2(\varepsilon, h_\varepsilon, H_\varepsilon, G_\varepsilon)\\ &= D_x\dot{\phi}_2 + D_y\dot{\phi}_2 H_\varepsilon - G_\varepsilon(\phi_1)(\mathcal{A}, \dot{\phi}_1) - H_\varepsilon(\phi_1)(D_x\dot{\phi}_1 + D_y\dot{\phi}_1 H_\varepsilon)\end{aligned}$$

are maps from X to $\mathcal{L}(X, Y)$. A linear equation for G_ε is obtained by differentiating (3.6) with respect to x and using (3.4) again

$$(3.9) \qquad G_\varepsilon(\phi_1)(\mathcal{A}, \mathcal{A}) - \mathcal{B}G_\varepsilon = \mathcal{F}_3$$

where

$$\begin{aligned}\mathcal{F}_3 = \mathcal{F}_3(\varepsilon, h_\varepsilon, H_\varepsilon) &= D^2_{xx}\phi_2 + 2D^2_{xy}\phi_2(\cdot, H_\varepsilon) + D^2_{yy}\phi_2(H_\varepsilon, H_\varepsilon)\\ &\quad - H_\varepsilon(\phi_1)(D^2_{xx}\phi_1 + 2D^2_{xy}\phi_1(\cdot, H_\varepsilon) + D^2_{yy}\phi_2(H_\varepsilon, H_\varepsilon))\end{aligned}$$

is a map from X to $\mathcal{L}^2(X, Y)$.

Now we proceed as follows. For each fixed real ε and for each fixed pair of maps $h_\varepsilon : X \to Y$ and $H_\varepsilon : X \to \mathcal{L}(X, Y)$, we solve (3.9) for G_ε. The solution is written in the form

$$(3.10) \qquad G_\varepsilon = \mathcal{G}(\varepsilon, h_\varepsilon, H_\varepsilon)$$

where $\mathcal{G}(\varepsilon, h_\varepsilon, H_\varepsilon)$ is a map from X to $\mathcal{L}^2(X, Y)$. Inserting this expression into (3.8), we obtain

$$(3.11) \quad \dot{H}_\varepsilon(\phi_1)\mathcal{A} - \mathcal{B}\dot{H}_\varepsilon + \mathcal{C}(\varepsilon, h_\varepsilon, H_\varepsilon, \mathcal{G}(\varepsilon, h_\varepsilon, H_\varepsilon))\dot{h}_\varepsilon = \mathcal{F}_2(\varepsilon, h_\varepsilon, H_\varepsilon, \mathcal{G}(\varepsilon, h_\varepsilon, H_\varepsilon)).$$

This relation together with (3.5) is linear equation for $(\dot{h}_\varepsilon, \dot{H}_\varepsilon)$, which we write as

$$(3.12) \qquad \begin{pmatrix}\dot{h}_\varepsilon\\ \dot{H}_\varepsilon\end{pmatrix} = \mathcal{H}(\varepsilon, h_\varepsilon, H_\varepsilon)$$

where the right-hand side is a map from X to $Y \times \mathcal{L}(X, Y)$. This is the desired differential equation.

By the derivation of this equation, every two times continuously differentiable function $h_\varepsilon(x)$ which satisfies (3.2), together with its partial derivative $H_\varepsilon(x) = D_x h_\varepsilon(x)$, is a solution. To show that vice versa a solution $(h_\varepsilon(x), H_\varepsilon(x))$ of (3.12) such that (3.3) and

$$(3.13) \qquad H_0(x) = 0 \quad \text{for } x \in X$$

are satisfied yields a solution of (3.2), we show that H_ε is actually the partial derivative of h_ε with respect to x, i.e., (3.4) is satisfied. Inserting (3.4) into (3.5) and integrating with respect to ε, we then get relation (3.2). Here we use the fact that by (3.3), $h_\varepsilon(x)$ solves (3.2) for $\varepsilon = 0$.

To prove (3.4), we differentiate (3.5) and (3.8) with respect to x, which gives

$$(3.14)_j \qquad \begin{aligned}&\dot{p}_j(\phi_1)\mathcal{A}(\varepsilon, h_\varepsilon, p_j) - \mathcal{B}\dot{p}_j = \mathcal{F}_4(\varepsilon, p_j, q_j),\\ &\dot{q}_j(\phi_1)(\mathcal{A}(\varepsilon, h_\varepsilon, p_j), \mathcal{A}) - \mathcal{B}\dot{q}_j + \mathcal{C}\dot{p}_j = \mathcal{F}_5(\varepsilon, p_j, q_j, q_{l(j)}),\end{aligned}$$

for $j = 1$, where

$$l(1) = 2, \quad p_1 = D_x h_\varepsilon, \quad q_1 = D_x H_\varepsilon, \quad q_2 = G_\varepsilon,$$

$$\begin{aligned}\mathcal{F}_4(\varepsilon, p_j, q_j) = {} & D_x\dot{\phi}_2 + D_y\dot{\phi}_2 p_j + D^2_{yx}\phi_2(\dot{h}_\varepsilon, \cdot) + D^2_{yy}\phi_2(p_j, \dot{h}_\varepsilon) \\ & - q_j(\phi_1)(\mathcal{A}(\varepsilon, h_\varepsilon, p_j), (\dot{\phi}_1 + D_y\phi_1\dot{h}_\varepsilon)) \\ & - H_\varepsilon(\phi_1)(D_x\dot{\phi}_1 + D_y\dot{\phi}_1 p_j + D^2_{yx}\phi_1(\dot{h}_\varepsilon, \cdot) + D^2_{yy}\phi_1(\dot{h}_\varepsilon, p_j))\end{aligned}$$

is a map from X to $\mathcal{L}(X, Y)$ and

$$\begin{aligned}\mathcal{F}_5(\varepsilon, p_j, q_j, q_{l(j)}) = {} & D^2_{xx}\dot{\phi}_2 + D^2_{xy}\dot{\phi}_2(\cdot, p_j + H_\varepsilon) + D^3_{xxy}\phi_2(\cdot, \cdot, \dot{h}_\varepsilon) \\ & + D^3_{xyy}\phi_2(\cdot, \dot{h}_\varepsilon, H_\varepsilon + p_j) + D^2_{yy}\dot{\phi}_2(p_j, H_\varepsilon) + D_y\dot{\phi}_2 q_j \\ & + D^3_{yyy}\phi_2(\dot{h}_\varepsilon, H_\varepsilon, p_j) + D^2_{yy}\phi_2(\dot{h}_\varepsilon, q_j) + D^2_{yx}\phi_2(\dot{H}_\varepsilon, \cdot) \\ & + D^2_{yy}\phi_2(\dot{H}_\varepsilon, p_j) \\ & - H_\varepsilon(\phi_1)(D^2_{xx}\phi_1 + D^2_{xy}\phi_1(\cdot, p_j + H_\varepsilon) \\ & + D^2_{yy}\phi_2(H_\varepsilon, p_j) + D_y\phi_1 q_j - D_x G_\varepsilon(\phi_1)(\mathcal{A}(\varepsilon, h_\varepsilon, p_j), \mathcal{A}, \dot{\phi}_1 \\ & \qquad\qquad + D_y\phi_1\dot{h}_\varepsilon) \\ & - G_\varepsilon(\phi_1)(\mathcal{A}, D_x\dot{\phi}_1 + D_y\dot{\phi}_1 p_j + D^2_{yx}\phi_1\dot{h}_\varepsilon + D^2_{yy}\phi_1(\dot{h}_\varepsilon, p_j)) \\ & - q_{l(j)}(\phi_1)(\dot{\phi}_1 + D_y\phi_1\dot{h}_\varepsilon, D^2_{xx}\phi_1 \\ & \qquad\qquad + D^2_{xy}\phi_1(\cdot, p_j + H_\varepsilon) + D^2_{yy}\phi(p_j, H_\varepsilon) + D_y\phi_1 q_j) \\ & - q_j(\phi_1)(\mathcal{A}(\varepsilon, h_\varepsilon, p_j), D_x\dot{\phi}_1 + D^2_{xy}\phi_1(\cdot, \dot{h}_\varepsilon) + D_y\dot{\phi}_1 H_\varepsilon \\ & \qquad\qquad + D^2_{yy}\phi_1(\dot{h}_\varepsilon, H_\varepsilon) + D_y\phi_1\dot{H}_\varepsilon) \\ & - H_\varepsilon(\phi_1)(D^2_{xx}\dot{\phi}_1 + D^2_{xy}\dot{\phi}_1(\cdot, p_j + H_\varepsilon) + D^3_{xxy}\phi_1(\cdot, \cdot, \dot{h}_\varepsilon) \\ & \quad + D^3_{xyy}\phi_1(\cdot, \dot{h}_\varepsilon, p_j + H_\varepsilon) + D^2_{yy}\dot{\phi}_1(p_j, H_\varepsilon) + D_y\dot{\phi}_1 q_j \\ & \quad + D^3_{yyy}\phi_1(p_j, \dot{h}_\varepsilon, H_\varepsilon) + D^2_{yy}\phi_1(\dot{h}_\varepsilon, q_j) + D^2_{yx}\phi_1(\dot{H}_\varepsilon, \cdot) \\ & \qquad\qquad + D^2_{yy}\phi_1(p_j, \dot{H}_\varepsilon))\end{aligned}$$

is a map from X to $\mathcal{L}^2(X, Y)$. Furthermore, taking relation (3.8) as it stands and differentiating (3.9) with respect to ε, we obtain the relations $(3.14)_2$, where

$$l(2) = 1 \quad \text{and} \quad p_2 = H_\varepsilon.$$

Note, that here we need the assumption that f and g are of class C^3.

We shall show that the subspace given by $p_1 = p_2$ and $q_1 = q_2$ is invariant under the flow defined by the system of equations $(3.14)_1$ and $(3.14)_2$ in (p_1, q_1, p_2, q_2)-space. Thus, the identities (3.4) and (3.7) follow, when they are satisfied for $\varepsilon = 0$. But this will be a consequence of the initial conditions (3.3) and (3.13).

To summarize, so far we have argued that the problem (3.2) is formally equivalent to an initial-value problem for the differential equation (3.12). Now we are going to justify this argument step by step and to solve the initial-value problem.

We introduce the following notation:

$$I = [-\varepsilon_0, \varepsilon_0],$$

$$\mathcal{D}(r, L, M) = \{(h, H) \in C^1_L(X, Y) \times C^1_M(X, \mathcal{L}(X, Y)) \mid \|h\|_1 \leqq r, \|H\|_0 \leqq r, \|H\|_1 \leqq L\}$$

where ε_0 is an arbitrary real number greater than one, and r, L, and M are positive constants which are specified later.

LEMMA 3.1. *Assume that the conditions of Theorem* 2.1 *are satisfied and let* $(h_\varepsilon, H_\varepsilon) = (h, H)$ *be any element of* $\mathcal{D}(r, L, M)$ *where* $\operatorname{cl}(B_r(0)) \subset U$. *Then, for any* ε *in* I, *the equation* (3.5) *has a unique solution* $h_\varepsilon = \mathcal{H}_1 = \mathcal{H}_1(\varepsilon, h, H)$ *with the following properties*:

(i) $\mathcal{H}_1 \in C^1_{K_1}(X, Y)$

where the constant K_1 can be chosen independently of L and M. Furthermore $\|\mathcal{H}_1\|_0 \leqq r_1$ *and* $\|\mathcal{H}_1\|_1 \leqq r_2$, *where*

$$r_1 \leqq K_2(\|g\|_0 + \|f\|_0), \qquad r_2 \leqq K_3(\|f\|_1 + \|g\|_1),$$

with some positive constants K_2 and K_3; K_2 does not depend on L and M. Moreover, $\mathcal{H}_1(0)=0$ ($\mathcal{H}_1(0)=0$ *and* $D_x\mathcal{H}_1(0)=0$), *provided that* (N2) *and* $h(0)=0$ ((N2), (N3), $h(0)=0$, $Dh(0)=0$, *and* $H(0)=0$) *hold.*

(ii) *The map* $(\varepsilon, h, H) \mapsto \mathcal{H}_1(\varepsilon, h, H): I \times \mathcal{D}(r, L, M) \to C^1(X, Y)$ *is continuous and satisfies a Lipschitz condition with respect to* (h, H) *with constant* K_4.

Proof. The unique solution of (3.5) is given by

$$\mathcal{H}_1 = -\sum_{j=0}^{\infty} \left(\prod_{i=0}^{j} \mathcal{B}^{-1}(\phi_1^i) \right) \mathcal{F}_1(\phi_1^j) \tag{3.15}$$

if (L1)$_4$ holds, and by

$$\mathcal{H}_1 = \sum_{j=0}^{\infty} \left(\prod_{i=1}^{j} \mathcal{B}(\phi_1^{-i}) \right) \mathcal{F}_1(\phi_1^{-j-1}) \tag{3.16}$$

if (L2)$_4$ holds. Here we use the estimates $\|\mathcal{A}(\varepsilon, h, Dh) - A\|_0 = O(\delta)$ and $\|\mathcal{B} - B\|_0 = O(\delta)$ as $\delta \to 0$. It follows that for sufficiently small $\delta > 0$ the map

$$\mathcal{B}(x): Y \to Y \quad (\text{where } x \in X) \quad (\phi_1(\varepsilon, \cdot, h(\cdot)): X \to X)$$

can be inverted and the estimate $\|\mathcal{B}^{-1} - B^{-1}\|_0 = O(\delta)$ ($\|D_x\phi_1(\varepsilon, \cdot, h(\cdot))^{-1} - A^{-1}\|_0 = O(\delta)$) holds. Hence,

$$\|\mathcal{B}^{-1}\|_0 \|D_x\phi_1(\varepsilon, \cdot, h(\cdot))\|_0^j < 1 \qquad (\|\mathcal{B}\|_0 \|D_x\phi_1(\varepsilon, \cdot, h(\cdot))^{-1}\|_0^j < 1)$$

for all $0 \leqq j \leqq 4$. A straightforward computation shows that the series in (3.15) ((3.16)) converges in $C^1(X, Y)$ and represents a solution of (3.5) for $\dot{h}_\varepsilon$. Uniqueness is easily seen by an a priori C^0 estimate.

The remaining properties of $\mathcal{H}_1$ which are stated in the lemma, are easily seen by inspection of the formulas in (3.15) and (3.16). We simply note that $\|\mathcal{F}_1\|_0 \leqq \tilde{K}_2(\|g\|_0 + r\|f\|_0)$ and $\|D_x\mathcal{F}_1\|_0 \leqq \tilde{K}_3(\|f\|_1 + \|g\|_1)$ holds with some constants $\tilde{K}_2$ and $\tilde{K}_3$, where $\tilde{K}_2$ does not depend on L and M. Moreover, $\mathcal{F}_1(0)=0$ (resp. $\mathcal{F}_1(0)=0$ and $D_x\mathcal{F}_1(0)=0$) provided that (N2) and $h(0)=0$ (resp. (N2), (N3), $h(0)=0$, $Dh(0)=0$, and $H(0)=0$) holds. We also remark that Lipschitz constants can be estimated by the sup-norm of derivatives. Since the Lipschitz constant of $D_x\mathcal{F}_1$ is close to

$$\|D_{xx}^2 g\|_0 + 2\|D_{xy}^2 g\|_0 r + \|D_{yy}^2 g\|_0 r^2 + \|D_{xx}^2 f\|_0 r^2 + 2\|D_{xy}^2 f\|_0 r^2 + \|D_{yy}^2 f\|_0 r^3$$

where δ is sufficiently small, K_1 can be chosen independently of L and M.

To prove continuity of the map in (ii), one uses the fact that each member of the series in (3.15) (resp. (3.16)) has this property and that the convergence is uniform with respect to $(\varepsilon, h, H) \in I \times \mathcal{D}(r, L, M)$. □

LEMMA 3.2. *Suppose that the assumptions of Lemma* 3.1 *are valid. Then* (3.9) *has a unique solution* $G_\varepsilon = \mathcal{G} = \mathcal{G}(\varepsilon, h, H)$ *with the following properties*:

(i) $\mathcal{G} \in C_{K_5}^1(X, \mathcal{L}^2(X, Y))$,

$\|\mathcal{G}\|_0 \leqq K_6$, *and* $\|\mathcal{G}\|_1 \leqq K_7$, *where* K_5, K_6 *and* K_7 *are certain constants*; K_6 *does not depend on L and M. Furthermore,* $\mathcal{G}(0, 0, 0) = 0$.

(ii) *The map* $(\varepsilon, h, H) \mapsto \mathcal{G}(\varepsilon, h, H): I \times \mathcal{D}(r, L, M) \to C^1(X, \mathcal{L}^2(X, Y))$ *is continuous and satisfies a Lipschitz condition with respect to* (h, H) *with constant* K_8. *Moreover, with* $C^0(X, \mathcal{L}^2(X, Y))$ *as range, it is continuously differentiable.*

Proof. As in the previous proof, the unique solution of (3.9) is given by

$$\mathcal{G} = -\sum_{j=0}^{\infty}\left(\prod_{i=0}^{j}\mathcal{B}^{-1}(\phi_1^i)\right)\mathcal{F}_3(\phi_1^j)\left(\prod_{i=j-1}^{0}\mathcal{A}(\phi_1^i),\ \prod_{i=j-1}^{0}\mathcal{A}(\phi_1^i)\right) \tag{3.17}$$

if (L1)$_4$ holds, and by

$$\mathcal{G} = \sum_{j=0}^{\infty}\left(\prod_{i=1}^{j}\mathcal{B}(\phi_1^{-1})\right)\mathcal{F}_3(\phi_1^{-j-1})\left(\prod_{i=j+1}^{1}\mathcal{A}^{-1}(\phi_1^{-i}),\ \prod_{i=j+1}^{1}\mathcal{A}^{-1}(\phi_1^{-i})\right) \tag{3.18}$$

if (L2)$_4$ holds. Again, all properties of $\mathcal{G}$ which are stated easily follow from these formulae. Note that $\mathcal{F}_3(0,0,0)=0$. To see that K_6 does not depend on L and M we note that

$$\begin{aligned}\|\mathcal{F}_3\|_0 \leqq |\varepsilon|(&\|D_{xx}^2 g\|_0 + 2\|D_{xy}^2 g\|_0 r + \|D_{yy}^2 g\|_0 r^2\\ &+ \|D_{xx}^2 f\|_0 r + 2\|D_{xy}^2 f\|_0 r^2 + \|D_{yy}^2 f\|_0 r^3). \qquad \square\end{aligned}$$

LEMMA 3.3. *Suppose that the assumptions of the above lemmas hold. Then* (3.11), *with* $h_\varepsilon = \mathcal{H}_1$ *from Lemma* 3.1 *and* $\mathcal{G}$ *from Lemma* 3.2, *has a unique solution* $H_\varepsilon = \mathcal{H}_2 = \mathcal{H}_2(\varepsilon, h, H)$ *which has the following properties*:

(i) $\mathcal{H}_2 \in C_{K_9}^1(X, \mathcal{L}(X, Y)), \quad \|\mathcal{H}_2\|_0 \leqq r_3, \quad \|\mathcal{H}_2\|_1 \leqq K_{10}$

where $r_3 \leqq K_{11}(\|f\|_1 + \|g\|_1)$ *and* K_9, K_{10}, K_{11} *are certain positive constants*; K_9 *can be chosen independently of* M *and* K_{10} *independently of* L *and* M. *Moreover,* $\mathcal{H}_2(0)=0$ *provided that* (N2), (N3), $h(0)=0$, *and* $H(0)=0$ *hold.*

(ii) *The map* $(\varepsilon, h, H) \mapsto \mathcal{H}_2(\varepsilon, h, H): I \times \mathcal{D}(r, L, M) \to C^1(X, \mathcal{L}(X, Y))$ *is continuous and satisfies a Lipschitz condition with respect to* (h, H) *with some constant* K_{12}.

Proof. Set

$$\mathcal{F}_6 = \mathcal{F}_6(\varepsilon, h, H) = \mathcal{F}_2(\varepsilon, h, H, \mathcal{G}) - \mathcal{C}(\varepsilon, h, H, \mathcal{G})\mathcal{H}_1.$$

Then the equation which is considered in Lemma 3.3 has a unique solution given by

$$\mathcal{H}_2 = \sum_{j=0}^{\infty}\left(\prod_{i=0}^{j}\mathcal{B}^{-1}(\phi^i)\mathcal{F}_6(\phi_1)\prod_{i=j-1}^{0}\mathcal{A}(\phi_1^i)\right) \tag{3.19}$$

if (L1)$_4$ holds, and by

$$\mathcal{H}_2 = \sum_{j=0}^{\infty}\left(\prod_{i=1}^{j}\mathcal{B}(\phi^{-i})\mathcal{F}_6(\phi^{-j-1})\prod_{i=j+1}^{1}\mathcal{A}^{-1}(\phi_1^{-i})\right) \tag{3.20}$$

if (L2)$_4$ holds. To prove its stated properties, we note that

$$\|\mathcal{F}_6\|_0 \leqq \tilde{K}_{11}(\|f\|_1 + \|g\|_1)$$

holds with some constant $\tilde{K}_{11}$. Furthermore,

$$\begin{aligned}\|D_x\mathcal{F}_6\|_0 \leqq &\|D_{xx}^2 g\|_0 + 2\|D_{xy}^2 g\|_0 r + \|D_{yy}^2 g\|_0 r\\ &+ \|D_{xx}^2 f\|_0 r + \|D_{xy}^2 f\|_0 r^2 + \|D_{yy}^2 f\|_0 + O(\delta),\end{aligned}$$

the Lipschitz constant of $D_x\mathcal{F}_6$ with respect to x is smaller than

$$\begin{aligned}&\|D_{xxx}^3 g\|_0 + 3r\|D_{xxy}^3 g\|_0 + 3r^2\|D_{xyy}^3 g\|_0 + r^3\|D_{yyy}^3 g\|_0\\ &+ r(3L + |\varepsilon|K_1)\|D_{yy}^2 g\|_0 + (3L + |\varepsilon|K_1)\|D_{xy}^2 g\|_0 + r\|D_{xxx}^3 f\|_0\\ &+ 3r^2\|D_{xxy}^3 f\|_0 + 3r^3\|D_{xyy}^3 f\|_0 + r^4\|D_{yyy}^3 f\|\end{aligned}$$

$$+r^2(4L+(L+K_6)\|\mathcal{A}(\varepsilon, h, H)\|_0+|\varepsilon|K_1)\|D^2_{yy}f\|_0$$
$$+r(5L+2(L+K_6)\|\mathcal{A}(\varepsilon, h, H)\|_0+|\varepsilon|K_1)\|D^2_{yx}f\|_0$$
$$+(L+(L+K_6)\|\mathcal{A}(\varepsilon, h, H)\|_0)\|D^2_{xx}f\|_0+O(\delta),$$
$$\|D_x\mathcal{A}\|_0\leqq|\varepsilon|(\|D^2_{xx}f\|_0+2\|D^2_{xy}f\|_0r+\|D^2_{yy}f\|_0r^2)+O(\delta)$$

and

$$\|D_x\mathcal{B}\|_0\leqq|\varepsilon|(\|D^2_{xy}g\|_0+\|D^2_{yy}g\|_0r+\|D^2_{xy}f\|_0r+\|D^2_{yy}f\|_0r^2)+O(\delta)\quad \text{as } \delta\to 0.$$

Therefore, K_{10} can be chosen independently of L and M, and K_9 independently of M. Moreover, $\mathcal{F}_6(0)=0$, provided that (N2), (N3), $h(0)=0$ and $H(0)=0$ holds. But this implies $\mathcal{H}_2(0)=0$. The rest of the proof is similar to the previous proofs. □

By Lemma 3.1 and Lemma 3.3, the right-hand side of the differential equation (3.12) is given by

$$\mathcal{H}(\varepsilon, h, H)=\begin{pmatrix}\mathcal{H}_1(\varepsilon, h, H)\\ \mathcal{H}_2(\varepsilon, h, H)\end{pmatrix}.$$

It is uniquely determined by the stated properties. Next we are going to solve this equation with initial values $h=0$ and $H=0$ at $\varepsilon=0$. To this end we select r, L and M such that

(I1) $$\operatorname{cl}(B_r(0))\subset U,\ L\geqq\varepsilon_0\max(K_1, K_{10})\quad\text{and}\quad M\geqq\varepsilon_0K_9(L).$$

We also assume that the conditions

(I2) $$r_i\leqq\frac{r}{\varepsilon_0}\qquad(i=1,2,3)$$

are valid. This is achieved by requiring δ to be sufficiently small since $r_i=O(\delta)$ as $\delta\to 0$. Note that r has been chosen independently of δ. On the other hand we point out that $\operatorname{cl}(B_{\varepsilon_0r_1}(0))\subset U$ has to be true, but not necessarily $\operatorname{cl}(B_r(0))\subset U$. Hence, under certain circumstances one can shrink U to make δ sufficiently small and work with functions h such that $h(x)\in U$ for all $x\in X$.

LEMMA 3.4. *Let the assumptions of Theorem* 2.1 *be true. Furthermore, suppose that the constants r, L, M and δ are chosen such that the conditions* (I1) *and* (I2) *are fulfilled. Then the differential equation* (3.12) *has a unique solution* $\varepsilon\mapsto(h_\varepsilon(x), H_\varepsilon(x))$ *in the interval I, which has the following properties*:

(i) $\varepsilon\mapsto h_\varepsilon\in C^1(I, C^1_L(X, Y))$, $\varepsilon\mapsto H_\varepsilon\in C^1(I, C^1(X, \mathcal{L}(X, Y)))$, *where* $(h_\varepsilon, H_\varepsilon)\in\mathcal{D}(r, L, M)$ *holds for all* ε *in I, and* $h_0(x)=0$, $H_0(x)=0$ *for all* $x\in X$. *Moreover,* $h_\varepsilon(0)=0$ $(h_\varepsilon(0)=0, D_xh_\varepsilon(0)=0$ *and* $H_\varepsilon(0)=0)$ *for all* ε *in I, provided that* (N2) ((N2) *and* (N3)) *holds.*

(ii) $D_xh_\varepsilon=H_\varepsilon$, $D_xH_\varepsilon=G_\varepsilon=\mathcal{G}(\varepsilon, h_\varepsilon, H_\varepsilon)$.

Proof. The proof of part (i) follows the lines of the proof of the usual Picard-Lindelöf theorem for ordinary differential equations. We look for a continuous solution $\varepsilon\mapsto(h_\varepsilon, H_\varepsilon)$ of the integral equation

(3.21) $$\begin{pmatrix}h_\varepsilon\\ H_\varepsilon\end{pmatrix}=\int_0^\varepsilon\mathcal{H}(\sigma, h_\sigma, H_\sigma)\,d\sigma\qquad(\varepsilon\in I).$$

This problem is equivalent to solving the initial-value problem (3.12), (3.3), and (3.13). In particular, a continuous solution of (3.21) is continuously differentiable. According to the previous lemmas, the right-hand side of (3.21) defines a contraction map $\mathscr{T}$ of the metric space

$$S=\{\varepsilon\mapsto(h_\varepsilon, H_\varepsilon)\in C^0(I, C^1_L(X, Y))$$
$$\times C^0(I, C^1_M(X, \mathscr{L}(X, Y)))|(h_\varepsilon, H_\varepsilon)\in\mathscr{D}(r, L, M)\text{ for all }\varepsilon\in I\}$$

into itself, where the metric is given by the norm defined by

$$\sup_{\varepsilon\in I}(e^{\gamma|\varepsilon|}\|h_\varepsilon\|_1, e^{\gamma|\varepsilon|}\|H_\varepsilon\|_1)$$

with

$$\gamma>\max(K_2, K_{12}).$$

Obviously, with this metric S is complete. Therefore $\mathscr{T}$ has a unique fixed point in S, which is the desired solution. Moreover, the set

$$S_0=\{\varepsilon\mapsto(h_\varepsilon, H_\varepsilon)\in S|h_\varepsilon(0)=0\text{ for all }\varepsilon\in I\}$$

or $S_1=\{\varepsilon\mapsto(h_\varepsilon, H_\varepsilon)\in S|h_\varepsilon(0)=0,\ D_xh_\varepsilon(0)=0$ and $H_\varepsilon(0)=0\}$ for all $\varepsilon\in I$ is closed and invariant under the map $\mathscr{T}$, provided that (N2) (resp. (N2) and (N3)) holds. Hence, the unique fixed point of $\mathscr{T}$ in S lies in $S_0(S_1)$. Thus, part (i) is proved.

To prove (ii), let $(h_\varepsilon, H_\varepsilon)$ be given by the solution of (3.21) constructed above and define G_ε to be $\mathscr{G}(\varepsilon, h_\varepsilon, H_\varepsilon)$ for ε in I. Note that $p_1(\varepsilon)=D_xh_\varepsilon$, $q_1(\varepsilon)=D_xH_\varepsilon$, $p_2(\varepsilon)=H_\varepsilon$ and $q_2(\varepsilon)=G_\varepsilon$ defines a solution of the system of equations in $(3.14)_1$ and $(3.14)_2$. These relations are fulfilled in the space $(C^0(I, C^0(X, \mathscr{L}(X, Y)))\times C^0(I, C^0(X, \mathscr{L}^2(X, Y))))^2$. Furthermore, by (3.3), (3.13) and $\mathscr{G}(0,0,0)=0$,

$$p_1(0)=p_2(0)=0,\qquad q_1(0)=q_2(0)=0 \tag{3.22}$$

holds for this solution. Therefore it remains to show that this implies $p_1(\varepsilon)=p_2(\varepsilon)$ and $q_1(\varepsilon)=q_2(\varepsilon)$ for all $\varepsilon\in I$.

Using the formulas (3.19) (resp. (3.20)) and (3.17) (resp. (3.18)), we can rewrite the relations in $(3.14)_j$ $(j=1,2)$ in the form

$$(3.23)_j\qquad \dot p_j(\varepsilon)=\mathscr{H}_3(\varepsilon, p_j(\varepsilon), q_j(\varepsilon)),\qquad q_j(\varepsilon)=\mathscr{H}_4(\varepsilon, p_j(\varepsilon), q_j(\varepsilon), q_{l(j)}(\varepsilon))$$

where the maps $\varepsilon\mapsto\mathscr{H}_3(\varepsilon, p_j(\varepsilon), q_j(\varepsilon)): I\to C^0(X, \mathscr{L}(X, Y))$ and $\varepsilon\to\mathscr{H}_4(\varepsilon, p_j(\varepsilon), q_j(\varepsilon), q_{l(j)}(\varepsilon))$ are continuous, and

$$\|\mathscr{H}_3(\varepsilon, p_1(\varepsilon), q_1(\varepsilon))-\mathscr{H}_3(\varepsilon, p_2(\varepsilon), q_2(\varepsilon))\|_0\leqq K_{13}\alpha(\varepsilon),$$
$$\|\mathscr{H}_4(\varepsilon, p_1(\varepsilon), q_1(\varepsilon), q_2(\varepsilon))-\mathscr{H}_4(\varepsilon, p_2(\varepsilon), q_2(\varepsilon), q_1(\varepsilon))\|_0\leqq K_{13}\alpha(\varepsilon),$$

holds for all ε in I with some positive constant K_{13} and

$$\alpha(\varepsilon)=\sup(\|p_1(\varepsilon)-p_2(\varepsilon)\|_0, \|q_1(\varepsilon)-q_2(\varepsilon)\|_0).$$

Integrating the equations in $(3.23)_j$ from 0 to ε, subtracting the integral relations which are obtained for $j=1$ and $j=2$ and using (3.22), we get the following estimate

$$\alpha(\varepsilon)\leqq\left|\int_0^\varepsilon K_{13}\alpha(\sigma)\,d\sigma\right|\qquad(\varepsilon\in I).$$

Gronwall's lemma yields $\alpha(\varepsilon)=0$ for all ε in I, i.e., $p_1(\varepsilon)=p_2(\varepsilon)$ and $q_1(\varepsilon)=q_2(\varepsilon)$.

Taking the map h_ε which has been constructed in Lemma 3.4 and setting $\varepsilon=1$, the existence part of Theorem 2.1 follows according to the discussion previous to the above lemmas.

The proof also yields uniqueness, but only within the class of families of maps h_ε (where $\varepsilon \in I$) which have the properties stated in Lemma 3.4. To prove the uniqueness assertion of Theorem 2.1 we therefore have to give a different argument. Here we can even weaken our assumptions considerably.

LEMMA 3.5. *Suppose the maps $f \in C^0(X \times U, X)$ and $g \in C^0(X \times U, Y)$ satisfy a Lipschitz condition with respect to y with constant $\delta > 0$. Furthermore, assume that $\|B^{-1}\| < 1$ (resp. $\|B\| < 1$)) holds. Then for each ε in I, (3.2) has at most one solution $h_\varepsilon = h \in C^0_L(X, Y)$ (resp. such that the map $x \mapsto \phi_1(\varepsilon, x, h(x)) : X \to X$ is surjective), where*

$$L < (\|B^{-1}\|^{-1} - 1 - \delta\varepsilon)/(\delta\varepsilon) \qquad (\textit{resp. } L < (1 - \|B\| - \delta\varepsilon)/(\delta\varepsilon)).$$

Proof. Assume that h and $\tilde{h}$ are two such solutions of equation (3.2). Then,

$$h - \tilde{h} = B^{-1}(h(\phi_1) - \tilde{h}(\phi_1) + \tilde{h}(\phi_1) - \tilde{h}(\phi_1(\varepsilon, \cdot, \tilde{h}(\cdot))) + \varepsilon g(\cdot, \tilde{h}(\cdot)) - \varepsilon g(\cdot, h(\cdot)))$$

$$(\text{resp. } h(\phi_1) - \tilde{h}(\phi_1) = B(h - \tilde{h}) + \tilde{h}(\phi_1(\varepsilon, \cdot, \tilde{h}(\cdot))) - \tilde{h}(\phi_1) + \varepsilon g(\cdot, h(\cdot)) - \varepsilon g(\cdot, \tilde{h}(\cdot))),$$

in which we use our standing convention that $\phi_1 = \phi_1(\varepsilon, \cdot, h(\cdot))$. Thus by the assumptions

$$\|h - \tilde{h}\|_0 \leqq \|B^{-1}\|(1 + \delta\varepsilon L + \delta\varepsilon)\|h - \tilde{h}\|_0 < \|h - \tilde{h}\|_0$$

$$(\text{resp. } \|h - \tilde{h}\|_0 \leqq (\|B\| + \delta\varepsilon L + \delta\varepsilon)\|h - \tilde{h}\|_0 < \|h - \tilde{h}\|_0)$$

follows, which implies $h = \tilde{h}$.

Remark 3.6. The initial-value problem (3.12), (3.3) and (3.13) also has a unique solution in a ball around the origin in the space $C^0_L(X, Y) \times C^0_M(X, \mathscr{L}(X, Y))$ with appropriate constants L and M, even under the weaker assumptions that f and g are of class C^2_{Lip} and that $(\text{L}1)_3$ or $(\text{L}2)_3$ holds. However, it is not obvious how to show $D_x h_\varepsilon = H_\varepsilon$, to make sure that the solution actually yields a solution of (3.2).

On the other hand, one can still use the deformation principle to prove existence of a C^2_{Lip} invariant manifold under the weaker assumptions mentioned above. This requires the solution of the nonlinear equation (3.6) for H_ε in some space $C^1_M(X, \mathscr{L}(X, Y))$ as a Lipschitz continuous function of ε in I and $h_\varepsilon \in C^1_L(X, Y)$. Here the identity $D_x h_\varepsilon = H_\varepsilon$ follows from the fact that $D_x h_\varepsilon$ as well as H_ε are solutions of the first equation in $(3.14)_j$ for p_j, if we set $q_j = D_x H_\varepsilon$. In general, one does not have an explicit representation for the solution of (3.6); one can, however, use the contraction mapping principle to solve it. Thus, this method is a combination of the usual fixed point method to construct invariant manifolds [15] and the pure deformation method which we have proposed in the present paper.

4. Proof of Corollary 2.2. Corollary 2.2 is a consequence of Theorem 2.1 together with the following lemma; a bootstrapping argument accomplishes our purpose.

LEMMA 4.1. (a) *Assume that $f \in C^k(X \times U, X)$ and $g \in C^k(X \times U, Y)$ holds for some $k \geqq 3$. Furthermore, let $(\text{L}1)_k$ or $(\text{L}2)_k$, $\|f\|_1 < \delta$, and $\|g\|_1 < \delta$ hold. Suppose that for fixed ε, $h_\varepsilon = h \in C^{k-1}(X, Y)$ is a solution of (3.2). Then $h \in C^k(X, Y)$ if δ is sufficiently small, generally depending on k and $\|h\|_1$ for fixed A and B.*

(b) *If f and g are of class C^k_{Lip} for some $k \geqq 2$ and $(\text{L}1)_{k+1}$ or $(\text{L}2)_{k+1}$, $\|f\|_1 < \delta$, and $\|g\|_1 < \delta$ holds, then any C^k solution of (3.2) is contained in $C^k_{\text{Lip}}(X, Y)$ for sufficiently small $\delta > 0$.*

Proof. Since h is at least of class C^2 and a solution of (3.3), by uniqueness of the solution of (3.9) we have $D^2_{xx}h = \mathscr{G}(\varepsilon, h, D_x h)$ with $\mathscr{G}$ given either by (3.17) or by (3.18). Hence, it remains to show that $h \in C^{k-1}(X, Y)$ implies $\mathscr{G}(\varepsilon, h, D_x h) \in C^{k-2}(X, \mathscr{L}^2(X, Y))$ in case (a), and $h \in C^k(X, Y)$ implies $\mathscr{G}(\varepsilon, h, D_x h) \in C^{k-2}_{\text{Lip}}(X, \mathscr{L}^2(X, Y))$ in case (b).

If f and g are of class C^k and $h \in C^{k-1}(X, Y)$ for some $k \geqq 2$, then the $(k-2)$nd derivative of each term in the series (3.17) (resp. (3.18)) exists and is continuous. It is easily proved by induction with respect to k that these derivatives are of the form $(j = 0, 1, 2, \cdots)$

$$(4.1)\qquad \begin{aligned} &\sum_{\Sigma\alpha = k-2} \left(\prod_{i=0}^{j} D^\alpha_x \mathscr{B}^{-1}(\phi^i_1)\mathscr{P}_{\alpha i} \right) D^\alpha_x \mathscr{F}_3(\phi^j_1)\mathscr{P}_{\alpha j} \\ &\qquad \cdot \left(\prod_{i=j-1}^{0} D^\alpha_x \mathscr{A}(\phi^i_1)\mathscr{P}_{\alpha i}, \prod_{i=j-1}^{0} D^\alpha_x \mathscr{A}(\phi^{-i}_1)\mathscr{P}_{\alpha i} \right) \end{aligned}$$

or

$$(4.2)\qquad \begin{aligned} &\sum_{\Sigma\alpha = k-2} \left(\prod_{i=1}^{j} D^\alpha_x \mathscr{B}(\phi^{-i}_1)\mathscr{P}_{\alpha i} \right) D^\alpha_x \mathscr{F}_3(\phi^{-j-1})\mathscr{P}_{\alpha(j+1)} \\ &\qquad \cdot \left(\prod_{i=j+1}^{1} D^\alpha_x \mathscr{A}^{-1}(\phi^{-i}_1)\mathscr{P}_{\alpha i}, \prod_{i=j+1}^{1} D^\alpha_x \mathscr{A}^{-1}(\phi^{-i}_1)\mathscr{P}_{\alpha i} \right) \end{aligned}$$

where

$$\mathscr{B} = \mathscr{B}(\varepsilon, h, D_x h), \quad \mathscr{F}_3 = \mathscr{F}_3(\varepsilon, h, D_x h), \quad \mathscr{A} = \mathscr{A}(\varepsilon, h, D_x h),$$

and $\mathscr{P}_{\alpha i}$ is an α-tuple of products with i factors of the form

$$D^\alpha_x \mathscr{A}(\phi^n_1)\mathscr{P}_{\alpha n} \quad (0 \leqq n \leqq i-1)$$

$$(\text{resp. } D^\alpha_x \mathscr{A}^{-1}(\phi^{-n}_1)\mathscr{P}_{\alpha n} \quad (1 \leqq n \leqq i)).$$

These sums have less than $\frac{1}{2}k!(j+1)^{k-2}$ terms, each of which is a "product" of less than $(k+1)j+2$ (resp. $(k+1)(j+1)$) factors, with at least $j-k+3$ (resp. $j-k+2$) factors $\mathscr{B}^{-1}$ (resp. $\mathscr{B}$), at most kj(resp. $k(j+1)$) factors $\mathscr{A}$(resp. $\mathscr{A}^{-1}$). Besides these factors there are at most $k-2$ factors which are derivatives of such factors, or of $\mathscr{F}_3$ of order less than or equal to $k-2$. Of course, $\mathscr{F}_3$ is itself a factor if no derivative of it is contained in the product.

Now assume that

$$(4.3)_k \qquad \|\mathscr{B}^{-1}\|_0 \|\mathscr{A}\|^i_0 < q < 1 \quad (\text{resp. } \|\mathscr{B}\|_0 \|\mathscr{A}^{-1}\|^i_0 < q < 1) \quad (0 \leqq i \leqq k)$$

holds, where q is some real number which does not depend on i. Then

$$K_{14}(j+1)^{k-2}q^{j-k+3} \quad (j > k-3) \quad (\text{resp. } K_{14}(j+1)^{k-2}q^{j-k+2}(j > k-2))$$

is an upper bound for the C^0 norm of the sum in (4.1) (resp. (4.2)), where the constant K_{14} depends on k. Consequently, the series of these sums over j converges uniformly with respect to x and represents the $(k-2)$nd derivative of $\mathscr{G}(\varepsilon, h, D_x h)$. But, by $(\text{L}1)_k$ (resp. $(\text{L}2)_k$), $(4.3)_k$ is satisfied for some number q, provided that δ is sufficiently small. Thus, part (a) of Lemma 4.1 follows.

Under the assumptions of part (b) the sum in (4.1) (resp. (4.2)) is contained in $C^0_{\text{Lip}}(X, \mathscr{L}^k(X, Y))$ for each j. Furthermore, if $(4.3)_{k+1}$ holds, by the above information about this sum, its Lipschitz constant L_j can be estimated from above by

$$K_{15}(j+1)^{k-1}q^{j-k+2} \quad (j > k-2) \quad (\text{resp. } K_{15}(j+1)^{k-1}q^{j-k+1}(j > k-1))$$

with some constant K_{15} that depends on k. But by $(\mathrm{L1})_{k+1}$ (resp. $(\mathrm{L2})_{k+1}$), the condition $(4.3)_{k+1}$ is satisfied for sufficiently small $\delta > 0$. It follows that $D_x^{k-2}\mathscr{G}(\varepsilon, h, D_x h) \in C^0_{\mathrm{Lip}}(X, \mathscr{L}^k(X, Y))$, since $\sum_{j=0}^{\infty} L_j < \infty$. Thus, part (b) of Lemma 4.1 holds. □

Remark 4.2. In case of center manifolds, the C^k_{Lip} result, even for $k=0$ and $k=1$, is the usual result which is obtained by a fixed point argument ([15], [18]). For $k=0$ one assumes that $(\mathrm{L1})_1$ or $(\mathrm{L2})_1$ holds and that $\|f\|_0$, $\|g\|_0$, and the Lipschitz constants for f and g are sufficiently small. For $k \geqq 1$ the assumptions are analogous to those of Theorem 2.1. For $k=2$, see also Remark 3.6.

The C^2_{Lip} center-manifold theorem together with Lemma 4.1(a) now yields the center manifold theorem in C^k spaces for any $k \geqq 3$.

Moreover, observe that for fixed h_ε in $C^1(X, Y)$, (3.6) can be solved for H_ε in the space $C^1(X, \mathscr{L}(X, Y))$, provided that f and g are of class C^2, $\|f\|_1$ and $\|g\|_1$ are sufficiently small, and $(\mathrm{L1})_2$ or $(\mathrm{L2})_2$ holds (cf., Remark 3.6). Thus it follows that the C^1_{Lip} center manifold is actually contained in the class C^2 in this case; but this is the C^2 center manifold theorem.

Acknowledgments. We thank Ethan Akin, Marty Golubitsky, Jack Hale, Morris Hirsch, Pat McSwiggen and Charles Pugh for their comments.

REFERENCES

[1] R. Abraham, J. Marsden and T. Ratiu, *Manifolds, Tensor Analysis and Application*, Addison-Wesley, London, 1983, 2nd ed., Springer-Verlag, New York, to appear.

[2] H. Amann, *Gewöhnliche Differentialgleichugen*, Walter de Gruyter-Verlag, Berlin, 1983.

[3] J. Carr, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.

[4] S. N. Chow and J. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

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

[6] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193-226.

[7] J. Hadamard, *Sur l'iteration et les solutions asymptotiques des équations differentielles*, Bull. Soc. Math. France, 29 (1901), pp. 224-228.

[8] J. Hale, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[9] J. Hale and J. Scheurle, *Smoothness of bounded solutions of nonlinear evolution equations*, J. Differential Equations, 56(1) (1985), pp. 142-163.

[10] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser-Verlag, Boston, 1982.

[11] M. M. Hirsch, C. C. Pugh and M. Shub, *Invariant Manifolds*, Lecture Notes in Mathematics 583, Springer-Verlag, Berlin, 1977; see also M. M. Hirsch and C. Pugh, *Stable manifolds and hyperbolic sets*, Proc. Symp. Pure Math., 14 (1970), pp. 133-164.

[12] M. C. Irwin, *Smooth Dynamical Systems*, Academic Press, New York, 1980; see also, *On the stable manifold theorem*, Bull. London Math. Soc., 2 (1970), pp. 196-198; and *On the smoothness of the composition map*, Quart. J. Math., Oxford Ser. (2) 23 (1972), pp. 113-133.

[13] A. Källen, *On the proof of the centre manifold theorem*, J. London Math. Soc., II, Ser. 26 (1982), pp. 169-173.

[14] A. Kelley, *The stable, center-stable, center, center-unstable and unstable manifolds*, J. Differential Equations, 3 (1967), pp. 546-570.

[15] J. Marsden and M. McCracken, *The Hopf bifurcation and its applications*, Applied Mathematical Science, Vol. 19, Springer-Verlag, New York, 1976.

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

[17] R. J. Sacker, *A new approach to the perturbation theory of invariant surfaces*, Comm. Pure Appl. Math., 18 (1965), pp. 717-732.

[18] K. R. Schneider, *On quasicentre manifolds of semilinear equations in Banach spaces*, Math. Nachr., 122 (1985), pp. 215-229.

[19] J. Sijbrand, *Studies in nonlinear stability and birfurcation theory*, Ph.D. thesis, Univ. of Utrecht, The Netherlands, 1981.

[20] P. McSwiggen, private communication.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

006

GLOBAL APPROXIMATION OF PERTURBED HAMILTONIAN DIFFERENTIAL EQUATIONS WITH SEVERAL TURNING POINTS*

HARRY GINGOLD† AND PO-FANG HSIEH‡

Abstract. Given a perturbed Hamiltonian system (E) $i\varepsilon V' = [H_0(x) + \varepsilon H_1(x, \varepsilon)]V$, where $H_0(x)$ is a Hermitian matrix analytic on an interval I. Any two eigenvalues of $H_0(x)$ are allowed to coalesce finitely many times in the interval I (however, no two of them are identical on I). Assume that $H_1(x, \varepsilon) \in C^1(I \times \bar{S}_c)$ with $S_c = (0, c]$. Let $U(x)$ be the unitary matrix such that $D(x) = U^{-1}(x)H_0(x)U(x)$ is diagonal. Assume that the entries of $U^{-1}H_1U - iU^{-1}U'$ are bounded on $I \times S_c$, absolutely integrable over I. Then, (E) is shown to have a fundamental matrix of the form $V = U(x)Y(I_n + P(x, \varepsilon))$ where Y is the exponential of a diagonal matrix, I_n is the n-dimensional unit matrix and $P(x, \varepsilon) \to 0$ as $\varepsilon \to 0^+$. This result is applied to prove an adiabatic approximation theorem in quantum mechanics and provide a criterion measuring the phenomenon of degeneracy by the orders of coalescing of eigenvalues. Two examples are given.

Key words. global approximation, singularly perturbed Hamiltonian system, several turning points, adiabatic approximation theorem

AMS(MOS) subject classifications. Primary 34E20; secondary 34E15, 81C12

1. Introduction. When dealing with a system of differential equations depending in a singular way on a parameter, it is not usually possible to obtain *one* global asymptotic expression to approximate the solution. This is so for a linear system even though the global existence of its solution is theoretically guaranteed. For example, if we are dealing with a singularly perturbed system $\varepsilon y' = A(x)y$ on an interval $[a, b]$ where all eigenvalues of $A(x)$ are distinct, but for some eigenvalues of $A(x)$, the real part of some of their differences change sign on $[a, b]$, the asymptotic solutions are obtained only for subintervals. In order to investigate the solution on an entire interval $[a, b]$, one needs so-called connection formulas (e.g. see W. Wasow [22], [26]).

However, there are certain singular differential systems for which it is possible to obtain global asymptotic formulas on an entire interval. This is so even if the coefficient matrix possesses eigenvalues which coalesce finitely many times on that interval. Such points are commonly called turning points of the differential equations.

The presence of symmetry properties in a physical system is of interest to a physicist. Given a Hamiltonian system in quantum mechanics, some of its symmetry properties are manifested in certain degeneracies of its energy levels. When the Hamiltonian is time-dependent, the appearance of symmetries can show up as coalescence of certain energy eigenvalues. In other words, this is equivalent to the presence of turning, or transition, points of the system. This is our motivation to study a Hamiltonian system with one or several turning points of any (finite) order. As a matter of fact, we will study systems which are slightly more general than a Hamiltonian system.

Consider the following n-dimensional matrix differential system:

$$i\varepsilon V' = [H_0(x) + \varepsilon H_1(x, \varepsilon)]V, \qquad ' = \frac{d}{dx} \tag{1.1}$$

where $H_0(x)$ is a Hermitian matrix, analytic and no two eigenvalues are identical on $I = [a, b]$, and $H_1(x, \varepsilon)$ is in the class of $C^1(I \times \bar{S}_c)$ with $S_c = (0, c]$. Here a may be

* Received by the editors May 28, 1985; accepted for publication (in revised form) October 1, 1986.

† Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506. The work of this author was supported in part by a grant from the National Aeronautics and Space Administration.

‡ Department of Mathematics, Western Michigan University, Kalamazoo, Michigan 49008. The work of this author was supported in part by a Faculty Research Fellowship from Western Michigan University.

$-\infty$ and b may be $+\infty$. By a theorem of linear algebra due to F. Rellich [17] there exists a unitary matrix $U(x)$, analytic on I such that

$$\begin{aligned} D_1(x) &= U^{-1}(x)H_0(x)U(x) \\ &= \operatorname{diag}\{\lambda_1(x), \lambda_2(x), \cdots, \lambda_n(x)\} \end{aligned} \tag{1.2}$$

where $\{\lambda_j(x) \mid j=1,2,\cdots,n\}$ are the eigenvalues of $H_0(x)$, which are known to be real and analytic on I. Let

$$Y = U^{-1}(x)V. \tag{1.3}$$

Then, the n by n matrix Y satisfies a differential equation

$$i\varepsilon Y' = [D_1(x) + \varepsilon R_1(x, \varepsilon)]Y \tag{1.4}$$

with

$$R_1(x, \varepsilon) = U^{-1}(x)H_1(x, \varepsilon)U(x) - iU^{-1}(x)U'(x). \tag{1.5}$$

Let

$$D(x, \varepsilon) = D_1(x) + \varepsilon \operatorname{diag} R_1(x, \varepsilon), \qquad R(x, \varepsilon) = R_1(x, \varepsilon) - \operatorname{diag} R_1(x, \varepsilon). \tag{1.6}$$

Also, let

$$\begin{aligned} &R_0(x, \varepsilon) = \operatorname{diag} R_1 = \operatorname{diag}\{r_1^0, r_2^0, \cdots, r_n^0\}, \\ &R(x, \varepsilon) = (r_{jk}), \qquad j, k = 1, 2, \cdots, n. \end{aligned} \tag{1.7}$$

Then $r_{jj} \equiv 0$ and (1.4) becomes

$$i\varepsilon Y' = [D(x, \varepsilon) + \varepsilon R(x, \varepsilon)]Y. \tag{1.8}$$

Assume that

$$\lambda_j(x) - \lambda_k(x) \neq 0 \quad \text{for } x \in I \quad (j \neq k; j, k = 1, 2, \cdots, n), \tag{1.9}$$

$$\|H_1(x, \varepsilon)\| \leqq k, \quad x \in I, \varepsilon \in S_c \quad (j, k = 1, 2, \cdots, n) \tag{1.10}$$

where k is a constant independent of (x, ε) and $\| \; \|$ is a suitable norm of a matrix, and that

$$\int_a^b \|H_1(x, \varepsilon)\| \, dx \quad \text{is uniformly bounded for } \varepsilon \in S_c, \tag{1.11}$$

$$\int_a^b \|H_1'(x, \varepsilon)\| \, dx \quad \text{is uniformly bounded for } \varepsilon \in S_c. \tag{1.12}$$

We shall prove the following.

THEOREM 1. *Under the assumptions* (1.9)-(1.12), *the fundamental matrix of* (1.8) *can be expressed as*

$$Y = Z(x, \alpha, \varepsilon)(I_n + P(x, \varepsilon)), \qquad I_n : n \times n \textit{ identity matrix} \tag{1.13}$$

where $Z(x, \alpha, \varepsilon)$ is a fundamental matrix of

$$i\varepsilon Z' = D(x, \varepsilon)Z, \qquad Z(\alpha, \alpha, \varepsilon) = I_n, \tag{1.14}$$

with $\alpha \in I$, and $P(x, \varepsilon)$ is an n by n matrix in the class $C^1(I \times S_{\hat{c}})$, $(0 < \hat{c} \leqq c)$, $\|P(x, \varepsilon)\| = O(\varepsilon^d)$, with $d > 0$, uniformly on I as $\varepsilon \to 0^+$, satisfying $P(\alpha, \varepsilon) = 0$.

If the transformation (1.13) is taken to be $Y = (I_n + P)Z$, then the resulting equation for P is longer and requires more complicated computation to find its solutions.

The points $x_0 \in I$ where $\lambda_j(x_0) = \lambda_k(x_0)$ for certain j, k $(j \neq k; j, k = 1, 2, \cdots, n)$ are called the turning points of (1.8).

An immediate result of Theorem 1 is that the system (1.1) has a fundamental solution

$$V(x, \varepsilon) = U(x) \exp\left\{-i\varepsilon^{-1} \int_\alpha^x D(t, \varepsilon)\, dt\right\}(I_n + P(x, \varepsilon)), \tag{1.15}$$

which is uniformly valid on I. As the interval I contains one or more turning points of the differential equation, i.e., points where some of $\{\lambda_j(x)\}$ coalesce, (1.15) may be considered a central connection formula valid at all turning points of I. (It is called "two point connection formula" by H. Turrittin [20] when there are only two singularities present.) As pointed out by H. Turrittin [20] and J. A. M. McHugh [14], the lateral connection formulas (or sectorial connection formulas in [20]) follow from central connection formulas; thus it is important to have (1.15).

Furthermore, when $a = -\infty$ and/or $b = +\infty$, (1.15) not only gives the asymptotic approximation of the solutions of (1.1) for $\varepsilon \to 0^+$, it also provides the asymptotic approximation of the solutions in terms of x as x tends to $a = -\infty$ and/or $b = +\infty$. In this case $P(-\infty, \varepsilon)$ or $P(\infty, \varepsilon)$ may be chosen to be zero. This is why it is called the *doubly asymptotic* formulas for the solutions of (1.1) (cf. W. Wasow [26]).

Similar to [8], a differential system of the form (1.8) which is taken into (1.14) by (1.13) with a matrix $P(x, \varepsilon)$ satisfying the properties described in Theorem 1 may be called a *globally almost diagonal system.* An entirely different method is used in [8] to prove results similar to Theorem 1 for a system without the factor i in the left-hand side of the equation.

Theorem 1 will be used in § 9 to prove an adiabatic approximation theorem in quantum mechanics for an n-dimensional slowly varying time-dependent Hamiltonian system with degenerate energy levels. This theorem was first proved by M. Born and V. Fock [2] and also studied later in the general setting by T. Kato [10] (also see R. L. Liboff [13] and A. Messiah [15]), even though the case of degenerate energy levels caused by crossing of eigenvalues is not rigorously studied in the general setting. However, a rigorous comprehensive asymptotic decomposition for the solutions of the Hamiltonian system in the presence of multidegenerate energy levels is yet to be derived (cf. W. Wasow [24], [25]). A special degeneracy of a special two-dimensional system was studied by K. O. Friedrich [3], [4]. Recently, H. Gingold [6], [7] provided a comprehensive asymptotic decomposition method for a two-dimensional system with general degeneracy. We will illustrate in this paper that his method can be generalized to an n-dimensional system.

Using the principle of superposition (e.g. see R. L. Liboff [13]) a criterion equivalent to the adiabatic approximation theorem will be established, which enables us to measure *quantitatively* the *qualitative* phenomenon of degeneracy (or symmetry) by the *order of degeneracy* of eigenvalues in a slowly varying time-dependent Hamiltonian system.

In § 10, two examples of Theorem 1 will be given, one for bounded and the other for unbounded I.

The method of taking the system (1.8) into (1.14) is an essential tool in the study of singularly perturbed differential equations. For instance, W. Wasow [23] used it to calculate an adiabatic invariant of a second order equation with a simple turning point and A. Leung and K. Meyer [11] used it to study that for Hamiltonian systems with distinct and purely imaginary eigenvalues.

If $H_0+\varepsilon H_1$ in (1.1) is an infinite (or a general Hamiltonian) matrix, we cannot expect Theorem 1 to hold. However, if the "*orders*" *of all turning points of this infinite system are bounded*, then an analogue of Theorem 1 is expected to hold under fairly general conditions.

The requirement that H_0 is analytic can be relaxed considerably if we apply the methods in H. Gingold [5].

2. Preliminary reduction. From equations (1.8), (1.13) and (1.14), we have

$$iP'=Z^{-1}(x,\alpha,\varepsilon)RZ(x,\alpha,\varepsilon)(I_n+P),\quad P(\alpha,\varepsilon)=0,\quad \alpha\in I. \tag{2.1}$$

Equation (2.1) can be written equivalently as

$$P(x,\varepsilon)=-i\int_\alpha^x Z^{-1}(t,\alpha,\varepsilon)R(t,\varepsilon)Z(t,\alpha,\varepsilon)(I_n+P(t,\varepsilon))\,dt. \tag{2.2}$$

Put

$$LP=-i\int_\alpha^x Z^{-1}RZP\,dt, \tag{2.3}$$

and

$$P_0=LI_n. \tag{2.4}$$

Then, (2.2) is expressible as

$$P=P_0+LP, \tag{2.5}$$

or,

$$P=P_0+LP_0+L^2P=LI_n+L^2I_n+L^2P. \tag{2.6}$$

From (1.14), since $D(x,\varepsilon)$ is diagonal,

$$Z(x,\alpha,\varepsilon)=\exp\left\{-i\varepsilon^{-1}\int_\alpha^x D(t,\varepsilon)\,dt\right\}. \tag{2.7}$$

Thus,

$$\begin{aligned}L^2P=\int_\alpha^x & Z^{-1}(s,\alpha,\varepsilon)R(s,\varepsilon)Z(s,\alpha,\varepsilon)\\ &\cdot\left\{\int_\alpha^s Z^{-1}(t,\alpha,\varepsilon)R(t,\varepsilon)Z(t,\alpha,\varepsilon)P(t,\varepsilon)\,dt\right\}ds,\end{aligned} \tag{2.8}$$

or, by changing the order of integration,

$$\begin{aligned}L^2P=\int_\alpha^x & \left\{\int_t^x Z^{-1}(s,\alpha,\varepsilon)R(s,\varepsilon)Z(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z^{-1}(t,\alpha,\varepsilon)R(t,\varepsilon)Z(t,\alpha,\varepsilon)P(t,\varepsilon)\,dt.\end{aligned} \tag{2.9}$$

Put

$$D(x,\varepsilon)=\operatorname{diag}\{d_1(x,\varepsilon),d_2(x,\varepsilon),\cdots,d_n(x,\varepsilon)\}. \tag{2.10}$$

Then, by (1.6) and (1.7),

$$d_j(x, \varepsilon) = \lambda_j(x) + \varepsilon r_j^0(x, \varepsilon), \qquad j = 1, 2, \cdots, n. \tag{2.11}$$

Put

$$L^2P = (A_{jk}) \quad \text{and} \quad P = (p_{jk}), \quad j, k = 1, 2, \cdots, n. \tag{2.12}$$

By (1.7), (2.7) and (2.10), we have

$$\begin{aligned}&(Z^{-1}(s, \alpha, \varepsilon)R(s, \varepsilon)Z(s, \alpha, \varepsilon))_{jk}\\ &=\begin{cases} r_{jk}(s, \varepsilon) \exp\left\{i\varepsilon^{-1} \int_\alpha^s [d_j(\eta, \varepsilon) - d_k(\eta, \varepsilon)]\, d\eta\right\} & \text{if } j \neq k,\\ 0 \quad \text{if } j = k. \end{cases}\end{aligned} \tag{2.13}$$

Then,

$$\begin{aligned}A_{jk} = \int_\alpha^x \sum_{l=1}^n \Bigg[\sum_{h=1}^n \Bigg(\int_t^x r_{jh}(s, \varepsilon) \exp\left\{i\varepsilon^{-1} \int_\alpha^s (d_j(\eta, \varepsilon) - d_h(\eta, \varepsilon))\, d\eta\right\} ds\Bigg)\\ \cdot r_{hl}(t, \varepsilon) \exp\left\{i\varepsilon^{-1} \int_\alpha^t (d_h(\eta, \varepsilon) - d_l(\eta, \varepsilon))\, d\eta\right\}\Bigg] p_{lk}(t, \varepsilon)\, dt.\end{aligned} \tag{2.14}$$

In order to prove Theorem 1, we have to establish that

$$|||L^2P||| \leqq L(\varepsilon)|||P||| \tag{2.15}$$

for a suitable norm $|||\ \ |||$ of a matrix, where $L(\varepsilon)$ is a quantity which depends only on ε and tends to 0 as $\varepsilon \to 0^+$. We will introduce an alternate stationary phase method for several turning points, in several steps, in §§ 3-7.

The method used in the proof of Theorem 1 is different from those employed by W. A. Harris, Jr. and D. A. Lutz [9], N. Levinson [12], Y. Sibuya [18] and W. Wasow [22], [26]. The integral operator L given by (2.3) depends on the fundamental matrix $Z(x, s, \varepsilon)$ of (1.14), given by (2.7), but (2.6) is linear in P. Thus, it is possible to find $P(x, \varepsilon)$ uniformly valid on I, even though it contains one or more turning points of (1.8). Therefore, the tedious construction of the connection formulas can be avoided.

3. Alternative to stationary phase method. Consider

$$J(a, b, \alpha) = \int_a^b r(s, \varepsilon) \exp\left\{i\varepsilon^{-1} \int_\alpha^s p(\eta, \varepsilon)\, d\eta\right\} ds, \qquad a \leqq \alpha \leqq b \tag{3.1}$$

where $r(x, \varepsilon)$ is in $C^1(I \times \bar{S}_c)$, and $p(x, \varepsilon)$ is real analytic on $I \times \bar{S}_c$. Assume that the following conditions are satisfied:

(i) $p(x, 0)$ vanishes at some points of I, but is not identically zero on I;

(ii) There exists a positive constant g_1 such that

$$|p'(x, \varepsilon)| \leqq g_1 \quad \text{for } x \in I, \quad \varepsilon \in S_c; \tag{3.2}$$

(iii) There exist two positive constants m_1 and m_2 such that

$$|r(x, \varepsilon)| \leqq m_1 \quad \text{for } x \in I, \quad \varepsilon \in S_c, \tag{3.3}$$

$$\int_a^b |r(s, \varepsilon)|\, ds \leqq m_2, \quad \int_a^b |r'(s, \varepsilon)|\, ds \leqq m_2 \quad \text{for } \varepsilon \in S_c. \tag{3.4}$$

The zeros of $p(x, 0)$ are called the "turning points" of the integral $J(a, b, \alpha)$.

First note that, by integration by parts, (3.1) is expressible as

$$\begin{aligned} J(a,b,\alpha)&=\left[-i\varepsilon r(s,\varepsilon)\frac{1}{p(s,\varepsilon)}\exp\left\{i\varepsilon^{-1}\int_\alpha^s p(\eta,\varepsilon)\,d\eta\right\}\right]_a^b \\ &\quad+i\varepsilon\int_a^b \frac{r'(s,\varepsilon)p(s,\varepsilon)-r(s,\varepsilon)p'(s,\varepsilon)}{[p(s,\varepsilon)]^2}\exp\left\{i\varepsilon^{-1}\int_\alpha^s p(\eta,\varepsilon)\,d\eta\right\}ds. \end{aligned} \tag{3.5}$$

We shall use (3.5) to estimate the integral $J(a,b,\alpha)$ as $\varepsilon\to 0^+$ in several steps, each proved as a lemma.

LEMMA 1. *For α, β, $t\in[a,b]$, $J(a,b,\alpha)$ satisfies*

$$J(a,b,\alpha)=J(a,t,\alpha)+J(t,b,\beta)\exp\left\{i\varepsilon^{-1}\int_\alpha^\beta p(\eta,\varepsilon)\,d\eta\right\}. \tag{3.6}$$

Consequently, $|J(a,b,\alpha)|$ is independent of α and

$$|J(a,b,\alpha)|\leqq|J(a,t,\alpha)|+|J(t,b,\beta)|\quad \textit{for all } \alpha,\beta,t\in[a,b]. \tag{3.7}$$

LEMMA 2. *Suppose that a is finite, $p(x,\varepsilon)$ is independent of ε and expressible as*

$$p(x,\varepsilon)\equiv p(x)=(x-a)^{\nu_a}\hat p(x), \tag{3.8}$$

where ν_a is a positive integer, $\hat p(x)$ is real analytic on I and satisfying

$$0<g_2\leqq|\hat p(x)|\quad \textit{for } x\in I, \tag{3.9}$$

with g_2 a positive constant. Assume that the conditions (3.3) *and* (3.4) *hold. Then, there exist three positive constants K_a, l_a and c_a satisfying $0<\nu_a l_a<1$, $0<c_a\leqq c$ such that*

$$|J(a+\varepsilon^{l_a},t,\alpha)|\leqq K_a\varepsilon^{1-\nu_a l_a} \tag{3.10}$$

for $a+\varepsilon^{l_a}\leqq t$, $\alpha\leqq b$, $\varepsilon\in S_{c_a}$.

LEMMA 2A. *Suppose that $a=-\infty$, $p(x,\varepsilon)$ is independent of ε and expressible as*

$$p(x,\varepsilon)\equiv p(x)=x^{-\nu_a}\hat p(x) \tag{3.11}$$

where ν_a is a positive integer, $\hat p(x)$ is real analytic on I and satisfying (3.9). *Assume that* (3.3) *and* (3.4) *hold. Furthermore, assume that $r(x,\varepsilon)$ is expressible as*

$$r(x,\varepsilon)=x^{-2}\hat r(x,\varepsilon) \tag{3.12}$$

for $x\in(-\infty,-q)$, $\varepsilon\in S_c$, where q is a suitable positive constant, and

$$|\hat r(x,\varepsilon)|\leqq m_3 \tag{3.13}$$

for $s\in(-\infty,-q)$ and $\varepsilon\in S_c$ with m_3 a positive constant. Then, there exist three positive constants K_a, d_a, and c_a with $0<d_a<1$, $0<c_a\leqq c$ such that

$$|J(-\infty,t,\alpha)|\leqq K_a\varepsilon^{d_a} \tag{3.14}$$

for $-\infty<t$, $\alpha\leqq b$, $\varepsilon\in S_{c_a}$.

LEMMA 3. *Suppose that b is finite, $p(x,\varepsilon)$ is independent of ε and expressible as*

$$p(x,\varepsilon)\equiv p(x)=(x-b)^{\nu_b}\hat p(x) \tag{3.15}$$

where ν_b is a positive integer, $\hat p(s)$ is real analytic on I and satisfies (3.9). *Under the conditions* (3.3) *and* (3.4), *there exist three positive constants K_b, l_b and c_b satisfying $0<\nu_b l_b<1$ and $0<c_b\leqq c$ such that*

$$|J(t,b-\varepsilon^{l_b},\alpha)|\leqq K_b\varepsilon^{1-\nu_b l_b} \tag{3.16}$$

for $a\leqq t$, $\alpha\leqq b-\varepsilon^{l_b}$, $\varepsilon\in S_{c_b}$.

LEMMA 3A. *Suppose that $b=+\infty$, $p(x,\varepsilon)$ is independent of ε and expressible as*

$$p(x,\varepsilon)\equiv p(x)=x^{-\nu_b}\hat{p}(x) \tag{3.17}$$

where ν_b is a positive integer, $\hat{p}(s)$ is real analytic on I and satisfying (3.9). *Assume that* (3.3) *and* (3.4) *hold. Furthermore, assume that $r(x,\varepsilon)$ is expressible as in* (3.12) *and satisfies* (3.13) *for $x\in(q,+\infty)$, $\varepsilon\in S_c$ where q is a suitable positive constant. Then, there exist three positive constants K_b, d_b, and c_b with $0<d_b<1$, $0<c_b\leqq c$ such that*

$$|J(t,+\infty,\alpha)|\leqq K_b\varepsilon^{d_b} \tag{3.18}$$

for $a\leqq t$, $\alpha<+\infty$, $\varepsilon\in S_{c_b}$.

LEMMA 4. *Suppose that a and b are both finite, $p(x,\varepsilon)$ is independent of ε and expressible in the form*

$$p(x,\varepsilon)\equiv p(x)=(x-a)^{\nu_a}(x-b)^{\nu_b}\hat{p}(x) \tag{3.19}$$

where ν_a and ν_b are positive integers. Under the conditions (3.3), (3.4) *and* (3.9), *there exist five positive constants K_{ab}, l_a, l_b, d_{ab} and c_{ab}, with $0<\nu_a l_a<1$, $0<\nu_b l_b<1$, $0<d_{ab}<1$ and $0<c_{ab}\leqq c$, such that*

$$|J(a+\varepsilon^{l_a},b-\varepsilon^{l_b},\alpha)|\leqq K_{ab}\varepsilon^{d_{ab}} \tag{3.20}$$

for $a\leqq\alpha\leqq b$, $\varepsilon\in S_{c_{ab}}$.

LEMMA 5. *Suppose that $p(x,\varepsilon)$ is independent of ε and expressible in the form*

$$p(x,\varepsilon)\equiv p(x)=\left[\prod_{j=1}^{m}(x-\alpha_j)^{\nu_j}\right]\hat{p}(x) \tag{3.21}$$

where $a\leqq\alpha_1<\alpha_2<\cdots<\alpha_{m-1}<\alpha_m\leqq b$ (equality may hold only when $a\neq-\infty$, $b\neq\infty$), and $\nu_j(j=1,2,\cdots,m)$ are positive constants. Under the assumptions (3.3), (3.4) *and* (3.9), *there exist positive constants K_1, d_1 and c_1 $(0<d_1<1$, $0<c_1\leqq c)$ such that*

$$|J(a,b,\alpha)|\leqq K_1\varepsilon^{d_1}\quad\textit{for } a\leqq\alpha\leqq b,\quad \varepsilon\in S_{c_1}. \tag{3.22}$$

Lemma 1 follows easily from (3.1). As a matter of fact,

$$\begin{aligned}
J(a,t,\alpha)&+J(t,b,\beta)\exp\left\{i\varepsilon^{-1}\int_\alpha^\beta p(\eta,\varepsilon)\,d\eta\right\}\\
&=\int_a^t r(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\int_\alpha^s p(\eta,\varepsilon)\,d\eta\right\}ds\\
&\quad+\int_t^b r(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\left[\int_\beta^s p(\eta,\varepsilon)\,d\eta+\int_\alpha^\beta p(\eta,\varepsilon)\,d\eta\right]\right\}ds\\
&=\left(\int_a^t+\int_t^b\right)r(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\int_\alpha^s p(\eta,\varepsilon)\,d\eta\right\}ds=J(a,b,\alpha).
\end{aligned}$$

Remarks. (1) The points b in Lemmas 2, 2A and 5 may be $+\infty$ and a in Lemma 3, 3A and 5 may be $-\infty$.

(2) If l_a is chosen to be $l_a=(1+\nu_a)^{-1}$, then

$$1-\nu_a l_a=(1+\nu_a)^{-1}=l_a. \tag{3.23}$$

Then (3.10) agrees with the estimate obtained by the traditional stationary phase method (e.g. see F. W. J. Olver [16, p. 101]). The estimate for (3.16) can be treated similarly. Thus, our method contains the traditional stationary phase method as a special case, but *does not need* to utilize the integration in the complex plane.

(3) If $p(x,0)$ is in the form of (3.21), $x=\alpha_j$ $(j=1,2,\cdots,m)$ are called "turning points" of $J(a,b,\alpha)$ of order ν_j, respectively. If $p(x,0)$ is in the form of (3.17), the order of the turning point $x=+\infty$ is

$$\nu_\infty=\nu_b-2. \tag{3.24}$$

Similarly, if $p(x,0)$ is in the form of (3.11), the order of the turning point $x=-\infty$ is

$$\nu_{-\infty}=\nu_a-2. \tag{3.25}$$

(4) For the system (1.8), if $x=x_0$ is a turning point arised in more than one entries of A_{jk} (cf. (2.14)), the highest order is taken to be the order of the system at this turning point.

(5) To avoid cumbersome notation, the subscript h for ν_h, l_h, K_h and d_h, in this paper, is to denote either a point on $[a,b]$ or an index of the points on $[a,b]$. As it is clear from the text, we utilize both notation without warning.

(6) N. Bleistein [1] and F. Ursell [21] studied the asymptotic expansions of $J(a,b,\alpha)$ in fractional powers when several zeros of $p(x,\varepsilon)$ are nearly coincident at $x=0$. Their methods differ from ours as they employ transformations of variables and special functions. Their method is more in the spirit of the method of steepest descent, while ours is a straightforward alternative to the stationary phase method.

4. Proof of Lemmas 2 and 3. Let us prove Lemma 2 and consider first the case when b is finite. By (3.3), (3.8) and (3.9), we have

$$\left|\frac{r(t,\varepsilon)}{p(t)}\right|+\left|\frac{r(a+\varepsilon^{l_a},\varepsilon)}{p(a+\varepsilon^{l_a})}\right|\leqq\frac{2m_1}{g_2}\varepsilon^{-\nu_a l_a} \tag{4.1}$$

for $a+\varepsilon^{l_a}\leqq t\leqq b$.

In order to estimate the second term of (3.5), note that, since $\hat{p}(x)$ is analytic on I, there exists a positive constant g_3 such that

$$|\hat{p}'(x)|\leqq g_3 \quad \text{for } x\in I. \tag{4.2}$$

By (3.4), (3.8) and (3.9), we have

$$\int_{a+\varepsilon^{l_a}}^{t}\left|\frac{r'(s,\varepsilon)}{p(s)}\right|ds\leqq\frac{m_2}{g_2}\varepsilon^{-\nu_a l_a}. \tag{4.3}$$

(i) If $\nu_a>1$, by (3.3) and (3.8), we have

$$\begin{aligned}\int_{a+\varepsilon^{l_a}}^{t}\left|\frac{r(s,\varepsilon)p'(s)}{[p(s)]^2}\right|ds&\leqq m_1\int_{a+\varepsilon^{l_a}}^{t}\left|\frac{\nu_a}{(s-a)^{\nu_a+1}\hat{p}(s)}+\frac{\hat{p}'(s)}{(s-a)^{\nu_a}[\hat{p}(s)]^2}\right|ds\\&\leqq m_1\left\{\frac{\nu_a}{g_2}\int_{a+\varepsilon^{l_a}}^{t}\frac{ds}{(s-a)^{\nu_a+1}}+\frac{g_3}{g_2^2}\int_{a+\varepsilon^{l_a}}^{t}\frac{ds}{(s-a)^{\nu_a}}\right\}\\&\leqq m_1\left\{\frac{\nu_a}{g_2}\left(\frac{1}{\varepsilon^{\nu_a l_a}}-\frac{1}{(t-a)^{\nu_a}}\right)\right.\\&\qquad\left.+\frac{g_3}{g_2^2}\left(\frac{1}{(\nu_a-1)\varepsilon^{(\nu_a-1)l_a}}-\frac{1}{(\nu_a-1)(t-a)^{\nu_a-1}}\right)\right\}\\&\leqq m_1\left\{\frac{\nu_a}{g_2}\frac{1}{\varepsilon^{\nu_a l_a}}+\frac{g_3}{g_2^2}\frac{1}{(\nu_a-1)\varepsilon^{(\nu_a-1)l_a}}\right\}\\&=\frac{m_1}{\varepsilon^{\nu_a l_a}g_2}\left\{\nu_a+\frac{g_3}{g_2(\nu_a-1)}\varepsilon^{l_a}\right\}\\&=K_1\varepsilon^{-\nu_a l_a}\quad\text{for }\varepsilon\in S_{c_a}\end{aligned} \tag{4.4}$$

where $K_1=(m_1/g_2)\{\nu_a+(g_3/(g_2(\nu_a-1)))c_a^{l_a}\}$ with c_a to be specified.

(ii) If $\nu_a=1$, then $p'(s)=\hat{p}(s)+(s-a)\hat{p}'(s)$ and

$$\begin{aligned}\int_{a+\varepsilon^{l_a}}^{t}\left|\frac{r(s,\varepsilon)p'(s)}{[p(s)]^2}\right|ds &\leqq m_1\left\{\int_{a+\varepsilon^{l_a}}^{t}\left|\frac{1}{(s-a)^2\hat{p}(s)}+\frac{\hat{p}'(s)}{(s-a)\hat{p}(s)^2}\right|ds\right\}\\ &\leqq m_1\left\{\frac{1}{g_2}\int_{a+\varepsilon^{l_a}}^{t}\frac{ds}{(s-a)^2}+\frac{g_3}{g_2^2}\int_{a+\varepsilon^{l_a}}^{t}\frac{ds}{(s-a)}\right\}\\ &\leqq m_1\left\{\frac{1}{g_2}\left(\frac{1}{\varepsilon^{l_a}}-\frac{1}{t-a}\right)+\frac{g_3}{g_2^2}(|\log(t-a)|+|\log\varepsilon^{l_a}|)\right\}\\ &\leqq\frac{m_1}{g_2}\left\{\frac{1}{\varepsilon^{l_a}}+\frac{2g_3}{g_2}|\log\varepsilon^{l_a}|\right\}\leqq K_2\varepsilon^{-\nu_a l_a}\end{aligned}\tag{4.5}$$

if $0<\varepsilon<\min\{1,(b-a)^{1/l_a},(b-a)^{-1/l_a}\}$, where

$$K_2=\frac{m_1}{g_2}\left\{1+\frac{2g_3}{eg_2}\right\}.$$

Here we use the fact that $0>x^l\log x>-e^{-1}$ for $0<x<1$ and positive constant l. Thus, in either case, if

$$c_a=\min\{c,1,(b-a)^{1/l_a},(b-a)^{-1/l_a}\},\tag{4.6}$$

there exists a positive constant K_a such that (3.10) is satisfied for $\varepsilon\in S_c$.

If $b=\infty$, and $t\to\infty$, (4.1) and (4.3) hold also. Instead of using (4.4) or (4.5), observe that

$$\begin{aligned}\left|\int_{a+\varepsilon^{l_a}}^{\infty}\frac{r(s,\varepsilon)p'(s)}{[p(s)]^2}ds\right| &\leqq m_1\left|\int_{a+\varepsilon^{l_a}}^{\infty}\frac{p'(s)}{[p(s)]^2}ds\right|\\ &=m_1\left|\left[\frac{1}{p(s)}\right]_{a+\varepsilon^{l_a}}^{\infty}\right|\leqq\frac{m_1}{g_2}\varepsilon^{-\nu_a l_a},\end{aligned}\tag{4.7}$$

for $\varepsilon\in S_{c_a}$ since $p(\infty)=\infty$ in this case. In this case $c_a=\min\{c,1\}$.

Similarly, Lemma 3 is proved for both finite a and $a=-\infty$.

Remark. The condition (4.2) is used, instead of (3.2), when $p(x)$ is in the form of (3.8). The same is true for the situations given in Lemmas 2A-5.

5. Proof of Lemmas 2A and 3A. Let us prove Lemma 2A and consider first the case that b is finite. Let

$$\gamma_a=\begin{cases}-\infty & \text{if } b\geqq 0,\\ b^{-1} & \text{if } b<0,\end{cases}\tag{5.1}$$

and l_a be a positive constant satisfying $0<\nu_a l_a<1$. Note that $|J(a,b,\alpha)|$ is independent of α. Then, by (3.7), we have

$$|J(-\infty,b,\alpha)|\leqq|J(-\infty,-\varepsilon^{-l_a},\alpha)|+|J(-\varepsilon^{-l_a},\gamma_a^{-1},\beta)|+|J(\gamma_a^{-1},b,\gamma)|,\tag{5.2}$$

for $-\infty<\alpha,\ \beta,\ \gamma\leqq b$.

By (3.5), (3.3) and (3.4), as $p(s)$ is analytic, nonzero for $\gamma_a^{-1} \leqq s \leqq b$, there is a nonnegative constant K_1 such that

$$|J(\gamma_a^{-1}, b, \gamma)| \leqq K_1 \varepsilon \quad \text{for } -\infty < \gamma \leqq b, \qquad \varepsilon \in S_c. \tag{5.3}$$

In order to estimate $|J(-\infty, -\varepsilon^{-l_a}, \alpha)|$ and $|J(-\varepsilon^{-l_a}, \gamma_a^{-1}, \beta)|$, put

$$s = z^{-1}. \tag{5.4}$$

Then, by (3.1), (5.4), (3.11), (3.12) and (3.13), we have

$$\begin{aligned} |J(-\infty, -\varepsilon^{-l_a}, \alpha)| &= \left| \int_0^{-\varepsilon^{l_a}} \hat{r}\left(\frac{1}{z}, \varepsilon\right) z^2 \exp\left\{ i\varepsilon^{-1} \int_{\alpha^{-1}}^{z} p\left(\frac{1}{\eta}\right)\left(-\frac{d\eta}{\eta^2}\right)\right\}\left(-\frac{dz}{z^2}\right)\right| \\ &\leqq m_3 \varepsilon^{l_a} \end{aligned} \tag{5.5}$$

for $\varepsilon \in S_{c_a}$, where $-\varepsilon^{l_a} < \alpha^{-1} < 0$ and

$$c_a = \min\{1, c, q^{-1/l_a}, (-b)^{-1/l_a}\}. \tag{5.6}$$

Also, from (3.5), we have

$$\begin{aligned} |J(-\varepsilon^{-l_a}, \gamma_a^{-1}, \beta)| \leqq \varepsilon & \left| \left[r\left(\frac{1}{z}, \varepsilon\right) \frac{1}{z^{\nu_a} \hat{p}(1/z)} \right]_{-\varepsilon^{l_a}}^{\gamma_a} \right| \\ & + \varepsilon \int_{\gamma_a}^{-\varepsilon^{l_a}} \left| \frac{(dr/dz)(1/z, \varepsilon) p(1/z) - r(1/z, \varepsilon)(dp/dz)(1/z)}{[p(1/z)]^2} \right| dz. \end{aligned} \tag{5.7}$$

By (3.3), (3.11) and (5.6)

$$\left| \frac{r(\gamma_a^{-1}, \varepsilon)}{\gamma_a^{\nu_a} \hat{p}(\gamma_a^{-1})} \right| + \left| \frac{r(-\varepsilon^{-l_a}, \varepsilon)}{(-\varepsilon^{l_a})^{\nu_a} \hat{p}(-\varepsilon^{-l_a})} \right| \leqq \frac{2m_1}{g_2} \varepsilon^{-\nu_a l_a} \tag{5.8}$$

for $\varepsilon \in S_{c_a}$. By (3.4), (3.11) and (5.6), we have

$$\int_{\gamma_a}^{-\varepsilon^{l_a}} \left| \frac{(dr/dz)(1/z, \varepsilon)}{z^{\nu_a} \hat{p}(1/z)} \right| dz \leqq \frac{m_2}{\varepsilon^{\nu_a l_a} g_2} \tag{5.9}$$

for $\varepsilon \in S_{c_a}$. Also, similar to (4.4) and (4.5), there exists a positive constant K_4 such that

$$\int_{\gamma_a}^{-\varepsilon^{l_a}} \left| \frac{r(1/z, \varepsilon)(dp/dz)(1/z)}{[p(1/z)]^2} \right| dz \leqq K_4 \varepsilon^{-\nu_a l_a} \tag{5.10}$$

for $\varepsilon \in S_{c_a}$.

Thus, by (5.3), (5.5), (5.7), (5.8), (5.9) and (5.10), we have (3.14) for a suitable positive constant K_a, and

$$d_a = \min\{l_a, 1 - \nu_a l_a\}. \tag{5.11}$$

If $b = +\infty$, then, by (3.7), we can apply

$$|J(-\infty, \infty, \alpha)| \leqq |J(-\infty, \hat{\alpha}, \alpha)| + |J(\hat{\alpha}, \hat{\beta}, \beta)| + |J(\hat{\beta}, \infty, \gamma)| \tag{5.12}$$

for suitable finite constants $\hat{\alpha}$, $\hat{\beta}$, α, β and $\gamma (\hat{\alpha} \leqq \hat{\beta})$. The estimate of the first is shown above, that of the last term can be obtained in a similar fashion, and that of the middle term can be obtained by Lemma 2, or repeatedly using Lemma 2, similar to the proof of Lemma 4, if necessary. Thus, Lemma 2A is proved.

In a similar way, Lemma 3A is proved.

6. Proof of Lemma 4. For $p(x)$ in the form of (3.19), let $0<\nu_a l_a<1$, $0<\nu_b l_b<1$ and

$$c_{ab}=\min\left\{c,1,\left(\frac{b-a}{4}\right)^{1/l_a},\left(\frac{b-a}{4}\right)^{1/l_b},\left(\frac{b-a}{4}\right)^{-1/l_a},\left(\frac{b-a}{4}\right)^{-1/l_b}\right\}. \tag{6.1}$$

Here, a and b are both finite. Then, by Lemma 2 and Lemma 3, there are two positive constants K_{1a} and K_{1b} such that

$$\begin{aligned}\left|J\left(a+\varepsilon^{l_a},\frac{a+b}{2},\alpha\right)\right|&\leqq K_{1a}\varepsilon^{1-\nu_a l_a},\\ \left|J\left(\frac{a+b}{2},b-\varepsilon^{l_b},\beta\right)\right|&\leqq K_{1b}\varepsilon^{1-\nu_b l_b},\end{aligned} \tag{6.2}$$

for α, $\beta\in[a,b]$, $\varepsilon\in S_{c_{ab}}$. Let

$$d_{ab}=\min\{1-\nu_a l_a,1-\nu_b l_b\}. \tag{6.3}$$

Then, by Lemma 2, there exists a suitable positive constant K_{ab} such that

$$|J(a+\varepsilon^{l_a},b-\varepsilon^{l_b},\alpha)|\leqq K_{ab}\varepsilon^{d_{ab}} \tag{6.4}$$

for $\alpha\in[a,b]$, $\varepsilon\in S_{c_{ab}}$. Thus Lemma 4 is proved.

7. Proof of Lemma 5. For $p(x)$ in the form of (3.21) first choose m constants l_1, $l_2,\cdots,l_m$ satisfying

$$0<\nu_j l_j<1,\qquad j=1,2,\cdots,m. \tag{7.1}$$

Let

$$\begin{aligned}\varepsilon_a&=\begin{cases}1 & \text{if } \alpha_1=a,\\ \min\left\{\left(\frac{\alpha_1-a}{2}\right)^{1/l_a},\left(\frac{\alpha_1-a}{2}\right)^{-1/l_a}\right\} & \text{if } a<\alpha_1,\end{cases}\\ \varepsilon_b&=\begin{cases}1 & \text{if } \alpha_m=b,\\ \min\left\{\left(\frac{b-\alpha_m}{2}\right)^{1/l_m},\left(\frac{b-\alpha_m}{2}\right)^{-1/l_m}\right\} & \text{if } \alpha_m<b,\end{cases}\end{aligned} \tag{7.2}$$

and

$$\begin{aligned}c_1=\min\Bigg\{c,1,\varepsilon_a,\varepsilon_b,&\left(\frac{\alpha_{j+1}-\alpha_j}{4}\right)^{1/l_j},\left(\frac{\alpha_{j+1}-\alpha_j}{4}\right)^{-1/l_j},\\ &\left(\frac{\alpha_{j+1}-\alpha_j}{4}\right)^{1/l_{j+1}},\left(\frac{\alpha_{j+1}-\alpha_j}{4}\right)^{-1/l_{j+1}}\Bigg|\ j=1,2,\cdots,m-1\Bigg\}.\end{aligned} \tag{7.3}$$

Also, let

$$\delta_j=\varepsilon^{l_j},\qquad j=1,2,\cdots,m, \tag{7.4}$$

and

$$\delta_a=\begin{cases}0 & \text{if } \alpha_1=a,\\ \delta_1 & \text{if } a<\alpha_1,\end{cases}\qquad \delta_b=\begin{cases}0 & \text{if } \alpha_m=b,\\ \delta_m & \text{if } \alpha_m<b.\end{cases} \tag{7.5}$$

Consider

$$I_1=[\alpha_1-\delta_a,\alpha_1]\cup[\alpha_1,\alpha_1+\delta_1]\left\{\bigcup_{j=2}^{m-1}[\alpha_j-\delta_j,\alpha_j+\delta_j]\right\} \tag{7.6}$$
$$\cup[\alpha_m-\delta_m,\alpha_m]\cup[\alpha_m,\alpha_m+\delta_b]$$

and

$$I_2=[a,b]-\mathrm{Int}\,(I_1). \tag{7.7}$$

By Lemma 1, $|J(a,b,\alpha)|$ is decomposed into

$$|J(a,b,\alpha)|\leqq J_1+J_2 \tag{7.8}$$

where J_1 is the sum of absolute values of the integrals over each subinterval of I_1 and J_2 is the sum of those over each subinterval of I_2.

To estimate each term in J_1, using (3.1) and (3.3), we can find a suitable positive constant $\hat{K}_1$ such that

$$J_1=\int_{I_1}|r(s,\varepsilon)|\,ds\leqq 2m_1\sum_{j=1}^{m}\varepsilon^{l_j}\leqq\hat{K}_1\varepsilon^{\hat{d}} \tag{7.9}$$

for $\varepsilon\in S_{c_1}$, where

$$\hat{d}=\min\{l_1,l_2,\cdots,l_m\}. \tag{7.10}$$

In order to estimate J_2, as $\hat{p}(x)$ satisfies (3.9), let

$$\hat{g}_0=\left(\inf_{a_j\leqq x\leqq\alpha_1}\prod_{j=2}^{m}|x-\alpha_j|^{\nu_j}\right)g_2,$$
$$\hat{g}_m=\left(\inf_{\alpha_m\leqq x\leqq b}\prod_{j=1}^{m-1}|x-\alpha_j|^{\nu_j}\right)g_2, \tag{7.11}$$
$$\hat{g}_j=\left(\inf_{\alpha_j\leqq x\leqq\alpha_{j+1}}\prod\nolimits_j'|x-\alpha_i|^{\nu_i}\right)g_2\qquad(j=1,2,\cdots,m-1)$$

where $\prod_j'$ is the product over all i's satisfying $1\leqq i\leqq j-1$, $j+1\leqq i\leqq m$. Then each $\hat{g}_j$ $(j=0,1,2,\cdots,m)$ is a positive constant. Note that J_2 is the sum

$$J_2=|\hat{J}(a,\alpha_1-\delta_a,\beta_1)|+\sum_{j=1}^{m-1}|J(\alpha_j+\delta_j,\alpha_{j+1}-\delta_{j+1},\beta_j)|+|\hat{J}(\alpha_m+\delta_m,b,\beta_m)| \tag{7.12}$$

where $\beta_1,\beta_2,\cdots,\beta_m\in[a,b]$. By (7.5)

$$\hat{J}(a,\alpha_1-\delta_a,\beta_1)=\begin{cases}0 & \text{if } a=\alpha_1,\\ J(a,\alpha_1-\delta_1,\beta_1) & \text{if } a<\alpha_1,\end{cases} \tag{7.13}$$
$$\hat{J}(\alpha_m+\delta_b,b,\beta_m)=\begin{cases}0 & \text{if } \alpha_m=b,\\ J(\alpha_m+\delta_m,b,\beta_m) & \text{if } \alpha_m<b.\end{cases}$$

By Lemma 3, using $\hat{g}_0$ for g_2 in (3.9), we have

$$|\hat{J}(a,\alpha_1-\delta_a,\beta_1)|\leqq\hat{K}_0\varepsilon^{1-\nu_a l_a}\quad\text{for }\beta_1\in[a,b],\quad\varepsilon\in S_{c_1} \tag{7.14}$$

where $\hat{K}_0$ is zero if $a=\alpha_1$ and positive if $a<\alpha_1$. Similarly, by Lemma 2 and using $\hat{g}_m$ for g_2 in (3.9), we have

$$|\hat{J}(\alpha_m+\delta_b,b,\beta_m)|\leqq\hat{K}_m\varepsilon^{1-\nu_b l_b}\quad\text{for }\beta_m\in[a,b],\quad\varepsilon\in S_{c_1} \tag{7.15}$$

where $\hat{K}_m$ is zero if $\alpha_m=b$, and positive if $\alpha_m<b$.

Also, by Lemma 4, using $\hat{g}_j$ for g_2 in (3.9), we have

$$|J(\alpha_j+\delta_j,\ \alpha_{j+1}-\delta_{j+1},\ \beta_j)| \leqq \hat{K}_j\varepsilon^{\hat{d}_j}, \qquad j=1,2,\cdots,m-1, \tag{7.16}$$

for $\beta_j \in [a, b]$, $\varepsilon \in S_{c_1}$, where $\hat{K}_j$ is a suitable positive constant, and

$$\hat{d}_j = \min(1-\nu_j l_j,\ 1-\nu_{j+1}l_{j+1}), \qquad j=1,2,\cdots,m-1. \tag{7.17}$$

Let

$$d_1 = \min(\hat{d},\ 1-\nu_1 l_1,\ 1-\nu_2 l_2,\cdots,1-\nu_m l_m). \tag{7.18}$$

Then, by (7.8), (7.9), (7.12), (7.14), (7.15) and (7.16), there is a positive constant K_1 such that (3.22) is satisfied. Thus, Lemma 5 is proved.

Remark. The value of d_1 may be actually larger if $r(x, 0)$ vanishes at the same point where $p(x, 0)$ vanishes, namely at the turning points of $J(a, b, \alpha)$.

8. Completion of proof of Theorem 1. Note first that, by the analyticity of H_0, H_1, U and the conditions of (1.10)-(1.12), all the conditions of Lemmas 2-5 are satisfied. For $\alpha \in I$, let

$$\Delta_{jk}(x,\varepsilon) := d_j(x,\varepsilon) - d_k(x,\varepsilon) = q_{jk}(x) + \varepsilon\tilde{q}_{jk}(x,\varepsilon) \tag{8.1}$$

and

$$\begin{aligned} J_{jk}(a,b,\alpha) &:= \int_a^b r_{jk}(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\int_\alpha^s \Delta_{jk}(\eta,\varepsilon)\,d\eta\right\}ds, \\ &= \int_a^b \hat{r}_{jk}(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\int_\alpha^s q_{jk}(\eta)\,d\eta\right\}ds, \end{aligned} \tag{8.2}$$

$$j, k = 1, 2, \cdots, n, \quad j \neq k,$$

where

$$\hat{r}_{jk}(x,\varepsilon) = r_{jk}(x,\varepsilon)\exp\left\{i\int_\alpha^x \tilde{q}_{jk}(s,\varepsilon)\,ds\right\}.$$

By (1.11) and (1.12), $\hat{r}_{jk}(x, \varepsilon)$ satisfies the conditions (3.3) and (3.4). We shall show first that there exists a function $G_{jk}(\varepsilon)$ such that

$$|J_{jk}(a,t,\alpha)| \leqq G_{jk}(\varepsilon) \quad \text{for } t \in I \tag{8.3}$$

where

$$G_{jk}(\varepsilon) \text{ tends to 0 as } \varepsilon \to 0^+. \tag{8.4}$$

(i) If all the turning points of $J_{jk}(a, b, \alpha)$ are finite in order and in number, then $q_{jk}(x)$ is expressible as

$$q_{jk}(x) = \left[\prod_{i=1}^{m_{jk}} (x-\alpha_i^{(jk)})^{\nu_i^{(jk)}}\right]\hat{q}_{jk}(x) \tag{8.5}$$

where m_{jk} and $\nu_i^{(jk)}$ are positive integers, $\alpha_i^{(jk)} \in I (i = 1, 2, \cdots, m_{jk})$, and $\hat{q}_{jk}(x)$ is real analytic on I satisfying

$$0 < g_{jk} \leqq |\hat{q}_{jk}(x)| \tag{8.6}$$

for $x \in I$ with a positive constant g_{jk}. Then, by Lemma 5, there exists $G_{jk}(\varepsilon)$ such that (8.3) and (8.4) are true. It is noteworthy that in this case, a may be $-\infty$ and b may be $+\infty$.

(ii) If $a = -\infty$, b is finite and $a = -\infty$ is a turning point of $J_{jk}(a, b, \alpha)$, then there exists a positive integer $\nu_0^{(jk)}$ and a positive constant ρ_1, $(-\rho_1 \leqq b)$, such that $q_{jk}(x)$ is expressible as

$$q_{jk}(x) = x^{-\nu_0^{(jk)}}\hat{q}_{jk}^0(x) \tag{8.7}$$

for $x \in (-\infty, -\rho_1)$, where $\hat{q}^0_{jk}(x)$ is real analytic and satisfies

$$0 < g^0_{jk} < |\hat{q}^0_{jk}(x)| \tag{8.8}$$

for $x \in (-\infty, -\rho_1)$ with g^0_{jk} a positive constant. Since $U(x)$ is a real analytic unitary matrix on I, and $r_{jk}(x, \varepsilon)$ is an entry of $R_1(x, \varepsilon)$ given by (1.15), it satisfies (3.12) and (3.13) for $x \in (-\infty, -\rho_2)$ with certain positive constant ρ_2 satisfying $-\rho_2 \leqq b$. Let

$$\rho = \max\{\rho_1, \rho_2\}. \tag{8.9}$$

Then, by Lemma 2A, $J_{jk}(-\infty, -\rho, \alpha)$ satisfies (3.14). Combine this with the method of step (i) and by (3.7) we have (8.3).

Similarly, if $b = +\infty$, a is finite and $b = +\infty$ is a turning point of $J_{jk}(a, b, \alpha)$, by Lemma 3A, we have (8.3).

If $a = -\infty$, $b = +\infty$ and both are turning points of $J_{jk}(-\infty, +\infty, \alpha)$, apply above discussion to $J_{jk}(-\infty, -\rho, \alpha)$, $J_{jk}(-\rho, \rho, \alpha)$, $J_{jk}(\rho, +\infty, \alpha)$ and by (3.7), we have (8.3).

Thus, (8.3) is true for all situations.

Now, let $\|P(x, \varepsilon)\|$ be a suitable norm of P and

$$|||P||| = \sup_{x \in I} \|P(x, \varepsilon)\|. \tag{8.10}$$

Then, by (2.14), (1.10), (1.11) and (8.3),

$$\begin{aligned} |A_{jk}| &\leqq \sum_{l=1}^{n} \left[\sum_{h=1}^{n} G_{jh}(\varepsilon) \left| \int_{\alpha}^{x} r_{hl}(t, \varepsilon) \exp\left\{ i\varepsilon^{-1} \int_{\alpha}^{t} \Delta_{hl}(\eta, \varepsilon)\, d\eta \right\} dt \right| \right] |||P||| \\ &\leqq \left[\sum_{l,h=1}^{n} G_{jh}(\varepsilon) \hat{G}_{hl} \right] |||P||| \end{aligned} \tag{8.11}$$

where $\hat{G}_{hl}$ are suitable positive constants. Let

$$L(\varepsilon) = \max_{1 \leqq j \leqq n} \left\{ \sum_{l,h=1}^{n} G_{jh}(\varepsilon) \hat{G}_{hl} \right\}. \tag{8.12}$$

Then, we have

$$|||L^2 P||| \leqq L(\varepsilon) |||P||| \tag{8.13}$$

where $L(\varepsilon)$ tends to 0 as $\varepsilon \to 0^+$.

Similarly, if we let

$$\hat{L}(\varepsilon) = \|\hat{G}_{jk}(\varepsilon)\|, \tag{8.14}$$

then, we have

$$|||P_0||| = |||LI_n||| \leqq \hat{L}(\varepsilon) \tag{8.15}$$

with $\hat{L}(\varepsilon)$ tending to 0 as $\varepsilon \to 0^+$.

Furthermore, if we choose c_3 such that

$$0 < L(\varepsilon) < 1 \tag{8.16}$$

for $\varepsilon \in S_{c_3}$, then (2.6) defines a contraction mapping and

$$\begin{aligned} |||P||| &\leqq |||P_0||| + |||L^2 I_n||| + |||L^2 P||| \\ &\leqq \hat{L}(\varepsilon) + L(\varepsilon)|||I_n||| + L(\varepsilon)|||P|||. \end{aligned} \tag{8.17}$$

Therefore,

$$|||P||| \leqq \frac{\hat{L}(\varepsilon) + L(\varepsilon)|||I_n|||}{1 - L(\varepsilon)} \tag{8.18}$$

for $x \in I$, $\varepsilon \in S_{\hat{c}}$, where $\hat{c} = \min\{c_1, c_2, c_3\}$. Thus, Theorem 1 is proved.

9. Adiabatic approximation theorem. In this section, we shall investigate an adiabatic approximation theorem for an n-dimensional Hamiltonian system with multidegenerate energy levels. This theorem was first proved by M. Born and V. Fock [2] in 1928, and later studied in a general setting by T. Kato [10]. Although some cases of crossing of energy levels were discussed, as pointed out by Wasow [24], [25], the rigorous proof of a general situation as in Theorem 1 was not previously given.

This adiabatic approximation theorem will be shown to be true by observing the asymptotic behavior of the probabilities that eigenstates of the system stay in their original states. A comprehensive study of the *size* of such probabilities for a multidegenerate system has not been done before. This asymptotic analysis also will allow the computing of adiabatic invariants in the presence of several turning points. Although this paper studies the Hamiltonian system of a general degeneracy, the case of most general degeneracy, such as the degeneracy of two identical eigenvalues or of an infinite system, is not yet included here.

Given a slowly varying Hamiltonian operator $H(\varepsilon t)$ which depends on $x=\varepsilon t$ for small but nonvanishing parameter ε, where $t\in[0,\infty]$, $\varepsilon\in[0,c]$; namely $x\in[0,\infty]$. Given a state of the system $v(t,\varepsilon)$, an n-column vector, governed by Schrödinger equation

$$i\frac{dv}{dt}=H(\varepsilon t)v, \tag{9.1}$$

or equivalently

$$i\varepsilon v'=H(x)v, \qquad '=\frac{d}{dx}. \tag{9.2}$$

LEMMA 6. *Assume that $H(x)$ is analytic on $[0,\infty]$ and its eigenvalues coalesce finitely many times on $[0,\infty]$. Then, the fundamental matrix of (9.2) is*

$$V(x,\varepsilon)=U(x)\exp\left\{-i\varepsilon^{-1}\int_0^x D(s,\varepsilon)\,ds\right\}(I_n+P(x,\varepsilon)) \tag{9.3}$$

where $U(x)$ is a unitary matrix, $D(x,\varepsilon)$ is a diagonal matrix such that

$$D(x,0)=\operatorname{diag}\{\lambda_1(x),\lambda_2(x),\cdots,\lambda_n(x)\} \tag{9.4}$$

with $\{\lambda_j(x)\,|\,j=1,2,\cdots,n\}$ eigenvalues of $H(x)$, and $\|P(x,\varepsilon)\|=O(\varepsilon^d)$, $(d>0)$, uniformly on $[0,\infty]$ as $\varepsilon\to0^+$ with $P(0,\varepsilon)=0$.

Before proceeding to prove this lemma, we first compare some terminologies in mathematics and physics. An eigenvalue of H is referred as an energy level of the operator H, while the eigenvector of H is referred as its eigenstate. A solution $v(t,\varepsilon)$ of the time-dependent system (9.1) is referred as a state of the system. The evolution of a state $v(t,\varepsilon)$ is the observation of how is the dependence of v on t.

To show this lemma, let $U(x)$ be the unitary matrix such that

$$D_1(x)=U^{-1}(x)H(x)U(x)=\operatorname{diag}\{\lambda_1(x),\cdots,\lambda_n(x)\}. \tag{9.5}$$

Put

$$y=U^{-1}(x)v. \tag{9.6}$$

Then, (9.2) becomes

$$i\varepsilon y'=[D_1(x)+\varepsilon R_1(x)]y \tag{9.7}$$

where

$$R_1(x) = -iU^{-1}(x)U'(x). \tag{9.8}$$

Let

$$D(x, \varepsilon) = D_1(x) + \varepsilon \operatorname{diag} R_1(x). \tag{9.9}$$

Then, by Theorem 1, Lemma 6 follows immediately.

Now, let

$$U(x) = (u_1(x), u_2(x), \cdots, u_n(x)), \tag{9.10}$$

$$V(x, \varepsilon) = (v_1(x, \varepsilon), v_2(x, \varepsilon), \cdots, v_n(x, \varepsilon)). \tag{9.11}$$

Then, by (9.5), $u_j(x)$ is an eigenstate corresponding to $\lambda_j(x)$, $(j = 1, 2, \cdots, n)$. Let $D(x, \varepsilon)$ given by (9.9) be

$$D(x, \varepsilon) = \operatorname{diag}\{\lambda_j(x) + \varepsilon r_j(x)\}, \tag{9.12}$$

let α in Theorem 1 be 0, and

$$\lambda_j(x, \varepsilon) = \exp\left\{-i\varepsilon^{-1}\int_0^x [\lambda_j(s) + \varepsilon r_j(s)]\, ds\right\}. \tag{9.13}$$

Also, let e_j denote the jth unit vector of the standard basis of R^n. Then, by the notation for $P(x, \varepsilon)$ given by (2.12), we have

$$\begin{aligned} v_j(x, \varepsilon) &= U(x)\exp\left\{-i\varepsilon^{-1}\int_0^x D(s, \varepsilon)\, ds\right\}(I_n + P(x, \varepsilon))e_j \\ &= U(x)\begin{pmatrix} \hat{\lambda}_1 & & & 0 \\ & \hat{\lambda}_2 & & \\ & & \ddots & \\ 0 & & & \hat{\lambda}_n \end{pmatrix}\begin{pmatrix} p_{1j} \\ \vdots \\ p_{j-1,j} \\ 1 + p_{jj} \\ p_{j+1,j} \\ \vdots \\ p_{nj} \end{pmatrix} \\ &= \hat{\lambda}_j(x, \varepsilon)u_j(x) + U(x)\begin{pmatrix} \hat{\lambda}_1 p_{1j} \\ \hat{\lambda}_2 p_{2j} \\ \vdots \\ \hat{\lambda}_n p_{nj} \end{pmatrix} \\ &= \hat{\lambda}_j(x, \varepsilon)u_j(x) + \sum_{k=1}^{n} \hat{\lambda}_k(x, \varepsilon)p_{kj}(x, \varepsilon)u_k(x). \end{aligned} \tag{9.14}$$

Let $q_k^{(j)}$ be the probability of the state $v_j(x, \varepsilon)$ to be in the eigenstate $u_k(x)$. Note that $\hat{\lambda}_j(x, \varepsilon) = 0(1)$, $(j = 1, 2, \cdots, n)$; then, by the superposition principle (e.g. see R. L. Liboff [13, Chap. 5]),

$$q_j^{(j)} = |\hat{\lambda}_j|^2|1 + p_{jj}|^2/\{|\hat{\lambda}_j|^2|1 + p_{jj}|^2 + \sum_{h=1}^{n}{}' |\lambda_n P_{hj}|^2\}, \qquad \textstyle\sum': \text{sum over } h \neq j, \tag{9.15}$$

$$q_k^{(j)} = |\hat{\lambda}_k p_{kj}|^2/\{|\hat{\lambda}_j|^2|1 + p_{jj}|^2 + \sum_{h=1}^{n}{}' |\lambda_n P_{hj}|^2\} \qquad (k \neq j). \tag{9.16}$$

Since $\|P(x, \varepsilon)\| = O(\varepsilon^d)$, $(d > 0)$, *uniformly* on $[0, \infty]$ as $\varepsilon \to 0^+$, we have the following.

LEMMA 7. *The probabilities* $q_j^{(j)}$ *and* $q_k^{(j)}$ $(j,\ k=1,2,3,\cdots,n;\ k\neq j)$ *satisfy*

$$q_j^{(j)}-1=O(\varepsilon^{2d}),\qquad q_k^{(j)}=O(\varepsilon^{2d}),\tag{9.17}$$

uniformly on $[0,\infty]$ *as* $\varepsilon\to 0^+$.

By (9.14), we have

$$v_j(0,\varepsilon)=U(0)e_j=u_j(0),\qquad j=1,2,\cdots,n;\tag{9.18}$$

namely, $v_j(0,\varepsilon)$ is an eigenstate corresponding to the eigenvalue $\lambda_j(0)$. Thus, we have the following adiabatic approximation theorem.

THEOREM 2. *Suppose that the eigenvalues of* $H(\varepsilon t)$, *a slowly varying selfadjoint Hamiltonian operator, coalesce finitely many times for* $x=\varepsilon t$, $t\in[0,\infty]$ *and* $\varepsilon\in(0,c]$ *with* c *a small positive constant. Assume that* $H(x)$ *is analytic for* $x\in[0,\infty]$. *If the system started to evolve at* $t=0$ *such that* $v_j(0,\varepsilon)$ *is an eigenvector (an eigenstate) of* $H(0)$ *corresponding to the eigenvalue (energy level)* $\lambda_j(0)$, *then* $v_j(x,\varepsilon)$, *the evolution of that state is approximately the eigenvector of* $H(\varepsilon t)$ *for* $t\in[0,\infty]$, *corresponding to the eigenvalue* $\lambda_j(\varepsilon t)$, *while* $\varepsilon\to 0^+$ $(j=1,2,\cdots,n)$.

As pointed out in Remark 2 in § 3, if we use the estimate that could be provided by the traditional stationary phase method, then, by the proof of Theorem 1, we have, for each difference of eigenvalues $\lambda_j(x)-\lambda_k(x)$, $(j\neq k)$,

$$d_{jk}=\min\{l_h,l_\infty\}\tag{9.19}$$

where

$$\begin{aligned}&l_h=(\nu_h+1)^{-1},\ \nu_h:\text{ order of finite turning point } x=x_h,\\ &l_\infty=\begin{cases}(\nu_\infty+3)^{-1} & \text{if } x=\infty \text{ is a turning point with order } \nu_\infty,\\ +\infty & \text{if } x=\infty \text{ is not a turning point.}\end{cases}\end{aligned}\tag{9.20}$$

Then, d in (9.17) is given by

$$d=\min\{d_{jk}\,|\,1\leqq j,\ k\leqq n, j\neq k\}.\tag{9.21}$$

From (9.17) we can see that $q_j^{(j)}$ is closer to 1 and $q_k^{(j)}$ $(k\neq j)$ are closer to 0 if d is larger, i.e., ν_h and/or ν_∞ are smaller. Thus, we can measure the *qualitative* phenomenon of degeneracy (or symmetry) *quantitatively* by the *order* of the degeneracy of eigenvalues (energy levels) in a slowly varying time dependent Hamiltonian system.

Therefore, it is tempting to speculate that the following criterion is valid in quantum mechanical systems.

CRITERION. *In a slowly varying time-dependent Hamiltonian system* (9.1), *if a state of the system starts to evolve from an eigenstate, the less the order of degeneracy (symmetry), the closer is (in an asymptotic sense as* $\varepsilon\to 0^+$*) the state to its initial eigenstate.*

10. Examples. To illustrate Theorem 1, consider the following two examples.

Example 1. Given the following three-dimensional system:

$$i\varepsilon Y'=\left[\begin{pmatrix}x^{q+2} & 0 & 0\\ 0 & x^q(2x-1) & 0\\ 0 & 0 & 3x^q(2x-3)\end{pmatrix}+\varepsilon\begin{pmatrix}0 & r_{12} & r_{13}\\ r_{21} & 0 & r_{23}\\ r_{31} & r_{32} & 0\end{pmatrix}\right]Y\tag{10.1}$$

for $x\in[0,5]$, where q is a positive integer, $r_{jk}=r_{jk}(x,\varepsilon)$ $(j,\ k=1,2,3;\ j\neq k)$ satisfy (1.10)-(1.12) for $x\in[0,5]$, $\varepsilon\in S$. Denote

$$\lambda_1(x)=x^{q+2},\quad \lambda_2(x)=x^q(2x-1),\quad \lambda_3(x)=3x^q(2x-3),\tag{10.2}$$

and

$$D_1(x)=\operatorname{diag}\{\lambda_1(x),\lambda_2(x),\lambda_3(x)\}. \tag{10.3}$$

Then,

$$\begin{gathered}\lambda_1(x)-\lambda_2(x)=x^q(x-1)^2,\quad \lambda_1(x)-\lambda_3(x)=x^q(x-3)^2,\\ \lambda_2(x)-\lambda_3(x)=-4x^q(x-2).\end{gathered} \tag{10.4}$$

Thus, $x=0$, $x=1$, $x=2$ and $x=3$ are turning points of orders q, 2, 1 and 2, respectively. By Theorem 1, (10.1) has a fundamental solution

$$Y(x,\varepsilon)=\exp\left\{-i\varepsilon^{-1}\int_\alpha^x D_1(s)\,ds\right\}(I_3+P(x,\varepsilon)) \tag{10.5}$$

with

$$\|P(x,\varepsilon)\|\leqq K\varepsilon^d, \tag{10.6}$$

uniformly on $[0,5]$ for $\varepsilon\in S_{\hat c}$. Here K, d and $\hat c$ are suitable positive constants. As each obtained by the traditional stationary phase method, for $\lambda_1(x)-\lambda_2(x)$, $l_0=1/(1+q)$, $l_1=\frac{1}{3}$; for $\lambda_1(x)-\lambda_3(x)$, $l_0=1/(1+q)$, $l_3=\frac{1}{3}$; and for $\lambda_2(x)-\lambda_3(x)$, $l_0=1/(1+q)$, $l_2=\frac{1}{2}$. Then, by (7.10) and (7.18) $d=\min(1/(1+q),\frac{1}{3},\frac{1}{2})$.

Example 2. Given

$$i\varepsilon Y'=\left[\begin{pmatrix}1 & 0 & 0\\ 0 & 2x/(x^2+1) & 0\\ 0 & 0 & -2x/(x^2+1)\end{pmatrix}+\varepsilon\begin{pmatrix}0 & r_{12} & r_{13}\\ r_{21} & 0 & r_{23}\\ r_{31} & r_{32} & 0\end{pmatrix}\right]Y \tag{10.7}$$

for $x\in(-\infty,\infty)$ where $r_{jk}(x,\varepsilon)$ $(j,k=1,2,3;j\neq k)$ satisfy (1.10)-(1.12) for $x\in(-\infty,\infty)$, $\varepsilon\in S_c$. In this case,

$$\lambda_1(x)-\lambda_2(x)=\frac{(x-1)^2}{x^2+1},\quad \lambda_1(x)-\lambda_3(x)=\frac{(x+1)^2}{x^2+1},\quad \lambda_2(x)-\lambda_3(x)=\frac{4x}{x^2+1}. \tag{10.8}$$

Thus, $x=0$, ±1 are turning points of order 1 and 2, respectively, while $x=\pm\infty$ are turning points of order -2 for $\lambda_1(x)-\lambda_2(x)$ and $\lambda_1(x)-\lambda_3(x)$, and of order -1 for $\lambda_2(x)-\lambda_3(x)$. Then, by Theorem 1 and the estimates obtained in Lemmas 2-5, (10.7) has a fundamental solution (10.5) satisfying (10.6) uniformly on $[-\infty,\infty]$ for $\varepsilon\in S_{\hat c}$, with suitable positive constants K, d and $\hat c$. Then, by (7.10) and (7.18), $d=\frac{1}{3}$.

Acknowledgment. The authors wish to express their sincere gratitude to the referee for valuable suggestions to improve this work.

REFERENCES

[1] N. Bleistein, *Uniform asymptotic expansions of integrals with many nearby stationary points and algebraic singularities*, J. Math. Mech., 17 (1967), pp. 533-560.

[2] M. Born and V. Fock, *Beweis des Adiabatensatzes*, Z. Phys., 5 (1928), pp. 165-180.

[3] K. O. Friedrich, *Special topics in analysis*, Lecture Notes, New York University, New York, 1953.

[4] ———, *On the adiabatic theorem in quantum theory*, Report IMM·NYU-218, New York University, New York, 1955.

[5] H. Gingold, *On continuous triangularization of matrix functions*, this Journal, 10 (1979), pp. 709-720.

[6] ———, *An asymptotic decomposition method applied to multi-turning point problems*, this Journal, 16 (1985), pp. 7-27.

[7] ———, *Asymptotic decompositions on an entire interval for two by two first order linear differential system with multi-coalesing turning points*, preprint, West Virginia University, Morgantown, W.VA.

[8] H. GINGOLD AND P. F. HSIEH, *Global simplification of a singularly perturbed almost diagonal system*, this Journal, 17 (1986), pp. 7-18.

[9] W. A. HARRIS, JR. AND D. A. LUTZ, *A unified theory of asymptotic integration*, J. Math. Anal. Appl., 57 (1977), pp. 571-586.

[10] T. KATO, *On the adiabatic theorem of quantum mechanics*, J. Phys. Soc. Japan, 5 (1955), pp. 435-439.

[11] A. LEUNG AND K. MEYER, *Adiabatic invariants for linear Hamiltonian systems*, J. Differential Equations, 17 (1975), pp. 32-43.

[12] N. LEVINSON, *The asymptotic nature of solutions of linear differential equations*, Duke Math. J., 15 (1948), pp. 111-126.

[13] R. L. LIBOFF, *Introductory Quantum Mechanics*, Holden-Day, San Francisco, 1980.

[14] J. A. M. MCHUGH, *An historical survey of ordinary linear differential equations with a large parameter and turning points*, Arch. Hist. Exact. Sci., 7 (1971), pp. 277-324.

[15] A. MESSIAH, *Quantum Mechanics, Vol.* II, Interscience, New York, 1961.

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

[17] F. RELLICH, *Störungstheorie der Spektralzerlegung*, I, *Mitteilung*, Math. Ann., 113 (1936), pp. 600-619.

[18] Y. SIBUYA, *Simplification of a system of linear ordinary differential equations about a singular point*, Funkcial. Ekvac., 4 (1962), pp. 39-56.

[19]. G. STRANG, *Linear Algebra and its Applications*, Academic Press, New York, 1976.

[20] H. L. TURRITTIN, *Solvable related equations pertaining to turning point problems*, in Asymptotic Solutions of Differential Equations and Their Applications, C. H. Wilcox, ed., John Wiley, New York, 1964, pp. 27-52.

[21] F. URSELL, *Integrals with a large parameter. Several nearly coincident saddle-points*, Proc. Camb. Phil. Soc., 72 (1972), pp. 49-65.

[22] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

[23] ———, *Calculation of an adiabatic invariant by turning point theory*, this Journal, 5 (1974), pp. 673-700.

[24] ———, *Some recent results in the theory of adiabatic invariants*, in International Conference on Differential Equations, H. A. Antociewiz, ed., Academic Press, New York, 1975, pp. 746-764.

[25] ———, *Adiabatic invariants in the asymptotic theory of ordinary linear differential equations*, in Asymptotic Methods and Singular Perturbations, R. E. O'Malley, Jr., ed., SIAM-AMS Proceedings, 10, Providence, RI, 1976, pp. 131-144.

[26] ———, *Linear Turning Point Theory*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

007

TIME DEPENDENT NONLINEAR OSCILLATIONS WITH MANY PERIODIC SOLUTIONS*

JAMES DAMON†

Abstract. We investigate the local structure of nonlinear operators defining time dependent nonlinear oscillations. We explicitly describe the local structure of such operators using as models stable mappings between finite-dimensional spaces. This structure theorem allows us to prove the existence of many periodic solutions for such nonlinear oscillations. We determine the maximum number of such periodic solutions occurring near equilibrium. As one consequence we deduce that there exist oscillations with polynomial nonlinear terms of degree $2n$ for which there exist $(n+1)^2$ periodic solutions.

Key words. nonlinear oscillations, periodic solutions, infinitesimal stability, Lyapunov-Schmidt reduction, local algebras, nonlinear Fredholm mappings

AMS(MOS) subject classifications. 34C15, 34C25, 58F22

Introduction. The goal of this paper is to derive results concerning the number of small amplitude periodic solutions of the equation

$$(*) \qquad x''+m^2\lambda^2x+P(x,x',t)=q(t)$$

for $q(t)$ near 0. We assume that P and q are periodic in t of period $\lambda'=2\pi/\lambda$ and that $P(0,0,t)=0$, so that $x(t)\equiv 0$ is an equilibrium solution to $(*)$ when $q=0$. Our method is to explicitly determine the local structure near 0 of the operator

$$(**) \qquad F(x(t))=x''+m^2\lambda^2x+P(x,x',t)$$

where $F: C^2_{\lambda'}\to C^0_{\lambda'}$ and $C^k_{\lambda'}$ denotes the C^k-periodic functions of period $\lambda'=2\pi/\lambda$ defined on $\mathbb{R}$. In this formulation, solutions of $F(x(t))=q(t)$ correspond to small amplitude periodic solutions to $(*)$.

First of all, the results state that for a generic choice of P (in a sense to be made precise) the operator F is infinitesimally stable, and any such infinitesimally stable operator is locally equivalent as a mapping to $f_0\times \mathrm{id}$, for f_0 an infinitesimally stable mapping between finite dimensional spaces (id denotes the identity mapping on a Banach space). Moreover, for any infinitesimally stable $f_0:\mathbb{R}^k, 0\to\mathbb{R}^k, 0$ with $\dim\ker df_0(0)=2$, there is an operator F of the form $(**)$ locally equivalent to $f_0\times\mathrm{id}$. This infinitesimal stability has several main consequences:

(1) Since $\dim\ker(dF(0))=2$, the maximum number of small amplitude periodic solutions to $(*)$ which do occur for some small $q(t)$ is given by the dimension of a certain local algebra associated to F (this follows from the above decomposition together with results in [DG] giving explicitly the maximum number of solutions to $f(x)=y$ near 0 for $f:\mathbb{R}^n, 0\to\mathbb{R}^p, 0$ an infinitesimally stable germ).

(2) The maximum number of small periodic solutions remains constant under small perturbations P (and the periodic solutions themselves persist under small perturbations of P and q).

* Received by the editors October 1, 1985; accepted for publication October 29, 1986. This work was supported in part by a grant from the National Science Foundation.

† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514.

(3) This maximum number can be determined from a finite part of the Taylor expansion of P.

For the special case of P polynomial in x and x' (with coefficients depending on t), these results imply that there exists such a P of degree $2n$ and an appropriate (open set of) $q(t)$ for which $(*)$ has $(n+1)^2$ periodic solutions.

These results can be contrasted with the results previously obtained for the case where P does not depend on t and $q=0$. Then $(*)$ can be converted to a first-order system and the number of periodic solutions of $(*)$ for P and its perturbations was studied in the context of (degenerate) Hopf bifurcation by Takens [T] and Golubitsky and Langford [GL]. Here periodic solutions are created by perturbation of some basic P. As such the case is treated as a bifurcation problem. In contrast, infinitesimal stability implies that the local structure of the operator does not change under small perturbations. This corresponds to the difference between using the Lyapunov–Schmidt reduction to study the bifurcations of $F^{-1}(0)$ under perturbation by a finite number of external parameters, and using it to determine that all possible deformations of the operator itself already occur along the "internal directions" of the eigenfunctions.

For the polynomial case of $\deg 2n$, the number of periodic solutions we obtain exceeds by a factor of $n+1$ the number obtained by Takens and others, e.g. [CH], [S1] and [LMP]. However, even the number of periodic solutions we give is not generally the best possible for P of degree n. For $n=2$ or 3 the methods of this paper allow us to construct P of degree n with n^2 periodic solutions for $(*)$. Very likely this holds for all n, although technical details have prevented the verification of the more general results. Beyond the examples of degree 2 and 3 which we explicitly give in § 8, we have not been able to explicitly give P of degree $2n-2$ with n^2-periodic solutions. However, we are able to give very explicitly the form of a polynomial P of degree n and a $q(t)$ such that $(*)$ has $2n$-periodic solutions.

It seems likely that the methods used here are applicable to some other nonlinear problems. The main feature of nonlinear oscillations which is used is that the eigenfunctions of the linearized operator form an algebra whose complexification is the algebra of finite Laurent series. Thus, similar results should hold for other nonlinear operators whose linearization is selfadjoint with eigenfunctions forming an algebra possessing appropriate analogous properties.

1. Infinitesimally stable germs and operators.

Since the questions which we shall consider are local, we shall consider two mappings to be equivalent at a point x_0 if they agree in a neighborhood of x_0. Such an equivalence class is referred to as a germ of a mapping. Furthermore, by translation we may assume $x_0=0$ and $f(0)=0$; then $f: E,\ 0\to F,\ 0$ will be used to denote the germ of f at 0. Here, E and F may denote finite- or infinite-dimensional spaces, depending on the circumstances.

For a smooth germ $f:\mathbb{R}^n,\ 0\to\mathbb{R}^p,\ 0$, let $\pi: T\mathbb{R}^p\to\mathbb{R}^p$ denote the projection of the tangent bundle. Then, a germ of a smooth vector field along f is a smooth germ $\zeta:\mathbb{R}^n,\ 0\to T\mathbb{R}^p$ such that $\pi\circ\zeta=f$. Then, f is *infinitesimally* stable (as defined by Mather [M2]) if given ζ, there are smooth germs of vector fields ξ on $\mathbb{R}^n$ and η on $\mathbb{R}^p$ so that $\zeta=\xi(f)+\eta\circ f$ (here $\xi(f)$ denotes the directional derivative of f with respect to ξ). Next, suppose that E and F are Frechet spaces with $E=E_1\oplus G$, $F=F_1\oplus G$ with $\dim_{\mathbb{R}} E_1,\ F_1,\ <\infty$. Suppose $f: E,\ 0\to F,\ 0$ is a smooth germ of the form $f(x,u)=(\bar{f}(x,u),u)$ with $x\in E_1$, $u\in G$. If there is a finite-dimensional subspace $G_1\subset G$ such that the germ $f_1=(f|E_1\oplus G_1): E_1\oplus G_1,\ 0\to F_1\oplus G_1,\ 0$ is infinitesimally stable, then we shall say that *f is infinitesimally stable*. Here and in what follows, "smooth" will mean "smooth in the Gateaux differentiable sense." More generally, a smooth germ

$g: E, 0 \to F, 0$ will be said to be infinitesimally stable if it is *locally equivalent* to an infinitesimally stable f above by $f = \psi \circ g \circ \varphi$ for ψ and φ germs of local diffeomorphisms of E and F.

Ambrosetti and Prodi [AP] and Berger, Church and Timourian [BC], [BCT] have determined the structure of certain nonlinear operators using infinite-dimensional versions of folds and cusps. Quite generally, the local structure of smooth infinitesimally stable germs can be given using a slight modification of Mather's theorem "Infinitesimal stability implies stability." This is given in [D1].

THEOREM 1.1. *With the above notation, if $f: E, 0 \to F, 0$ is infinitesimally stable, then f is locally equivalent to $f_1 \times \mathrm{id}_{G_2}$ (where G_2 is a closed complement to G_1 in G).*

A corollary follows.

COROLLARY 1.2. *Suppose E, F and H are Banach spaces, $f: E, 0 \to F, 0$ is infinitesimally stable, and $h: E \oplus H, 0 \to F \oplus H, 0$ has the form $h(x, u) = (\bar{h}(x, u), u)$ with $\bar{h}(x, 0) = f(x)$ for $x \in E$, $u \in H$. Then, h is locally equivalent to $f \times \mathrm{id}_H$.*

To use these results, we recall several fundamental properties of infinitesimally stable germs due to Mather. If we apply the finite-dimensional analogue of the Lyapunov-Schmidt procedure to a germ $f: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$, then we may choose local coordinates $(x_1, \cdots, x_s, u_1, \cdots, u_q)$ for $\mathbb{R}^n$ and $(y_1, \cdots, y_t, u_1, \cdots, u_q)$ for $\mathbb{R}^p$ so that f has the form $f(x, u) = (\bar{f}(x, u), u)$, with $\bar{f}(x, 0)$ $(= f_0(x))$ having rank 0 at 0. In the terminology of singularity theory, such an f is called an unfolding of f_0 and the u_i are the unfolding parameters. With f in this form it now becomes a formal calculation to verify infinitesimal stability.

Let $\mathscr{E}_s$ denote the ring of smooth germs $g: \mathbb{R}^s, 0 \to \mathbb{R}$; it has a maximal ideal $\mathscr{m}_s$ consisting of germs vanishing at 0. In $(\mathscr{E}_s)^{(t)} = \mathscr{E}_s \oplus \mathscr{E}_s \cdots \oplus \mathscr{E}_s$ (t copies), which we can view as an $\mathscr{E}_s$-module, we consider the submodule L generated by $\partial f_0/\partial x_1, \cdots, \partial f_0/\partial x_s$, and the submodule $(I(f_0))^{(t)}$ (t-copies of $I(f_0) =$ ideal in $\mathscr{E}_s$ generated by the coordinate functions of f_0, $\{y_i \circ f_0, 1 \leq i \leq t\}$). By $\mathscr{m}_s^k$ we mean the ideal of $\mathscr{E}_s$ generated by monomials of degree $= k$. Then, the quotient $\mathscr{E}_s/\mathscr{m}_s^k$ is a finite-dimensional vector space with basis given by the monomials of degree $< k$. Last, letting $v = (v_1, \cdots, v_q)$, $\partial f/\partial v_i|_{v=0}$ is a germ on $\mathbb{R}^s$, and hence via its coordinate functions can be viewed as an element of $(\mathscr{E}_s)^{(t)}$, in fact, of $(\mathscr{m}_s)^{(t)}$. Then Proposition 1.8 of [M4] can be applied.

THEOREM 1.3 (Verification criterion). *f is infinitesimally stable iff $\partial \bar{f}/\partial v_1|_{v=0}, \cdots, \partial \bar{f}/\partial v_q|_{v=0}$ span the quotient space*

$$N(f_0) = (\mathscr{m}_s)^{(t)}/(L + I(f_0)^{(t)} + (\mathscr{m}_s^{p+1})^{(t)}).$$

As this space is finite-dimensional, this becomes a very computable criterion using part of the Taylor series of f.

An important invariant of a germ is its local algebra. For $f: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$, we define the local algebra $Q(f) = \mathscr{E}_n/I(f)$, where, just as earlier, $I(f)$ is the ideal generated by the coordinate functions of f. If f is an unfolding of f_0 as above, then $I(f) = I(f_0) + \mathscr{m}_q \cdot \mathscr{E}_n$; hence $Q(f) \xrightarrow{\sim} Q(f_0)$. Furthermore, for infinitesimally stable germs there is the very fundamental classification theorem of Mather [M4, Thm. A].

THEOREM 1.4 (Classification by local algebras). *If $f, g: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$ are infinitesimally stable germs with isomorphic local algebras then f and g are locally equivalent at 0.*

Furthermore, given an algebra Q satisfying certain conditions it is possible to construct an infinitesimally stable germ f, with $Q(f) \simeq Q$. For example, consider $Q \simeq \mathscr{E}_s/I$ with $I \subset \mathscr{m}_s^2$ generated by $f_{01}, \cdots, f_{0t}$ and suppose for simplicity that $\dim_{\mathbb{R}} Q < \infty$. Then, we use the f_{0i} as coordinate functions to define $f_0: \mathbb{R}^s, 0 \to \mathbb{R}^t, 0$.

Then, $N(f_0)$ is a subspace of a quotient of $(Q)^{(t)} \xrightarrow{\sim} (Q(f_0))^{(t)}$. Hence, there are $\varphi_1, \cdots, \varphi_q \in (m_s)^{(t)}$ which project to a basis for $N(f_0)$. Then, $f: \mathbb{R}^{s+q}, 0 \to \mathbb{R}^{t+q}, 0$ defined by

$$f(x_1, \cdots, x_s, v_1, \cdots, v_q) = \left(f_0(x) + \sum_{i=1}^{q} v_i\varphi_i, v_1, \cdots, v_q\right) \tag{1.5}$$

is infinitesimally stable by Theorem 1.1 and has a local algebra $Q(f) \xrightarrow{\sim} Q(f_0) \xrightarrow{\sim} Q$.

This leads to the third property.

THEOREM 1.6 (Normal form for infinitesimally stable germs). *Given an algebra Q satisfying certain conditions* (*which include the case* $\dim_{\mathbb{R}} Q < \infty$) (*see* [M4, *Thm.* B]), *then there is an infinitesimally stable germ f of the form* (1.5) *with local algebra $Q(f) \xrightarrow{\sim} Q$.*

These three results allow one to begin with an infinitesimally stable germ g, compute its local algebra $Q(g)$, construct a normal form f with local algebra $Q(f) \xrightarrow{\sim} Q(g)$. Then, by the classification theorem, f and g are locally equivalent, so the specifically given germ f can be used instead to study the local structure of g.

One consequence of these results is the following:

If $f: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$ is infinitesimally stable and $n \leqq p$ then by Mather [M4] $\dim_{\mathbb{R}} Q(f) < \infty$. Let N be the smallest integer so that $m_n^N \cdot Q(f) = 0$.

COROLLARY 1.7. *In the above situation if $h: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$ with $h = (h_1, \cdots, h_p)$ and $h_j \in m_n^{N+1}$, $1 \leqq j \leqq p$, then $f_1 = f + h$ is locally equivalent to f.*

Proof. By the classification theorem it is sufficient to show that $f + h$ is infinitesimally stable and $Q(f+h) \xrightarrow{\sim} Q(f)$. We may assume that we have chosen coordinates for $\mathbb{R}^n$ and $\mathbb{R}^p$ so that f has the form $f(x, u) = (\bar{f}(x, u), u)$ with $d\bar{f}(0) = 0$. We must further change coordinates to put $f_1 = f + h$ in this form; however, this will only change the coordinate functions of f by terms in m_n^{N+1}. Thus, f_1 has the form $(\bar{f}_1(x, u), u) = (\bar{f}(x, u) + g(x, u), u)$ with $g = (g_1, \cdots, g_t)$ and $g_i \in m_n^{N+1}$. Then, by the conditions on g,

$$I(f) = I(f_1) + m_n^{N+1} \quad \text{or} \quad I(f) = I(f_1) + m_n \cdot I(f).$$

By Nakayama's Lemma, $I(f) = I(f_1)$.

Second, with v_i denoting x_i or u_i,

$$\frac{\partial(\bar{f} + g)}{\partial v_i} \equiv \frac{\partial \bar{f}}{\partial v_i} \operatorname{mod} m_n^N.$$

Thus, in the verification criterion

$$N(\bar{f}(x, 0)) = N(\bar{f}_1(x, 0))$$

and

$$\left.\frac{\partial \bar{f}}{\partial u_i}\right|_{u=0} = \left.\frac{\partial f_1}{\partial u_i}\right|_{u=0} \quad \text{in } N(\bar{f}_1(x, 0)).$$

Thus, f_1 is infinitesimally stable and since $I(f_1) = I(f)$, $Q(f_i) \xrightarrow{\sim} Q(f)$. □

One consequence of the infinitesimal stability of a germ $f: \mathbb{R}^n, 0 \to \mathbb{R}^p, 0$ with $n \leqq p$ occurs when we wish to count solutions x near 0 to the equation $f(x) = y$ for y near 0. We define the *real multiplicity* of f:

$m(f) \overset{\text{def}}{=} \max \{k$: for all sufficiently small open neighborhoods V of 0 in $\mathbb{R}^p$ and U of 0 in $\mathbb{R}^n$, there is a $y \in V$ so that $\operatorname{card}(f^{-1}(y) \cap U) = k\}$.

For infinitesimally stable germs, $m(f)$ has several important properties summarized in the following theorem (see [DG] or [D2]).

THEOREM 1.8. *For infinitesimally stable f*

(1) *$m(f)$ only depends on $Q(f)$, i.e., it will be the same for different infinitesimally stable f between different dimensional spaces as long as they have isomorphic local algebras.*

(2) *In the case that* $\dim_{\mathbb{R}} \ker (df(0)) \leqq 2$ *or $Q(f)$ satisfies other conditions in* [DG] *then*

$$m(f) = \delta(f) \overset{\text{def}}{=} \dim_{\mathbb{R}} Q(f).$$

(3) *Even if f is not infinitesimally stable at* 0 *but $\delta(f) < \infty$, then $m(f) \leqq \delta(f)$ (see* [GG]).

We conclude this section by deducing from Theorem 1.1 and Corollary 1.2 consequences for germs of smooth Fredholm mappings $f: E, 0 \to F, 0$ between Banach spaces E and F. By the Lyapunov-Schmidt procedure, we may assume f has the form $\bar{f}(x, u) = (\bar{f}(x, u), u)$ for $u \in G$, $x \in E_1$ and $E \xrightarrow{\sim} E_1 \oplus G$, $F \xrightarrow{\sim} F_1 \oplus G$. Then, $f_0(x) = \bar{f}(x, 0): E_1, 0 \to F_1, 0$ is a smooth mapping between finite-dimensional spaces and so has a well-defined local algebra $Q(f_0)$.

LEMMA 1.9. *If $Q(f_0)$ is isomorphic to an algebra of an infinitesimally stable germ, then the algebra obtained by carrying out the Lyapunov-Schmidt procedure in a different way is still isomorphic to $Q(f_0)$.*

Thus, the algebra is an invariant of f and so, for such an f, we may denote the algebra by $Q(f)$ and call it the *local algebra of the smooth germ of a Fredholm mapping f.*

Note. In the case that f is Fredholm of index $\leqq 0$, then the condition on $Q(f_0)$ is equivalent to $\dim_{\mathbb{R}} Q(f_0) < \infty$ by another result of Mather [M4].

Proof. There is a germ $g: E \oplus \mathbb{R}^k, 0 \to F \oplus \mathbb{R}^k, 0$ defined by $g(x, u, v) = (\bar{f}(x, u) + \sum_{i=1}^{k} v_i \varphi_i, u, v)$, where $f_1(x, v) = (f_0(x) + \sum_{i=1}^{n} v_i \varphi_i, v)$ is a normal form for an infinitesimally stable germ with algebra $Q(f_0)$. Then, by Theorem 1.1, g is locally equivalent to $f_1 \times \mathrm{id}_G$ by $g = \Psi \circ (f_1 \times \mathrm{id}_G) \circ \Phi$ with Ψ, Φ germs of diffeomorphisms satisfying $\Psi | F_1 \oplus \mathbb{R}^n = \mathrm{id}$, $\Phi | E_1 \oplus \mathbb{R}^n = \mathrm{id}$.

If a Lyapunov-Schmidt procedure were applied using different local coordinates then in place of E_1 and F_1 we would obtain finite-dimensional submanifolds E_1' and F_1' such that $T_0 E_1' = E_1$ and $T_0 F_1' \oplus G = F$. Then, $f_0': E_1', 0 \to F_1', 0$ would be obtained as the restriction of g to E_1'. Hence, if E_1'' and F_1'' denote the images of E_1' and F_1' via Φ^{-1} and Ψ^{-1}, then f_0' is locally equivalent to $f_0'': E_1'', 0 \to F_1'', 0$ with f_0'' the restriction of $f_1 \times \mathrm{id}_G$ to E_1''. By the properties of Φ and Ψ, if p_1 and p_2 denote the linear projections $E \oplus \mathbb{R}^n \to E_1 \oplus \mathbb{R}^n$ and $F \oplus \mathbb{R}^n \to F_1 \oplus \mathbb{R}^n$ then $p_1 | T_0 E_1''$ is an isomorphism with image transverse to $\mathbb{R}^n$ and similarly for $p_2 | T_0 F_1''$. Let $E_1''' = p_1(E_2'')$, $F_1''' = p_2(F_1'')$ and $f_0''' = f_1 \times \mathrm{id} | E_1'''$. Then, $f_1 = p_2 \circ (f_1 \times \mathrm{id}_G)$ is constant on the fibers of p_1 so f_0''' is locally equivalent to f_0''. Finally, as E_1''' and F_1''' are transverse to $\mathbb{R}^n$ at 0, it follows that f_1 can be represented as an unfolding of f_0'''. Hence, $Q(f_0''') \xrightarrow{\sim} Q(f_1) \xrightarrow{\sim} Q(f_0)$. As f_0''' is locally equivalent to f_0', $Q(f_0''') \xrightarrow{\sim} Q(f_0')$ and the result follows. □

We can then summarize the consequences for germs of smooth nonlinear Fredholm maps.

THEOREM 1.10. *Let $f: E, 0 \to F, 0$ be an infinitesimally stable germ of a mapping of Banach spaces. Then*

(1) *f is locally equivalent to an infinitesimally stable germ $g: E, 0 \to F, 0$ iff $Q(f) \xrightarrow{\sim} Q(g)$ and their Fredholm indices are equal.*

(2) *Let $\mathfrak{m}$ denote the maximal ideal of $Q(f)$ and suppose $\mathfrak{m}^N \cdot Q(f) = 0$. If $h: E, 0 \to F, 0$ is a smooth germ such that $\|h\| \leqq O(\|x\|^{N+1})$, then $f + h$ is locally equivalent to f.*

(3) *If $df(0)$ has index $\leqq 0$ (as a Fredholm mapping) then $m(f)$ only depends on $Q(f)$ (i.e. it is the same for any other infinitesimally stable germ $f': E', 0 \to F', 0$ even with a different index $\leqq 0$ as long as $Q(f') \xrightarrow{\sim} Q(f)$).*

(4) *As in (3), and moreover if $\dim_{\mathbb{R}} \ker (df(0)) \leqq 2$ or $Q(f)$ satisfies other conditions in* [DG], *then $m(f) = \delta(f) \overset{\text{def}}{=} \dim_{\mathbb{R}} (Q(f))$.*

(5) *Even if f is not infinitesimally stable, but $Q(f)$ is isomorphic to the local algebra of an infinitesimally stable germ and the Fredholm index of $f \leqq 0$, then $m(f) \leqq \delta(f)$.*

2. Statements of main theorems. We use the notation and terminology of the preceding section to state the main results. We consider solutions to

$$x'' + m^2\lambda^2 x + P(x, x', t) = q(t) \tag{2.1}$$

with P continuous in (x, x', t), smooth in (x, x') but *without* linear terms and periodic of period $2\pi/\lambda$ in t. Associated to (2.1) we have the operator

$$f(x) = x'' + m^2\lambda^2 x + P(x, x', t) \tag{2.2}$$

as an operator $f: C^2_{(2\pi/\lambda)} \to C^0_{(2\pi/\lambda)}$ (recall that $C^k_{(2\pi/\lambda)}$ denotes the space of $2\pi/\lambda$-periodic C^k-functions on $\mathbb{R}$).

Many of the results about solutions of (2.1) for $q(t)$ near 0 follow from results about the operator f.

The first result is a weak form of genericity.

THEOREM 1. *Given $N \geqq 2$ with m and λ fixed, there is an integer r such that we may consider the space of the first r Fourier coefficients of the coefficient functions $a_{ij}(t)$ of terms $x^i x'^j$ of P with $i+j \leqq N$ for operators of the form* (2.2). *Then, for any such operator f, there is an open dense set of the space of such Fourier coefficients such that the operator is infinitesimally stable at* 0. *Thus, as mappings these operators are locally equivalent near* 0 *to mappings of the form $f_0 \times \mathrm{id}$ where $f_0: \mathbb{R}^k, 0 \to \mathbb{R}^k, 0$ is an infinitesimally stable germ* (*and* id *denotes the identity mapping on some Banach space*).

There is also a partial converse.

THEOREM 2. *Given any infinitesimally stable germ $f_0: \mathbb{R}^k, 0 \to \mathbb{R}^k, 0$ with $\dim_{\mathbb{R}} \ker (df(0)) = 2$, then there is an operator f of the form* (2.2) *such that as a mapping f is locally equivalent* (*near* 0) *to $f_0 \times \mathrm{id}$.*

Thus, the entire theory of Σ_2 stable map germs (i.e. mapping germs f with $\dim \ker df(0) = 2$) is needed to describe the possible structure for operators of the form (2.2).

Let f be such an infinitesimally stable operator of the form (2.2) and let Q denote its local algebra with maximal ideal $\mathfrak{m}$ so that $\mathfrak{m}^l \cdot Q = 0$ and $\dim_{\mathbb{R}} Q = \delta(Q)$. Then we may draw the following conclusions.

THEOREM 3. (1) *There are $q(t) \in C^0_{(2\pi/\lambda)}$ such that there are $\delta(Q)$* (*isolated*) *periodic solutions to* (2.1) *of period* $(2\pi/\lambda)$. *For $q(t)$ in a sufficiently small neighborhood of* 0, *this is the maximum number of solutions close to zero which can occur.*

(2) *These periodic solutions persist under deformations in the following sense*:

Let $P(x, x', t, \nu)$ and $Q(t, \nu)$ be continuous functions, smooth in (x, x', ν) (*respectively ν*) *for $\nu \in B_a(0) \subset E$, a Banach space, and periodic of period $(2\pi/\lambda)$ in t, so that $P(x, x', t, 0) = P(x, x', t)$ and $Q(t, 0) = q(t)$* (*as in* 1). *Then there are $\delta(Q)$ families of periodic solutions $x^{(i)}(t, \nu)$ to the equation*

$$x'' + m^2\lambda^2 x + P(x, x', t, \nu) = Q(t, \nu) \tag{2.3}$$

defined for $\|\nu\| < \varepsilon$ for some ε with $0 < \varepsilon < a$.

(3) *If $H(x, x', t)$ is continuous, smooth in (x, x') and periodic in t of period $(2\pi/\lambda)$, so that*

$$|H(x, x', t)| = O(\|(x, x')\|)^{l+1}$$

where $m^l \cdot Q = 0$ by hypothesis, then there still exists $q_H(t)$ so that

$$x'' + m^2\lambda^2 x + P(x, x', t) + H(x, x', t) = q_H(t) \tag{2.4}$$

has $\delta(Q)$ periodic solutions of period $(2\pi/\lambda)$.

As a corollary of this, we have the existence of polynomial P for which (2.1) has a specific number of periodic solutions.

THEOREM 4. *There is a polynomial P of total degree $2n-2$ in x and x' and a $q(t)$ such that* (2.1) *has n^2 periodic solutions.*

A specific example of a polynomial P with a given number of periodic solutions is given by the following.

Let

$$P(x, x', t) = \sin(\lambda t)(x^2 + x'^2) + \sum_{j=3}^{n+1} p_j(t)x^j \tag{2.5}$$

where

$$p_j(t) = a_j \cos((j-1)\lambda t) + b_j \sin((j-1)\lambda t) + c_j \cos(j\lambda t).$$

Then, as a corollary of Theorem 3, we have the following.

THEOREM 5. *For almost all c_j, and then for specific a_j and b_j, f obtained using* (2.5) *is infinitesimally stable with local algebra $\mathbb{R}[[x, y]]/(x^2 + y^2, x^{n+1})$. There is $q(t)$ which has only nonzero Fourier coefficients of degree $\leqq 2n$ so that with the above P,* (2.1) *has $2n+2$ (isolated) periodic solutions of period $(2\pi/\lambda)$.*

3. Outline of the method and some derivative computations. We begin by outlining the method we use and describing the role that subsequent sections play in carrying out the outline. The first step is to apply the Lyapunov-Schmidt procedure to the operator (2.2). For simplicity, we refer to the case $m = \lambda = 1$ so that $f: C^2_{2\pi} \to C^0_{2\pi}$. The form of Lyapunov-Schmidt we use is slightly different from its standard form. For our operator f, $df(0)$ is Fredholm of index 0 with $K = \ker df(0)$ spanned by $\{\cos(t), \sin(t)\}$, which also spans a subspace complementary to image $(df(0))$. Let W (respectively W') denote the L^2-orthogonal complement to K in $C^2_{2\pi}$ (respectively $C^0_{2\pi}$). Then: $df(0): W \xrightarrow{\sim} W'$. We define $\Phi: C^2_{2\pi} \to C^0_{2\pi}$ by $\Phi(x, u) = x + f(x, u)$ for $(x, u) \in K \oplus W$. By the inverse function theorem, Φ^{-1} is defined in a neighborhood U of 0; and $F = f \circ \Phi^{-1}: U \to C^0_{2\pi}$ has the form $F(x, u) = (\bar{F}(x, u), u)$ for $(x, u) \in K \oplus W'$. Then, we shall refer to this procedure of obtaining F from f as the Lyapunov-Schmidt procedure.

Our goal is to apply the verification criterion (1.3) to F to deduce that, under certain circumstances, F is infinitesimally stable. The results needed to apply the criterion will be developed in the remainder of this section and in §§ 4 and 5. The verification criterion requires us to compute certain derivatives of $\bar{F}$ at 0, and establish certain surjectivity properties. By the rule for derivatives of compositions, to do this we must be able to compute the derivatives of f at 0. The remainder of this section will be concerned with computations of such derivatives. Lemmas 3.10 and 3.15 provide the key surjectivity results (while Corollaries 3.12, 3.13 and Lemma 3.14 are mainly used for examples).

Second, the verification criterion is to be applied to the restriction of F to a finite-dimensional subspace V_m spanned by $\{\cos(rt), \sin(rt)\}_{r=0}^{m}$. We wish to know that terms involving eigenfunctions $\cos(st)$ and $\sin(st)$ for s sufficiently large (say $\geqq M$) will not affect the derivative computations after restriction. This is achieved by the noninterference Lemma 4.1 and its Corollary 4.9. They allow us, for example, to restrict our original f to a mapping from V_M to itself and apply Lyapunov-Schmidt to the finite-dimensional mapping. This allows us to use standard techniques for working with finite-dimensional mappings and to inductively establish the verification procedure by successively adding terms without altering the preceding surjectivity conditions.

Last, in § 5 we prove that adding a "model operator" to any operator f produces one satisfying the verification criterion. This gives a "universal model" for establishing infinitesimal stability (Theorem 5.4 and Corollary 5.9).

Once these results are in place, the proofs of the theorems follow very simply by the arguments in § 6. Sections 7 and 8 contain specific examples already referred to. We now proceed to the derivative computations.

We begin by simplifying notation and assuming $\lambda = 1$. Then, we consider derivatives of nonlinear operators from $C^2_{2\pi}$ to $C^0_{2\pi}$ which are sums of terms of the form

$$(a_1 \cos kt + a_2 \sin kt) \cdot x^i (x')^j. \tag{3.1}$$

Although we will be working with real mappings and derivatives, computations are easier to see conceptually using complex notation. Furthermore, in certain steps we will wish to show that certain real forms span a vector space of multilinear forms. It will be easier to show that the forms span the corresponding complex space of multilinear forms.

We use local coordinates (x_1, x_2) for the subspace spanned by $\cos(t)$ and $\sin(t)$. We let $\omega = \cos(t) + i\sin(t)$ with complex conjugate $\bar{\omega}$. Also, we let $z = x_1 + ix_2$ so that $x_1 \cos t + x_2 \sin t = (1/2)(\bar{z}\omega + z\bar{\omega})$. Similarly, for $x = a_1 + ia_2$, $a_1 \cos kt + a_2 \sin kt = (1/2)(\bar{x}\omega^k + x\bar{\omega}^k)$. Since $d/dt(\omega) = i\omega$ and $d/dt(\bar{\omega}) = -i\bar{\omega}$,

$$\frac{d}{dt}\left(\frac{1}{2}(\bar{z}\omega + z\bar{\omega})\right) = \left(\frac{i}{2}\right)(\bar{z}\omega - z\bar{\omega}).$$

Given a polynomial in ω and $\bar{\omega}$, $Q(\omega, \bar{\omega})$, we introduce the function which gives the constant term of Q,

$$\mathrm{Cst}\,(Q(\omega, \bar{\omega})) = Q(0, 0).$$

Then, we can compute the L^2-component of $Q(\omega, \bar{\omega})$ for $\cos(kt)$ or $\sin(kt)$ by taking L^2-inner products with $(1/\pi)\cos(kt)$ or $(1/\pi)\sin(kt)$. This can alternately be computed as

$$\mathrm{Cst}\,((\omega^k + \bar{\omega}^k) \cdot Q(\omega, \bar{\omega})) \quad \text{or} \quad \mathrm{Cst}\,((-i)(\omega^k - \bar{\omega}^k) \cdot Q(\omega, \bar{\omega})). \tag{3.2}$$

More generally, if we let (u_k, v_k) denote coordinates for the subspace $\langle\cos(kt), \sin(kt)\rangle$, and let $w_k = u_k + iv_k$, then from (3.2),

$$w_k = 2 \cdot \mathrm{Cst}\,(\omega^k \cdot Q(\omega, \bar{\omega})). \tag{3.3}$$

Now consider $f: C^2_{2\pi} \to C^0_{2\pi}$ defined by

$$f(x) = (a_1 \cos(rt) + a_2 \sin(rt)) \cdot x^j (x')^k. \tag{3.4}$$

A standard computation yields

$$\frac{1}{m!}d^m f(0)(l_1, \cdots, l_m)$$

(3.5) $$= \begin{cases} 0, & m < j+k, \\ \frac{1}{m!}(a_1 \cos(rt) + a_2 \sin(rt)) \sum l_{\sigma(1)} \cdots l_{\sigma(j)} l'_{\sigma(j+1)} \cdots l'_{\sigma(k)} & \\ \qquad \text{(summed over } \sigma \in S_m) & \text{if } m = j+k. \end{cases}$$

We will consider several special cases of (3.5).

Case 1. If we evaluate (3.5) for $l_i = (1/2)(\bar{z}\omega + z\bar{\omega})$ for all i we obtain

(3.6) $$w_1 = \text{Cst}\,(\omega \cdot (\bar{\alpha}\omega^r + \alpha\bar{\omega}^r) \cdot (1/2)^m \cdot i^k(\bar{z}\omega + z\bar{\omega})^j(\bar{z}\omega - z\bar{\omega})^k).$$

Case 2. If $j = m$, $k = 0$, $l_1 = \cdots = l_{m-1} = (1/2)(\bar{z}\omega + z\bar{\omega})$ and $l_m = (1/2)(\bar{\beta}\omega^s + \beta\bar{\omega}^s)$ with $\beta = b_1 + ib_2$, then for (3.5)

(3.7) $$w_q = \text{Cst}\,(\omega^q \cdot (\bar{\alpha}\omega^r + \alpha\bar{\omega}^r) \cdot (\bar{\beta}\omega^s + \beta\bar{\omega}^s) \cdot (1/2)^m \cdot (\bar{z}\omega + z\bar{\omega})^{m-1}).$$

Case 3. Last, if $k = 1$, $l_1 = \cdots = l_j = (1/2)(\bar{z}\omega + z\bar{\omega})$ (and $m = j+1$) and $l_{j+1} = (1/2)(\bar{\beta}\omega^s + \beta\bar{\omega}^s)$, then for (3.5)

(3.8) $$w_1 = \text{Cst}\,(\omega \cdot (\bar{\alpha}\omega^r + \alpha\bar{\omega}^r) \cdot \Psi)$$

where

(3.9) $$\Psi = (i/j+1)\{j(\bar{\beta}\omega^s + \beta\bar{\omega}^s)(\bar{z}\omega + z\bar{\omega})^{j-1}(\bar{z}\omega - z\bar{\omega}) + s(\bar{\beta}\omega^s - \beta\bar{\omega}^s)(\bar{z}\omega + z\bar{\omega})^j\}.$$

Let (3.6), with α replaced by $\alpha_{jk}^{(r)}$, be denoted by $\Psi_{jk}^{(r)}$. Also, let $S^m(x_1, x_2)$ denote the vector space of real homogeneous polynomials of degree m in (x_1, x_2). Let $S_{\mathbb{C}}^m(x_1, x_2)$ denote the corresponding complex vector space. The $\Psi_{jk}^{(r)}$ are elements of $S^m(x_1, x_2) \times S^m(x_1, x_2)$ with $m = j+k$.

LEMMA 3.10. *$\{\Psi_{jk}^{(r)}: j+k = m, 0 \leqq r \leqq m\}$ span $S^m(x_1, x_2) \times S^m(x_1, x_2)$.*

Proof. We first observe that $S_{\mathbb{C}}^m(x_1, x_2)$ can be identified with $S^m(x_1, x_2) \times S^m(x_1, x_2)$ by $g + ih \mapsto (g, h)$. Via this identification, $\{z^j\bar{z}^k: j+k = m\}$ is a basis over $\mathbb{C}$ (since $x_1 = (1/2)(z + \bar{z})$, $x_2 = (i/2)(z - \bar{z})$ we may expand $x_1^n x_2^l = (1/2)^{n+l}(-i)^l(z+\bar{z})^n(z-\bar{z})^l$ to represent $x_1^n x_2^l$ as a linear combination of $z^j\bar{z}^k$). Thus viewing the elements $\Psi_{jk}^{(r)}$ as elements of $S_{\mathbb{C}}^m(x_1, x_2)$, it is sufficient to represent each $z^j\bar{z}^k$ as a complex linear combination of such elements. Consider

(3.11) $$\begin{aligned} w_1 &= \text{Cst}\,(\omega \cdot (\bar{\alpha}\omega^r + \alpha\bar{\omega}^r) \cdot (z^j\bar{\omega}^j)(\bar{z}^k\omega^k)) \\ &= \text{Cst}\,((\bar{\alpha}\omega^{r+1} + \alpha\bar{\omega}^{r-1})\omega^{k-j}) \cdot z^j\bar{z}^k. \end{aligned}$$

If $k \geqq j$ and $r = 1 + k - j$, then $w_1 = \alpha z^j\bar{z}^k$. If α represents an arbitrary complex number, we obtain the complex subspace spanned by $z^j\bar{z}^k$. If instead $j > k$, we use $r' = j - k - 1$ and obtain $\bar{\alpha}' z^j\bar{z}^k$. Again if α' is an arbitrary complex number then we obtain the subspace spanned by $z^j\bar{z}^k$.

Now, using in (3.11)

$$\bar{z}\omega = (1/2)(\bar{z}\omega + z\bar{\omega}) - i((i/2)(\bar{z}\omega - z\bar{\omega})),$$

$$z\bar{\omega} = (1/2)(\bar{z}\omega + z\bar{\omega}) + i((i/2)(\bar{z}\omega - z\bar{\omega}))$$

and expanding we obtain $z^j\bar{z}^k$ as a complex linear combination of $\{\Psi_{jk}^{(r)}\}$. □

We obtain the following as corollaries of the method.

COROLLARY 3.12. *In* (3.6) *with $r = m-1$, $j = m$, $k = 0$, $w_1 = (1/2)^m\bar{\alpha}z^m$ modulo terms of the form $(z \cdot \bar{z}) \cdot h$.*

COROLLARY 3.13. *For $f(x)=\sin(t)(x^2+x'^2)$, with $l_i=(1/2)(\bar z\omega+z\bar\omega)$ $i=1,2$, the w_1 component of $(1/2!)d^2f(0)(l_1,l_2)$ equals $iz\bar z$. If $l_2=(1/2)(\bar w_3\omega^3+w_3\bar\omega^3)$, it equals*

$$(1/4)(-i)(\bar z w_3)=\tfrac14(v_3x_2-u_3x_1,\ u_3x_2+v_3x_1).$$

For Case 2 we have the following lemma.

LEMMA 3.14. (1) *When $q=1$ and $r+s>m$ expression* (3.7) *equals*

$$\left(\frac12\right)^m\cdot\left\{\binom{m-1}{a}\alpha\bar\beta\bar z^{m-1-a}z^a+\binom{m-1}{a-1}\bar\alpha\beta z^{m-a}\bar z^{a-1}\right\}$$

if $r-s-1=m-1-2a$ and $0\leqq a\leqq m$; otherwise it equals 0.

(2) *When $r>s>q>m-1$, it equals*

$$\left(\frac12\right)^m\cdot\alpha\bar\beta\binom{m-1}{a}\bar z^{m-1-a}z^a$$

if $r-s-q=m-1-2a$ and $0\leqq a\leqq m-1$; otherwise it is 0 (*note by convention* $\binom{m-1}{m}=\binom{m-1}{-1}=0$).

Proof.

$$\omega^q(\bar\alpha\omega^r+\alpha\bar\omega^r)(\bar\beta\omega^s+\beta\bar\omega^s)=\bar\alpha\bar\beta\omega^{r+s+q}+\alpha\beta\bar\omega^{r+s-q}+\bar\alpha\beta\omega^{r-s+q}+\alpha\bar\beta\bar\omega^{r-s-q}.$$

In the first case, only the last two terms can have exponent $\leqq m-1$ in absolute value. In the second case, only the last term is possible. Thus, if $r-s-q=m-1-2a$, then (2) is as stated and we have the extra term if $q=1$. □

LEMMA 3.15. *Expression* (3.8), *with $r>j+1$, $s\geqq0$, is equal to*

$$\left(\frac{i}{j+1}\right)\left[-\bar\alpha\beta\left\{(s-j)\binom{j}{a}+(s+j)\binom{j}{a-1}\right\}\bar z^{j-a}z^a\right.$$
$$\left.+\alpha\bar\beta\left\{(s-j)\binom{j}{a-1}+(s+j)\binom{j}{a-2}\right\}z^{j+1-a}\bar z^{a-1}\right]\tag{3.16}$$

if $s=r+j+1-2a$ and $0\leqq a\leqq j$;

and equals 0 *if $s(\geqq0)$ is not of the above form.*

Proof. We first compute Ψ in (3.8):

$$\Psi=\left(\frac{i}{j+1}\right)(\bar z\omega+z\bar\omega)^{j-1}\{j(\bar\beta\omega^s+\beta\bar\omega^s)(\bar z\omega-z\bar\omega)+s(\bar\beta\omega^s-\beta\bar\omega^s)(\bar z\omega+z\bar\omega)\}.$$

Thus,

$$\omega(\bar\alpha\omega^r+\alpha\bar\omega^r)\cdot\Psi=\left(\frac{i}{j+1}\right)(\bar z\omega+z\bar\omega)^{j-1}\cdot\Psi_1\tag{3.17}$$

where a computation yields

$$\Psi_1=(s+j)(\bar\alpha\bar\beta\bar z\omega^{r+s+2}-\alpha\beta z\bar\omega^{r+s}+\alpha\bar\beta\bar z\omega^{s+2-r}-\beta\bar\alpha z\bar\omega^{s-r})$$
$$+(s-j)(\bar\alpha\bar\beta z\omega^{r+s}-\alpha\beta\bar z\bar\omega^{r+s-2}+\alpha\bar\beta z\omega^{s-r}-\bar\alpha\beta\bar z\bar\omega^{s-r-2}).$$

From (3.17), the formula for Ψ_1 yields, for (3.8) with $s=r+j+1-2a$, the expression

$$-\bar\alpha\beta\left\{(s-j)\binom{j}{a}\bar z\cdot\bar z^{j-1-a}z^a+(s+j)\cdot\binom{j}{a-1}z\bar z^{j-1-(a-1)}\cdot z^{a-1}\right\}$$
$$+\alpha\bar\beta\left\{(s-j)\binom{j}{a-1}\cdot z\cdot z^{j-1-(a-1)}\bar z^{a-1}+(s+j)\binom{j}{a-2}\cdot\bar z\cdot z^{j-1-(a-2)}\bar z^{a-2}\right\}$$

(where $\binom{j}{c}=0$ if $c<0$). This gives the desired formula upon rearrangement. If s is not of this form, then (3.17) is without constant terms and so (3.8) is zero. □

4. A noninterference lemma. We consider an operator f obtained by adding to $x''+x$ a finite number of terms of the form $(a \sin kt + b \cos kt)\cdot x^i(x')^j$ with $i+j>1$ (and say $i+j \leqq n$). Let

$$V_m = \text{vector space spanned by } \{\cos(kt), \sin(kt), 0 \leqq k \leqq m\}.$$

Then, we shall show that given $m>0$, there is an $M>0$ so that the n-jet of the germ obtained by applying the Lyapunov-Schmidt procedure to f and restricting to V_m does not depend on the Fourier coefficients of order $\geqq M$ of the image of the original f. In essence, the higher eigenfunctions do not interfere with the nonlinear behavior of the Lyapunov-Schmidt form of f on V_m.

More specifically, let $\Phi: C^2_{2\pi} \to C^0_{2\pi}$ be defined by $\Phi(x, u) = x + f(x, u)$ for $x \in K = \ker(df(0))$ and u belonging to the L^2-orthogonal complement in $C^2_{2\pi}$. Then, $f\circ\Phi^{-1}: C^0_{2\pi} \to C^0_{2\pi}$ gives the Lyapunov-Schmidt reduction. We let W_m denote the L^2-orthogonal complement to V_m in $C^0_{2\pi}$, and P_m denote the projection onto V_m along W_m. Define $f_2 = P_M \circ f \mid V_M$ (with M to be determined). Also, let $\Phi': V_M \to V_M$ be defined by $\Phi' = P_M \circ \Phi \mid V_M$ so that $\Phi'(x, u') = x + f_2(x, u')$. Then, $f_2 \circ \Phi'^{-1}: V_M \to V_M$ is the Lyapunov-Schmidt reduction for f_2. Finally, we let $f_1 = f_2 \circ \Phi'^{-1} \mid V_m : V_m \to V_m$ (here $m<M$). Note $f_1(V_m) \subseteq V_m$ by Lyapunov-Schmidt. Then, the n-jets of f_1 and $f\circ\Phi^{-1}$ are related by the following lemma.

LEMMA 4.1. *Given $m>0$, there is an integer $M>m$ such that for $1 \leqq j \leqq n$*

$$\text{(i)} \qquad d^j(f\circ\Phi^{-1})(0) \mid V_m \times\cdots\times V_m = d^j f_1(0),$$

$$\text{(ii)} \qquad d^j(P_m \circ f \circ \Phi^{-1})(0) \mid V_m \times\cdots\times V_m \times W_M \equiv 0.$$

Proof. First we prove (i). Let

$$K = \max\{k: \cos(kt) \text{ or } \sin(kt) \text{ appears as a coefficient of a term of } f\}.$$

We inductively define a sequence of integers by $m_1 = m$, and $m_j > (n+1)m_j + K$ for $1 < j \leqq n+1$ and let $M = m_{n+1}$. We claim this M will suffice for (i).

Let Q_j denote the projection onto W_{m_j} along V_{m_j} and let P_{m_j} be denoted just by P_j. Then, we first observe

$$Q_{j+1} \circ d^j f(0) \mid V_{m_j} \times\cdots\times V_{m_j} \equiv 0. \tag{4.2}$$

This follows from (3.5) since to have a component of $\cos(st)$ or $\sin(st)$ for $d^j f(0)$, we must have $s \leqq j\cdot m_j + K < m_{j+1}$ with $j \leqq n$.

Next, temporarily let P_1' denote the projection onto $\langle\cos(t), \sin(t)\rangle$ along its L^2 orthogonal complement in $C^2_{2\pi}$. Then, $\Phi = P_1' + f$ so

$$d^j\Phi(0) = \begin{cases} P_1' + df(0), & j=1, \\ d^j f(0), & j>1. \end{cases}$$

Thus,

$$Q_{j+1}\circ d^j\Phi(0) \mid V_{m_j}\times\cdots\times V_{m_j} = \begin{cases} (Q_{j+1}\circ P_1') + (Q_{j+1}\circ df(0) \mid V_{m_1}), & j=1, \\ Q_{j+1}\circ d^j f(0) \mid V_{m_j}\times\cdots\times V_{m_j}, & j>1 \end{cases}$$

or

$$Q_{j+1}\circ d^j\Phi(0) \mid V_{m_j}\times\cdots\times V_{m_j} = 0, \qquad j \geqq 1.$$

Thus,

$$d^j\Phi(0)\,|\,V_{m_j}\times\cdots\times V_{m_j}=P_{j+1}\circ d^j\Phi(0)\,|\,V_{m_j}\times\cdots\times V_{m_j} \tag{4.3}$$
$$=d^j\Phi'(0)\,|\,V_{m_j}\times\cdots\times V_{m_j},\qquad j\geqq 1$$

and

$$d^j\Phi(0)(V_{m_j}\times\cdots\times V_{m_j})\subset V_{m_{j+1}},\qquad j\leqq n. \tag{4.4}$$

Let L^l denote the l-multilinear operator.

$$L^l\overset{\text{def}}{=}(d\Phi(0))^{-1}\circ d^l\Phi(0).$$

By an expression in L of depth r we mean a term of the form L^l if $r=1$, while for $r>1$ we mean an expression (i.e. a composition) $L^s(L^{j_1},\cdots,L^{j_s})$ where each L^{j_i} is an expression of depth $r-1$. We claim that if $\Psi(L)$ is an expression in L of depth r then

$$\text{Image}\,(\Psi(L)\,|\,V_{m_1}\times\cdots\times V_{m_1})\subset V_{m_r}.$$

This follows by induction using (4.4) and the fact that $d\Phi(0)$ preserves the V_{m_j}. Furthermore, by another induction argument, if we replace L^l by

$$L'^l=(d\Phi'(0))^{-1}\circ d^l\Phi'(0)$$

then

$$\Psi(L)\,|\,V_{m_1}\times\cdots\times V_{m_1}=\Psi(L'). \tag{4.5}$$

Now,

$$d^lf_1(0)=d^l(f_2\circ\Phi'^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1}$$
$$=P_1\circ d^l(f\circ\ \Phi'^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1}.$$

We finally claim that

$$d^l(f\circ\Phi'^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1}=d^l(f\circ\Phi^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1}. \tag{4.6}$$

Then by the Lyapunov-Schmidt procedure, $Q_2(f\circ\Phi^{-1})=Q_2$; hence we may compose the left-hand side of (4.6) with P_1 and equality remains; the first result will follow. For this final claim we proceed in two steps. First, by the product rule $d^l(f\circ\ \Phi^{-1})(0)$ is a sum of terms of the form

$$d^rf(0)(d^{l_1}\Phi^{-1}(0),\cdots,d^{l_r}\Phi^{-1}(0))$$

with coefficients only depending upon r, $l_1,\cdots,l_r$ (see e.g. [F] or [R]).

Thus, if

$$d^l(\Phi^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1}=d^l(\Phi'^{-1})(0)\,|\,V_{m_1}\times\cdots\times V_{m_1} \tag{4.7}$$

and both have images in V_{m_l}, then by the product rule and (4.2) we have equality in (4.6).

Lastly, to establish (4.7) we again apply the product rule to

$$\Phi\circ\Phi^{-1}=\text{id}\quad\text{or}\quad\Phi_1\circ\Phi_1^{-1}=\text{id}_{V_{m_{n+1}}}$$

to obtain that $d^l\Phi^{-1}(0)$ is a sum of terms with well-determined coefficients which are expressions in L, as defined above. A similar statement applies to $d^l\Phi'^{-1}(0)$ and L'. Thus by (4.5), (4.7) is valid.

For (ii), we let the M from (i) be denoted by m' and reapply (i) for m' in place of m, obtaining $m'_j > (n+1)m'_{j-1} + K$ and $M' = m'_{n+1}$. This implies $m'_j - (n \cdot m'_1 + K) > m'_{j-1}$. By the product rule, to prove (ii), it is sufficient to prove that the terms

$$P_1 \circ d^r f(0)(d^{l_1}(\Phi^{-1})(0), \cdots, d^{l_r}(\Phi^{-1})(0)) \,|\, V_{m_1} \times \cdots \times V_{m_1} \times W_{M'}$$

are identically zero (recall $P_1 \overset{\text{def}}{=} P_{m_1}$). Since

$$d^l(\Phi^{-1})(0)(V_{m_1} \times \cdots \times V_{m_1}) \subset V_{m_l} \subset V_{m'}, \qquad l \leqq n,$$

it is sufficient to show that

$$(4.8) \qquad d^l(\Phi^{-1})(0)(V_{m'} \times \cdots \times V_{m'} \times W_{M'}) \subset W_{m'_2}$$

since then

$$d^r f(0)(V_{m'} \times \cdots \times V_{m'} \times W_{m'_2}) \subset W_{m'_1}$$

by the above inequality for m'_j. However, as we already mentioned, $d^l(\Phi^{-1})(0)$ is a sum of terms each of which is an expression in L of depth $l \leqq n$. At each stage, the image will be one lower $W_{m'_j}$; hence, each term will be in $W_{m'_2}$. Thus, so will (4.8). □

As a corollary of the method of proof we have the following additional result.

COROLLARY 4.9. *Given f, M, and m as in Lemma* 4.1, *let*

$$g(x, x', t) = \sum b_{ij}(t) x^i x'^j$$

summed for $i + j = k$ (with fixed $k \leqq n$) with $b_{ij}(t) \in W_M$. Let g_1 be obtained from $f + g$ by applying Lyapunov-Schmidt; then

(i) $d^j(f \circ \Phi^{-1})(0) \,|\, V_m \times \cdots \times V_m = d^j g_1(0) \,|\, V_m \times \cdots \times V_m, \qquad 1 \leqq j \leqq k,$

(ii) $P'_1 \circ d^k g(0) \circ d\Phi^{-1}(0) \,|\, V_m \times \cdots \times V_m \times W_M = d^k g_1(0) \,|\, V_m \times \cdots \times V_m \times W_M.$

Remark. In fact, a slightly more careful analysis allows us to conclude (i) for $1 \leqq j \leqq 2k - 2$.

5. A universal model. Consider a family of operators of the form

$$f(x, \nu) = x'' + \lambda^2 x + P(x, x', \nu, t)$$

with $\nu \in E$, and $\|P\| = O(\|x\|^2 + \|x'\|^2)$ (with P continuous, smooth in (x, x', ν), and periodic in t of period $2\pi/\lambda$). We may simultaneously apply Lyapunov-Schmidt and obtain a family which we may further restrict to some V_m.

$$F: U \to V_m \times E, \qquad F(x, u, \nu) = (\bar{F}(x, u, \nu), u, \nu)$$

for U an open neighborhood of 0 in $V_m \times E$ with $x \in \langle \sin(t), \cos(t) \rangle$, and u belonging to the orthogonal complement of $\langle \sin(t), \cos(t) \rangle$ in V_m. Looking at the nth order Taylor expansion of F with respect to x, we have in local coordinates $x = (x_1, x_2)(x^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2}$ for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2))$

$$\bar{F}(x, u, \nu) = \left(\sum_{1 \leqq |\boldsymbol{\alpha}| \leqq n} v_{\boldsymbol{\alpha}} x^{\boldsymbol{\alpha}}, \sum_{1 \leqq |\boldsymbol{\alpha}| \leqq n} v'_{\boldsymbol{\alpha}} x^{\boldsymbol{\alpha}} \right) + H \quad \text{with } \|H\| = O(\|x\|^{n+1} + \|x'\|^{n+1})$$

and $v_{\boldsymbol{\alpha}}$, $v'_{\boldsymbol{\alpha}}$ functions of (u, ν). If we let

$$\{v_{\boldsymbol{\alpha}}, v'_{\boldsymbol{\alpha}}: 1 \leqq |\boldsymbol{\alpha}| \leqq n\}$$

denote coordinates for a space T, this defines a mapping $v: U' \to T$ sending $(u, \nu) \mapsto (v_{\boldsymbol{\alpha}}(u, \nu), v'_{\boldsymbol{\alpha}}(u, \nu))$: $1 \leqq |\boldsymbol{\alpha}| \leqq n$. Here $U' = U_1 \times U_2$ is a neighborhood of 0 in $\langle \sin(t), \cos(t) \rangle^{\perp} \oplus E$. We let T_1 denote the subspace of T defined by the vanishing of

$\{v_\alpha, v'_\alpha: |\alpha|=1\}$. Given an operator $f_0(x)=x''+\lambda^2x+P(x,x',t)$ we consider a family of operators of the form

$$f(x,\nu)=f_0(x)+G(x,\nu) \tag{5.1}$$

where $G(x,\nu)=G_1+G_2$ with

$$G_1=\sum_{\substack{1\leqq k\leqq i+j\\ 2\leqq i+j\leqq n}}(a_{ij}^{(k)}\cos(kt)+b_{ij}^{(k)}\sin(kt))x^ix'^j \tag{5.2}$$

and

$$G_2=\sum_{i=2}^{n+1}(c_i\sin(r_it)+d_i\cos(r_it)+c'_i\sin(r'_it)+d'_i\cos(r'_it))x^{i-1}x'. \tag{5.3}$$

Let $\nu=(\nu_1,\nu_2)$, where ν_1 denotes the set of all coefficients $\{a_{ij}^{(k)}, b_{ij}^{(k)}\}$ and ν_2 denotes the set of all coefficients $\{c_i, d_i, c'_i, d'_i\}$. Let E_i denote the subspace with local coordinates ν_i and $E=E_1\oplus E_2$.

THEOREM 5.4. *There are integers r_i, r'_i such that for any f_0:*

(i) *For any $\nu_2\in E_2$, $v|(U_2\cap E_1): U_2\cap E_1\to T_1$ is a submersion at* 0.

(ii) *For any $\nu_1\in E_1$ and almost all $\nu_2\in E_2$ $v|U_1: U_1\to T$ is a submersion at* 0.

Proof. We repeatedly apply Lemma 4.1 $2n-1$ times beginning with $m_0=n+1$ and at the jth step increasing $K=m_{j+1}$ to obtain $m_0, m_1, \cdots, m_{2n-1}$. We choose r_i and r'_i so that $|r_i-r'_i|>4(i+1)$ and $m_0<m_1<r_1$, $r'_1<m_2<m_3<r_2$, $r'_2<m_4\cdots<m_{2n-1}<r_{n+1}$, r'_{n+1}. For (i), we first observe that the only derivative of f in which the $a_{ij}^{(k)}$ or $b_{ij}^{(k)}$ appear is the $i+j$th derivative. Thus, for $d^l\bar{F}(0)=P'\circ d^l(f\circ\Phi^{-1})$, we obtain from the formula for the derivative of compositions that the coefficients $a_{ij}^{(k)}$ and $b_{ij}^{(k)}$ do not appear in $d^l\bar{F}(0)$ if $l<i+j$. If $l=i+j$, then they only appear in the term

$$P'\circ d^lf\circ d\Phi^{-1}(=P'\circ d^lf \text{ when restricted to } \langle\sin(t),\cos(t)\rangle)$$

(recall P' denotes projection onto $\langle\sin(t),\cos(t)\rangle$). We remark that they also appear in $df\circ d^l(\Phi^{-1})$; however composing with P' annihilates this term.

Hence, evaluating (3.6) we obtain, with $\alpha_{jk}^{(l)}=a_{jk}^{(l)}+ib_{jk}^{(l)}$, that the $i+j$th-order terms of $\bar{F}$ involving $a_{ij}^{(k)}$ and $b_{ij}^{(k)}$ equal $\Psi_{ij}^{(k)}$, in the notation of Lemma 3.10. Thus, by Lemma 3.10 and the preceding discussion, the derivatives

$$\left.\frac{\partial\bar{F}}{\partial a_{ij}^{(k)}}\right|_{\nu_1=0}, \left.\frac{\partial\bar{F}}{\partial b_{ij}^{(k)}}\right|_{\nu_1=0} : i+j=s, \qquad 1\leqq k\leqq i+j$$

span $S^s(x_1,x_2)\times S^s(x_1,x_2)$ modulo higher order terms. The union of such derivatives for $2\leqq s\leqq n$, span T_1 so (i) holds.

For (ii), we let $\alpha_j=d_j+ic_j$, $\alpha'_j=d'_j+ic'_j$, and $w_s=u_s+iv_s$, where $s=r_j+j+1-2a$, $0\leqq a\leqq j$, or a corresponding formula for s' with r_j replaced by r'_j, $2\leqq j\leqq n$. We also let U_0 denote $U_1\cap V'$, where V' is the subspace spanned by the eigenfunctions $\cos(s''t)$, $\sin(s''t)$ for $s''=s$ or s' satisfying the preceding conditions. Then we shall show that (ii) holds even for the subset U_0. It is sufficient to show that

$$\left.\frac{\partial v}{\partial u_{s''}}\right|_{\langle\sin t,\cos t\rangle}, \quad \left.\frac{\partial v}{\partial v_{s''}}\right|_{\langle\sin t,\cos t\rangle} \tag{5.5}$$

span T for almost all choices of coefficients $\{\alpha_j, \alpha'_j\}$, where we undertand that in (5.5) s'' runs through all allowable values. By the way that the r_j and r'_j were chosen, the w_s and $w_{s'}$ are all distinct as j and a range over the allowable values just given. We first show that the lowest-order terms of the derivatives involving u_s and v_s (or $u_{s'}$, $v_{s'}$)

span T. We apply Lemma 3.15, with α replaced by α_j and α_j' and β replaced by w_s or $w_{s'}$. It is sufficient to show that (3.16) with $\alpha\bar{\beta}$ respectively real and purely imaginary, and for both s and s', spans a 4-dimensional real subspace to obtain both $\bar{z}^{j-a}z^a$ and $z^{j+1-a}\bar{z}^{a-1}$ in both coordinates. By subtracting versions of (3.16) for $\alpha\bar{\beta}$ real and s (respectively s') we obtain a real multiple of

$$(5.6)\qquad i\cdot(s-s')\left\{-\left(\binom{j}{a}+\binom{j}{a-1}\right)\bar{z}^{j-a}z^a+\left(\binom{j}{a-1}+\binom{j}{a-2}\right)z^{j+1-a}\bar{z}^{a-1}\right\}$$

or equivalently,

$$i\cdot(r_j-r_j')\left\{-\binom{j+1}{a}\bar{z}^{j-a}z^a+\binom{j+1}{a-1}z^{j+1-a}\bar{z}^{a-1}\right\}.$$

Subtracting a multiple of (5.6) from (3.16), we obtain a multiple of

$$(5.7)\qquad i\left[-\left\{-\binom{j}{a}+\binom{j}{a-1}\right\}\bar{z}^{j-a}z^a+\left\{-\binom{j}{a-1}+\binom{j}{a-2}\right\}z^{j-1-a}\bar{z}^{a-1}\right].$$

Then, adding and subtracting multiples of (5.6) and (5.7) yields multiples of

$$i\left(\binom{j}{a-1}\bar{z}^{j-a}z^a+\binom{j}{a-2}z^{j+1-a}\bar{z}^{a-1}\right)\quad\text{and}\quad i\left(\binom{j}{a}\bar{z}^{j-a}z^a-\binom{j}{a-1}z^{j+1-a}\bar{z}^{a-1}\right),$$

which are always linearly independent if $a \geqq 1$ (or they give just $i\binom{j}{0}\bar{z}^j$ if $a=0$). For $\alpha\bar{\beta}$ purely imaginary we obtain real multiples instead. This gives the claim. These elements will, in general, generate the four-dimensional subspace for a Zariski open subset of α_j, α_j', which we have just seen is nonempty.

Now we proceed to show inductively that the elements of (5.5) corresponding to r_j, r_j', $2 \leqq j \leqq k-1$ and allowable a, span T modulo the space T_k defined by the vanishing of $\{v_{\boldsymbol{\alpha}}, v_{\boldsymbol{\alpha}}' \colon |\boldsymbol{\alpha}| \leqq k\}$. This is true for T modulo T_1 by the preceding discussion since the elements of (5.5) for r_2 and r_2' map to a basis for T modulo T_1; and so adding fixed linear terms to them will not change this for a generic choice of coefficients c_2, d_2, c_2', d_2'. If now the result is true for $k<l$, then the terms of (5.5) corresponding to r_{l+1} and r_{l+1}' are of the form $\varphi_i+\varphi_i'+\varphi_i''$, where φ_i have zero coefficients $v_{\boldsymbol{\alpha}}$, $v_{\boldsymbol{\alpha}}'$ for $|\boldsymbol{\alpha}| \geqq l$, φ_i' only has nonzero coefficients for $|\boldsymbol{\alpha}|=l$, and $\varphi_i'' \in T_l$. By the earlier discussion the φ_i' span T_{l-1}/T_l. By induction, we may subtract off constant multiples of the terms given by the induction hypothesis to obtain $\varphi_i=0$. Then, in terms of a basis for T_{l-1}/T_l, the φ_i' have the matrix representation

$$(5.8)\qquad B+c_{l+1}A_1+d_{l+1}A_2+c_{l+1}'A_3+d_{l+1}'A_4$$

where B and A_i are fixed with generic sums of the A_i being nonsingular; thus a determinant argument implies (5.8) is nonsingular for almost all choices of c_{l+1}, etc. This completes the induction step. After n steps, we obtain the conclusion for (ii).

As a corollary, we have the following.

COROLLARY 5.9. *Let $P=0$ with the preceding notation. If W is a neighborhood of 0 in E, then there exists a neighborhood V of 0 in T_1 such that for every $h \in V$, there exists $\nu=(\nu_1,\nu_2) \in W$ so that $v(0,\nu)=h$ and $v\,|\,U_1$ is a submersion at 0.*

Proof. We prove the result by induction on n. It is vacuously true if $n=1$. Suppose it is true for $n-1$. We let $v^{(n-1)}$, $E_1^{(n-1)}$, $E_2^{(n-1)}$ denote the corresponding objects for $n-1$. Then, $E_i=E_i^{(n-1)}\oplus E_i'$, where E_i' denotes the subspace corresponding to $\{a_{ij}^{(k)}, b_{ij}^{(k)} \colon i+j=n\}$ if $i=1$ or $\{c_{n+1}, c_{n+1}', d_{n+1}, d_{n+1}'\}$ if $i=2$. Given $W \subset E_1 \oplus E_2$ we may find $W_i \subset E_i$ so that $W_1 \times W_2 \subset W$. We may further assume $W_i = W_i' \times W_i''$ according to the above decompositions for E_i. We also decompose $\nu \in W$, $\nu = (\nu_1, \nu_2)$ with $\nu_i = (\nu_i', \nu_i'')$.

Observe that $v(W_1'')$ is a neighborhood of 0 in T_{n-1} (this follows from the proof of (i) in Theorem 5.4). We choose a neighborhood V_1 of 0 in T_{n-1} so that if $h_1, h_2 \in V_1$, then $h_1 - h_2 \in v(W_1'')$. Then by continuity there exists a neighborhood Z of 0 in E such that the following composition maps into V_1

$$Z \xrightarrow{v} T_1 \xrightarrow{\text{pr}} T_{n-1}$$

(here pr is projection onto the nth derivative, which is well defined since we have fixed coordinates $\mathbf{x}$).

Now, if $Z = Z_1 \times Z_2$ and $Z_i = Z_i' \times Z_i''$ as above, then by the inductive assumption applied to $Z_1' \times Z_2'$, there is a V' satisfying the result. Here $V' \subset T_1/T_{n-1}$. We let $V = V' \times V_1$ (using the splitting of $T_1 = T_1/T_{n-1} \oplus T_{n-1}$ given by the fixed coordinates).

We claim V satisfies the result. Given $h \in V$, $h = h_{n-1} + h_n$ via the splitting. By induction, there exist $\nu_{n-1} \in Z_1' \times Z_2'$ so that $v^{(n-1)}(0, \nu_{n-1}) = h_{n-1}$ and $v^{(n-1)} \mid U_1^{(n-1)}$ is a submersion at 0. Then, $v(0, \nu_{n-1}) = h_{n-1} + w_n$ where $w_n \in V_1$. Thus, $h_n - w_n \in v(W_1'')$; hence, $h_n - w_n = v(\nu_1'')$, for some $\nu_1'' \in W_1''$. Let $\nu_n^{(1)} = (\nu_{n-1}, (\nu_1'', 0))$.

Then, by the proof of (i) in Theorem 5.4,

$$\begin{aligned} v(0, \nu_n^{(1)}) &= v(0, \nu_{n-1}) + v(0, \nu_1'') \\ &= h_{n-1} + w_n + h_{n-1} - w_n = h. \end{aligned}$$

Also, by Corollary 4.9,

$$d^n\bar{F}(x, u, \nu_n^{(1)})(0) \big|_{V_{m_0} \times \cdots \times V_{m_0} \times V_{m_{2n-2}}} = d^n\bar{F}(x, u, \nu_{n-1})(0) \big|_{V_{m_0} \times \cdots \times V_{m_0} \times V_{m_{2n-2}}},$$

and

$$d^j\bar{F}(x, u, \nu_n^{(1)})(0) = d^j\bar{F}(x, u, \nu_{n-1})(0), \qquad 1 \leqq j \leqq n-1.$$

Thus, $v \mid U_1^{(n-1)}$ is still a submersion at 0.

Finally, let $\nu_n = (\nu_{n-1}, (\nu_1'', \nu_2''))$. By the conditions on r_{n+1} and r_{n+1}' and Corollary 4.9, $v(0, \nu_n) = v(0, \nu_n^{(1)}) = h$. Also, by Corollary 4.9,

$$\begin{aligned} d^{n+1}\bar{F}(x, u, \nu_n)(0) \big|_{V_{m_0} \times \cdots \times V_{m_0} \times W_{m_{2n-1}}} &= d^{n+1}\bar{F}(x, 0, \nu_2'')(0) \big|_{V_{m_0} \times \cdots \times W_{m_{2n-1}}}, \\ d^j\bar{F}(x, u, \nu_n)(0) \big|_{V_{m_{2n-2}}} &= d^j\bar{F}(x, u, \nu_n^{(1)})(0) \big|_{V_{m_{2n-2}}}, \qquad 1 \leqq j \leqq n. \end{aligned}$$

Thus, by the proof of (ii) of Theorem 5.4, for almost all ν_2'', $v \mid U_1^{(n)}$ is a submersion at 0. Picking such a $\nu_2'' \in Z_2''$ completes the proof. □

6. Proofs of Theorems 1–4. For Theorem 1, it is sufficient to prove local openness and density near any specific operator f_0. Given N there is an $r > r_i$ given in Theorem 5.4. First we prove density. Consider $f = f_0 + G$ as in that theorem.

We claim that we may choose ν sufficiently small so that (ii) of Theorem 5.4 holds and $\mathfrak{m}^{N+1} \cdot Q(f) = 0$. Then, condition (ii) implies that f is infinitesimally stable by the verification criterion.

To establish the claim, we first apply (ii) of Theorem 5.4 with $\nu_1 = 0$ to obtain that $v \mid U_1$ is a submersion at 0. Now we let our original f_0 be replaced by $f(x, \nu_2)$ and still denote it by f_0. For this new f_0 we apply Theorem 5.4(i) to conclude that $v \mid U_2 \cap E_1$ is a submersion at 0. Thus, all k-jets sufficiently close to $v(0)$ are in the image of v. Furthermore, for small enough ν_1, $v \mid U_1$ remains a submersion at 0. Then, by the classification of algebras in [M6], there are arbitrarily small deformations of f_0 so that $I(\bar{F}(x, 0, \nu_1)) \supset \mathfrak{m}_x^{N+1}$ for $N \geqq 2$. Thus, we may pick a sufficiently small ν_1 to achieve this. For $\nu = (\nu_1, \nu_2)$ we have our result. □

This proof works for $m=1$. However, for $m>1$, we can consider $C^k_{2\pi/m\lambda}\subset C^k_{2\pi/\lambda}$. The argument just given works for $C^k_{2\pi/m\lambda}$. Then, we may consider the operator for $C^k_{2\pi/\lambda}$ as an extension of the operator for $C^k_{2\pi/m\lambda}$, since all coefficient functions constructed using Theorem 5.4 would be periodic of period $2\pi/m\lambda$. Then, the proof of Theorem 5.4 still follows in exactly the same way since ultimately the proof reduces to considering the restriction to $V_{m_1}\subset C^0_{2\pi/m\lambda}$, and the same arguments apply.

For openness, we consider $f=f_0+H$, where

$$H=\sum_{\substack{2\leqq i+j\leqq N\\ 0\leqq k\leqq r}}(a^{(k)}_{ij}\cos(kt)+b^{(k)}_{ij}\sin(kt))x^i x'^j.$$

We apply Lyapunov-Schmidt to

$$F: C^2_{2\pi/\lambda}\times T\to C^0_{2\pi/\lambda}\times T,\qquad (x(t),\nu)\to(f(x,\nu),\nu)$$

where ν denotes the Fourier coefficients $\{a^{(k)}_{ij}, b^{(k)}_{ij}\}_{i,j,k}$.

This gives $F': U\to C^0_{2\pi/\lambda}\times T$ where U is a neighborhood of 0 in $C^0_{2\pi/\lambda}\times T$. We restrict F' to $U\cap(V_{m_1}\times T)$. The coefficients of $x^\alpha u^\beta v^\gamma$ in the expression for F' are smooth functions in the Fourier coefficients. Then, the n-jet of F' will define an infinitesimally stable germ if it belongs to the open subset St^n of the set of all stable n-jets (see [M5]). Thus, the set of Fourier coefficients for which the n-jet of F' defines an infinitesimally stable germ will be the inverse image of this set and hence (locally) open in a neighborhood of 0 in T. The union of such sets for all n is again open. □

For Theorem 2, we use the universal model. To obtain a given infinitesimally stable germ f_0 up to equivalence, it is sufficient to construct one with a local algebra isomorphic to $Q(f_0)$. As f_0 is infinitesimally stable, it is finitely $\mathscr{K}$-determined (see [M4]) hence $\dim_{\mathbb{R}} Q(f_0)<\infty$ [M4]. Thus, if $Q(f_0)$ has maximal ideal $\mathscr{m}$, there is an l with $\mathscr{m}^l\cdot Q(f_0)=0$. It then follows using Nakayama's lemma, that for a germ f_1 to have $Q(f_1)\xrightarrow{\sim}Q(f_0)$ it is sufficient that their l-jets lie in the same $\mathscr{K}^l$-orbit (see [M4]). However, any $\mathscr{K}^l$-orbit contains 0 in its closure. Thus, we may apply Corollary 5.9 to obtain ν yielding an l-jet in the $\mathscr{K}^l$-orbit for which $v\,|\,U_1$ is a submersion at 0. The operator so constructed has local algebra isomorphic to $Q(f_0)$ by the preceding discussion; and furthermore, since $v\,|\,U_1$ is a submersion, the F obtained after Lyapunov-Schmidt satisfies the verification criterion (since $I(\bar{F}(x,0,\nu))\subset\mathscr{m}^l_x$). Thus, F is infinitesimally stable.

Theorem 3 is an immediate consequence of Corollary 1.2 and Theorem 1.10 applied to infinitesimally stable operators of the form (2.2).

Last, Theorem 4 follows because for $N=2n-2$ it is possible to construct an infinitesimally stable operator f with local algebra $\xrightarrow{\sim}\mathbb{R}[[x,y]]/(x^n,y^n)$, which has $\dim=n^2$. Then, Theorem 3 gives the desired result. □

7. Proof of Theorem 5. Again as in earlier proofs we assume $\lambda=1$ to slightly simplify notation. If f denotes the operator (2.2) with $(m=1,\lambda=1)$ P as in (2.5), we wish to examine the restriction $P_{2n}\circ f_1\,|\,V_{2n}$ with P_{2n} denoting projection onto V_{2n} along its L^2-orthogonal complement in $C^0_{2\pi}$, and f_1 denoting f after Lyapunov-Schmidt is applied. By the noninterference lemma, there is an $m>2n$ so that if g denotes $P_{2n}\circ f'_1\,|\,V_{2n}$, where f'_1 is obtained by applying Lyapunov-Schmidt to $P_m\circ f\,|\,V_m$, then g and $P_{2n}\circ f_1\,|\,V_{2n}$ have the same $n+1$th order Taylor expansions. Thus, it is sufficient to replace f by $f'=P_m\circ f\,|\,V_m$.

Thus, we are considering a smooth mapping between finite dimensional subspaces. We use the following notation. We let $\mathscr{C}_{x,w}$ denote the algebra of real-valued smooth germs on V_m at 0 (here recall $x=(x_1,x_2)$ and $w_j=u_j+iv_j$ as in § 3). Also, $\mathscr{C}_x$ denotes the algebra of germs only depending on (x_1, x_2), and $\mathscr{C}_w$ denotes the algebra of germs only depending on $\{(u_j, v_j): 2 \leqq j \leqq m\} \cup \{u_0\}$. These algebras have maximal ideals of germs vanishing at 0, denoted by $\mathfrak{m}_{x,w}$, $\mathfrak{m}_x$, and $\mathfrak{m}_w$. Then, Theorem 5 will follow from two lemmas.

LEMMA 7.1. *The mapping f' is given in local coordinates by*

$$f'(x, u, v) = \left(x_1^2 + x_2^2 + h,\ P_1(x_1) + x_2 P_2(x_1) + \sum_{i=2}^{2n} (c_i u_i \varphi_i + c_i' v_i \varphi_i') + H_0,\ q\right), \tag{7.2}$$
$$q = (q_0, q_3, \cdots, q_{2m})$$

where

(i) $$h = (1/4)(-u_3 x_1 + v_3 x_2) + h_1(x, u, v),$$
$$H_0 = (1/4)(u_3 x_2 + v_3 x_1) + H_0',$$

(ii) *both h_1 and H_0' have terms of degree $\geqq 3$ and h_1 consists of terms at least quadratic in* **x**,

(iii) $$H_0' \in (x_1^2 + x_2^2) \cdot \mathscr{C}_{x,w} + (\mathfrak{m}_w^2 + \mathfrak{m}_x^{n+2}) \cdot \mathscr{C}_{x,w},$$

(iv) $$P_1(x_1) = A x_1^{n+1} + \sum_{i=3}^{n} A_i x_1^i,$$
$$P_2(x_1) = B x_1^n + \sum_{i=2}^{n-1} B_i x_1^i,$$

(v) $$\left.\begin{aligned} q_{2j} &\equiv d_j u_j \\ q_{2j-1} &\equiv d_j v_j \end{aligned}\right. \quad \operatorname{mod} \mathfrak{m}_{x,w}^2 \cdot \mathscr{C}_{x,w},$$
$$d_j = 1 - j^2, \quad j \neq 0 \quad \textit{and} \quad d_0 = 1,$$

(vi) *all monomials of f' which have coefficients involving a_j, b_j or c_j have total degree in (x, u, v) at least equal to j; and A, B, A_i, B_i, C_{2i}, C_{2i}' are fixed nonzero multiples of a_{n+1}, b_{n+1}, a_i, b_i, c_i,*

(vii) *Finally, if $in(\varphi_i)$ denotes the lowest degree nonzero terms in the Taylor expansion of φ_i, then $\{(0, x_1), (0, x_2)\} \cup \{in(\varphi_{2i}), in(\varphi_{2i}')\}_{i=1}^n$ span the quotient space $\mathfrak{m}_x/((x_1^2 + x_2^2) \cdot \mathscr{C}_x + \mathfrak{m}_x^{n+1}$.*

From this lemma we can deduce the next one.

LEMMA 7.3. *For f' of the form in the preceding lemma, the corresponding g obtained after applying Lyapunov-Schmidt is infinitesimally stable at* 0 *and has local algebra*

$$Q(g) \xrightarrow{\sim} \mathbb{R}[[x, y]]/(x^2 + y^2, x^{n+1}).$$

Remark. Of course f' itself is infinitesimally stable with the same local algebra; however, it is much easier to show this for g (after further changing coordinates).

Theorem 5 follows from the lemmas, except for the special condition that $q(t)$ can be chosen to have only nonzero Fourier coefficients of degree $\leqq 2n$. In fact, $q(t)$ may be chosen in V_{2n}. This is because Lyapunov-Schmidt is applied to f by composing with Φ^{-1} so the target space is not changed. Then, the restriction of g to V_{2n} is proven to be infinitesimally stable, so by Theorem 1.8, there are $q(t) \in V_{2n}$ arbitrarily close to zero so that $g^{-1}(q)$ contains $2n+2$ points near 0. Thus, so must f.

Proof of Lemma 7.1. The proof of Lemma 7.1 follows from the derivative computations of § 3. The form of the coordinate for sin (t) follows from Corollary 3.13.

For the second term, we can first use Corollary 3.12 to obtain terms of the form $w_1=(1/2)^m\bar{\alpha}\cdot z^m$, where $\alpha=a_m+ib_m$. This gives, for the coefficient of $\cos(t)$,

$$(1/2)^m(a_m\operatorname{Re}(z^m)-b_m\operatorname{Im}(z^m)).$$

Now, we can modify this by adding terms of the form $h\cdot(x_1^2+x_2^2)=h\cdot z\bar{z}$. For example,

$$\operatorname{Re}(z^m)\equiv\operatorname{Re}(z^m+\bar{z}z^{m-1})\equiv 2x_1z^{m-1}\bmod (x_1^2+x_2^2)\mathscr{C}_x.$$

Continuing in this fashion gives

$$\operatorname{Re}(z^m)\equiv\operatorname{Re}(2^{m-1}x_1^{m-1}z)=2^{m-1}x_1^m\bmod (x_1^2+x_2^2)\cdot\mathscr{C}_x.$$

Similarly, mod $(x_1^2+x_2^2)\cdot\mathscr{C}_x$,

$$\operatorname{Im}(z^m)\equiv\operatorname{Im}(2ix_2z^{m-1})=2x_2\operatorname{Re}(z^{m-1})\equiv 2^{m-1}x_2x_1^{m-1}.$$

This gives the terms $P_1+x_2P_2$. The linear term of H_0 also comes from Corollary 3.13. We obtain the terms $C_iu_i\varphi_i+C_i'v_i\varphi_i'$ by applying Lemma 3.14 with $r=m$ so that the nonzero terms for the coefficient of $\cos(t)$ in the expression for f will be given in the lemma by $s=2a$, $0\leqq a\leqq m$. Thus, there will be no terms in $\cos(t)$ coefficient of f of the form $u_{2m}x_1^{j_1}x_2^{j_2}$ with $j_1+j_2<m$. For the degree m-terms in x, we obtain by the same corollary, for $s=2m$, $\bar{\alpha}\beta\bar{z}^{m-1}$ with $\alpha=c_m$ and $\beta=u_m+iv_m$.

This gives for the $\cos(t)$ term

$$(1/2)^m\cdot c_m\cdot(u_m\operatorname{Re}(\bar{z}^{m-1})-v_m\operatorname{Im}(\bar{z}^{m-1})).$$

By the same reasoning as above, $\operatorname{Re}(\bar{z}^{m-1})$ and $\operatorname{Im}(\bar{z}^{m-1})$ are congruent (modulo $(x_1^2+x_2^2)\cdot\mathscr{C}_x$) to $2^{m-2}\cdot x_2^{m-1}$ and $2^{m-2}x_1^{m-2}x_2$. These will be exactly the lowest order terms in which u_{2m} and v_{2m} can appear linearly; thus they are in (φ_m), in (φ_m'). Hence, (vii) follows. The conditions on the coefficients A_i, B_i, etc. follow from the above discussion. Furthermore since a_i, b_i, c_i only appear in the ith derivative, they will still only appear in derivatives of order $\geqq i$ after we apply Lyapunov-Schmidt. Hence, (vi) follows. Finally, the conditions on the functions q_i just state that the first derivative of f is $x''+x$. □

Proof of Lemma 7.3. The Lyapunov–Schmidt procedure corresponds to changing coordinates via

$$(x_1,x_2,u_0',u_1',\cdots,u_m',v_2',\cdots,v_m')=(x_1,x_2,q_0(x,u,v),\cdots,q_{2m}(x,u,v)).$$

Then

$$(u_i,v_i)=(d_i^{-1}u_i'+q_{2i-1}'(x,u,v),d_i^{-1}v_i'+q_{2i}'(x,u,v))\quad\text{with } q_j'\in\mathscr{m}_{x,w}^2. \tag{7.4}$$

We claim that g can be written in the form

$$\begin{gathered}g(x,u',v')=\left(x_1^2+x_2^2+h',P_1+x_2P_2+\sum_{i=2}^{2n}C_i^{(1)}\cdot u_i\cdot\varphi_i+C_i^{(2)}v_i\varphi_i'+H_1,u',v'\right)\\ \text{with } u'=(u_0',\cdots,u_m'),\quad v'=(v_1',\cdots,v_m')\end{gathered} \tag{7.5}$$

obtained from the change of coordinates. Here,

$$h'=(1/2^5)(u_3'x_1-v_3'x_2)+h_1',\qquad H_1=(1/2^5)(-u_3x_2-v_3x_1)+H_1'$$

with H_1', $h_1'\in\mathscr{m}_{x,w'}^3$ and the terms of h_1' still at least quadratic in x. Also, $C_j^{(1)}=d_{2j}^{-1}C_j$, $C_j^{(2)}=d_{2j}^{-1}C_j$ (recall $d_r=1-r^2$).

Again monomials $x^\alpha u'^\beta v^\gamma$ with coefficients involving a_i, b_i, or c_i still have total degree $\geqq i$. Moreover, we may write

$$H_1'=H_1''+H_1'''\quad\text{with } H_1'''\in(x_1^2+x_2^2)\cdot\mathscr{C}_{x,w'}+(\mathscr{m}_{w'}^2+\mathscr{m}_x^{n+2})\cdot\mathscr{C}_{x,w'}.$$

Also, the monomials $x^\alpha u'^\beta v'^\gamma$ in H_1'' with coefficients involving a_i, b_i, or c_i have total degree $>i$.

To show that this claim is correct, we note that the simplest way to carry out the change of coordinates for the first two coordinate functions is to repeatedly use (7.4) to replace each u_i or v_i that appears in the lowest degree terms. If initially a monomial $x^\alpha u^\beta v^\gamma$ has total degree $=j$, then after substitution and expansion, each term which still contains some u_j or v_j has total degree $>j$, while there will be a term of the form $c \cdot x^\alpha u'^\beta v'^\gamma$. Furthermore, for any term with coefficient involving a_i, b_i or c_i in $q_k'(x, u, v)$, the term has total degree $\geqq i$; hence, after substitution and expansion, it will contribute to terms of total degree $\geqq j-1+i$. As $j \geqq 2$, we may continue inductively.

Next, any term in $(x_1^2+x_2^2) \cdot \mathscr{C}_{x,w}$ will still belong to $(x_1^2+x_2^2) \cdot \mathscr{C}_{x,w'}$ after substitution. A similar remark holds for $\mathfrak{m}_x^{n+2}\mathscr{C}_{x,w}$. On the other hand, a monomial in $\mathfrak{m}_w^2 \cdot \mathscr{C}_{x,w}$ yields after substitution, a term in $\mathfrak{m}_{w'}^2 \cdot \mathscr{C}_{x,w'}$ plus monomials of total degree at least one greater. Thus, g has the form (7.5) as claimed.

We further change coordinates

$$x_1 = x_1' + \Psi_1(x', u', v'), \qquad x_2 = x_2' + \Psi_2(x', u', v') \tag{7.6}$$

so that

$$x_1^2 + x_2^2 + h_1'(x, u', v') = x_1'^2 + x_2'^2.$$

We can show this in two steps by first applying the parametrized Morse lemma to write

$$x_1^2 + x_2^2 + h_1'(x, u', v') = x_1'^2 + x_2'^2 + h_2(x_1', x_2', u', v') \tag{7.7}$$

with h_2 quadratic in x_1' and x_2' with coefficients in $\mathscr{C}_{w'}$. Since elements of $1+\mathfrak{m}_{w'}$ are invertible and have square roots, we can make a second local change of coordinates which is linear with coefficients in $\mathscr{C}_{w'}$ by diagonalizing $x_1'^2 + x_2'^2 + h_2$ viewed as a function of x_1', x_2'. This gives a new (x_1', x_2') with the desired properties.

From the first change of coordinates, we can insure that terms in Ψ_i involving a_j, b_j or c_j have total degree $\geqq j-1$ (this is not altered by the second change). This can be seen either from the explicit form for constructing change of coordinates in the Morse lemma, or by using a result of Mather [M3] for $\mathscr{R}^{(k)}$-equivalence. This latter method concerns change of coordinates with $k-1$-jet = id. Let $h_1'^{(k)}$ denote the terms of h_1' of degree $\leqq k$ (with $k \geqq 3$). Then, Mather's result states that $x_1^2+x_2^2+h_1'^{(k+1)}$ is equivalent to $x_1^2+x_2^2+h_1'^{(k)}$ by a local change of coordinates with $k-1$ jet = id if, for $0 \leqq t_0 \leqq 1$ with $\rho^{(k)} = h_1'^{(k+1)} - h_1'^{(k)}$,

$$\rho^{(k)} \in \mathfrak{m}_x^k \cdot \Delta(x_1^2 + x_2^2 + h_1'^{(k)} + (t_0 + t)\rho^{(k)}) \tag{7.8}$$

(where $\Delta(f_t)$ is the ideal in $\mathscr{C}_{x,w',t}$ generated by $\partial f_t/\partial x_1$, $\partial f_t/\partial x_2$). However, by Nakayama's lemma, the right-hand side of (7.8) equals $\mathfrak{m}_x^{k+1} \cdot \mathscr{C}_{x,w',t}$. Thus, the inclusion holds. Composing coordinate changes for decreasing k, $3 \leqq k \leqq n+1$ gives the result about Ψ_i.

When we change coordinates using (7.6) we obtain

$$\begin{aligned} g(x', u', v') = \Big(& x_1'^2 + x_2'^2 + h_2,\ P_1(x_1') + x_2' P_2(x_1') \\ & + \sum_{j=2}^{2n} (C_j^{(1)} u_j \varphi_j(x') + C_j^{(2)} v_j \varphi_j'(x')) + H_2,\ \mathbf{u}', \mathbf{v}' \Big) \end{aligned} \tag{7.9}$$

where now

$$h_2 = (1/2^5)(u_3'\sigma_1 - v_3'\sigma_2), \quad H_2' = (1/2^5)(-u_3'x_2' - v_3'x_1') \quad \text{for } H_2 = H_2' + H_2'' + H_2'''.$$

$H_2'' \in (x_1'^2 + x_2'^2) \cdot \mathscr{C}_{x',w'} + (m_{w'}^2 + m_{x'}^{n+2}) \cdot \mathscr{C}_{x',w'}$ and H_2''' consists of terms involving a_j, b_j and c_j and of total degree $>j$. Here σ_i denotes x_i as a function of (x', u', v').

For the claim concerning H_2', note that $P_1 + x_2 \cdot P_2$ has terms of degree $\geqq 3$; hence after substitution using (7.6), terms involving a_i, b_i or c_i will increase degree by at least two. A similar remark applies to φ_{2i} or φ_{2i}' since deg $in(\varphi_{2i}) =$ deg $in(\varphi_{2i}') = i$. Last, an element of $(m_{w'}^2 + m_x^{n+2}) \cdot \mathscr{C}_{x,w'}$ will be in $(m_{w'}^2 + m_{x'}^{n+2}) \cdot \mathscr{C}_{x',w'}$ after substitution; while an element of $(x_1^2 + x_2^2) \cdot \mathscr{C}_{x,w'}$ differs from an element of $(x_1'^2 + x_2'^2) \cdot \mathscr{C}_{x',w'}$ by an element of $h_1'(x, u', v') \cdot \mathscr{C}_{x',w'}$, which is at least quadratic in x. Again the degree of terms involving a_i, b_i, or c_i will increase by at least 1.

Thus, we can inductively choose a_i and b_i so that the terms of degree i of $P_1 + xP_2 + H_2$ not involving (u', v') belong to $(x_1'^2 + x_2'^2) \cdot \mathscr{C}_{x',w'}$. Furthermore we may write the remaining terms of H_2 as

$$\sum_{j=1}^{2n+2} l_j(u', v') \cdot \psi_j \bmod m_{x',w'}^{n+2} \quad \text{where } l_j = \sum_{i=2}^{2n} a_{ij} u_i' + b_{ij} v_i'$$

and $\{\psi_j, j = 1, \cdots, 2n+2\}$ denotes $\{in(\varphi_{2i}), in(\varphi_{2i}')\}_{i=1}^n \cup \{(0, x_1'), (0, x_2')\}$ and the coefficient functions of $(0, x_1')$ and $(0, x_2')$ are $(1/2)^5$ times $-v_3'$ and $-u_3'$ respectively. Moreover, the preceding discussion implies that l_j does not involve c_i unless deg $(\psi_j) >$ deg $(in(\varphi_{2i})) =$ deg $(in(\varphi_{2i}'))$. Thus, for generic a_n, b_n, and c_i the coefficients A' of $x_1'^{n+1}$ and B' of $x_2' x_1'^n$ are not zero. We also claim that for generic c_i the elements

$$\left\{ C_{2i}^{(1)} in(\varphi_{2i}) + \sum_{j=1}^{2n} a_{2i,j}\psi_j,\ C_{2i}^{(2)} in(\varphi_{2i}') + \sum_{j=1}^{2n} b_{2i,j}\psi_j \right\}_{i=1}^{n} \tag{7.10}$$

span $D_n = m_{x'}^2 / (m_{x'}^{n+1} + (x_1'^2 + x_2'^2) \cdot \mathscr{C}_{x'})$. This follows by replacing D_n by D_k and proceeding by induction on $k \leqq n$ using the fact that $\{in(\varphi_{2i}), in(\varphi_{2i}')\}_{i=1}^k$ span D_k. In the inductive step, we use that $C_{2i}^{(1)}$ and $C_{2i}^{(2)}$ are fixed multiples of c_i so that we must know that a matrix of the form

$$\begin{pmatrix} at + b_{11} & b_{12} \\ b_{21} & bt + b_{22} \end{pmatrix}$$

with a, $b \neq 0$ is nonsingular for generic t, which it clearly is.

Then, for such generic values for a_n, b_n, c_i and specific values of a_j, b_j, $j < n$,

$$g(x', 0) = (x_1^2 + x_2'^2, A' x_1'^{n+1} + B' x_2' x_1'^n + (x_1'^2 + x_2'^2) \cdot R(x'), 0)$$

and has local algebra

$$\mathbb{R}[[x_1', x_2']]/(x_1'^2 + x_2'^2, A' x_1'^{n+1} + B' x_2' x_1'^n) \xrightarrow{\sim} \mathbb{R}[[x, y]]/(x^2 + y^2, x^{n+1})$$

(by the classification in [M4] provided $A' \neq 0$).

Last, if we write $g(x', u', v') = (\bar{g}(x', u', v'), u', v')$, then

$$\left\{ \left.\frac{\partial \bar{g}}{\partial u_{2i}'}\right|_{u',v'=0}, \left.\frac{\partial \bar{g}}{\partial v_{2i}'}\right|_{u',v'=0} \right\}_{i=1}^{n}$$

are exactly (7.9), while

$$\left.\frac{\partial \bar{g}}{\partial u_3}\right|_{u',v'=0} = (1/2)^5 (\sigma_1(x', 0), -x_2') \quad \text{and} \quad \left.\frac{\partial \bar{g}}{\partial v_3}\right|_{u',v'=0} = (1/2)^5 (-\sigma_2(x', 0), -x_1').$$

By the verification criterion Theorem 1.3, g is infinitesimally stable. □

8. Examples of degree n with n^2-periodic solutions. We earlier referred to examples for $n=2$ and 3 of polynomial P of degree n for which there were $q(t)$ arbitrarily close to 0 for which (2.1) has n^2-periodic solutions. Here we briefly describe the examples.

$n=2$. Let

$$P(x,t)=(a+b\cos(t)+c\sin(t))x^2.$$

Then, a computation similar to those made in §§ 4 and 5 shows that if $a\neq 0$, $(b,c)\neq(0,0)$, the operator f in (2.2) (with $m=\lambda=1$) is infinitesimally stable with local algebra $Q(f)\xrightarrow{\sim}\mathbb{R}[[x,y]]/(x^2+y^2,xy)$. Thus, there are $q(t)$ arbitrarily close to 0 for which (2.1) has $\delta(Q(f))=4$ periodic solutions (more generally see [D1, § 3]). Moreover we may find such a $q(t)$ which has Fourier coefficients equal to zero except for $\sin(rt)$, $\cos(rt)$, $0\leqq r\leqq 2$.

$n=3$. Let

$$P(x,x',t)=(a_1\cos(4t)+a_2\cos(2t))x^3+(b_1(t)x^2+b_2(t)x)\cdot x'$$

where for $i=1,2$

$$b_i(t)=c_i\cos(r_it)+d_i\sin(r_it)+c_i'\cos(r_i't)+d_i'\sin(r_i't).$$

We claim that for generic a_i, c_i, c_i', d_i, d_i' and sufficiently general r_i and r_i', there are $q(t)$ arbitrarily close to 0 for which (2.1) (with $m=\lambda=1$) has 9 periodic solutions close to 0. This time the operator f so obtained may not be infinitesimally stable, so we have to modify our arguments slightly.

First, note that the second term of P is G_2 of (5.3) with $n=3$. However, now we have a much smaller version of G_1 in the first term. This is because the quadratic terms in G_2 can affect at most the cubic terms in x_1 and x_2 (using our earlier notation) for the germ g obtained from f by Lyapunov-Schmidt. The first term will contribute to the cubic terms of g. By a calculation using (3.5), we obtain from the first term of P, the following cubic terms in the $\langle\cos(t),\sin(t)\rangle$ coordinates for g,

$$a_4(\alpha x_1^3-3x_1x_2^2,\ \alpha x_2^3-3x_1^2x_2) \tag{8.1}$$

where $a_4\alpha=a_4+a_2$ (assuming $a_4\neq 0$). By the classification of Mather [M4], pairs of generic cubics in two variables contain one modulus parameter. A computation shows that α is the modulus parameter for $\alpha\neq 0$. Such pairs of generic cubics form a Zariski open subset of the space of pairs of cubics. Hence, for sufficiently small values of the constants c_i, d_i, etc., it follows that α (and hence a_2, for a_4 fixed) is still a modulus when we include the contribution to the cubic terms from G_2. Then, by the verification criterion, the operator

$$\tilde f: C^2_{2\pi}\times\mathbb{R}\to C^0_{2\pi}\times\mathbb{R},\qquad \tilde f(x(t),a_2)\to(f(x(t)),a_2)$$

is infinitesimally stable at $(0,a_2)$ for c_i, d_i, etc. sufficiently small, $a_4\neq 0$, and $a_2\neq -a_4$. Thus, it is so for generic values of c_i, d_i, etc., by the same argument given in § 6. Thus, for generic c_i, $d_i,\cdots,a_2$, $\tilde f$ has real multiplicity at $(0,a_2)$ equal to dimension of the local algebra of $\tilde f$ at $(0,a_2)$, which is 9. Thus, there is $(q(t),a_2')$ arbitrarily close to $(0,a_2)$ for which there are 9 solutions to $\tilde f(x(t),a_2')=(q(t),a_2')$, i.e., 9 periodic solutions to (2.1) for the specific value of a_2'. This does not quite prove what was claimed, since when we pick a point $(q_1(t),a_2'')$ still closer with 9 solutions we have no guarantee that $a_2''=a_2'$. Thus, for a fixed a_2, we do not know that $q(t)$ can be chosen arbitrarily close to 0 so there are still 9 solutions for the given a_2. However, this is true because by [D3, § 9, Thm. 4], for almost all values α (or a_2), the infinitesimally stable germ is topologically a product mapping along the α-direction. This implies the desired result.

This argument does not work in the general case of degree n; however, there is an alternate way to obtain the case $n = 3$ which suggests that the result should be true for general n, provided certain technical points in the computation can be handled.

Acknowledgment. Special gratitude is expressed to Jorge Sotomayor for his valuable comments and suggestions concerning earlier and cruder forms of these results.

REFERENCES

[AP] A. Ambrosetti and G. Prodi, *On the inversion of some differentiable maps with singularities*, Annali di Math., 93 (1972), pp. 231-246.

[BC] M. Berger and P. Church, *Complete integrability and perturbation of a nonlinear Dirichlet problem* I, Indiana Math. J., 28 (1979), pp. 935-952.

[BCT] M. Berger, P. Church and J. Timourian, *An application of singularity theory to non-linear elliptic partial differential equations*, Proc. Sympos. Pure Math., 40 (1983), pp. 119-126.

[CH] S. Chow and J. Hale, *Methods in Bifurcation Theory*, Springer-Verlag, Heidelberg-New York, 1982.

[D1] J. Damon, *On a theorem of Mather and the local structure of non-linear Fredholm maps*, Proc. Sympos. Pure Math., 45 (1986), pt. I, pp. 339-352.

[D2] ———, *The relation between C^∞ and topological stability*, Bol. Soc. Brasil. Math., 8 (1977), pp. 1-38.

[D3] ———, *Finite determinacy and topological triviality* II: *sufficient conditions and topological stability*, Compositio Math., 47 (1982), pp. 101-132.

[DG] J. Damon and A. Galligo, *A topological invariant for stable map germs*, Invent. Math., 32 (1976), pp. 103-132.

[F] L. Fraenkel, *Formulae for high derivatives of composite functions*, Proc. Cambridge Philos. Soc., 83 (1978), pp. 159-165.

[GG] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1973.

[LMP] A. Lins, W. deMelo and C. C. Pugh, *On Lienhard's equation*, Geometry and Topology, Rio de Janeiro, Lecture Notes in Math., 597, Springer-Verlag, Berlin-New York, 1977, pp. 335-357.

[M1] J. Mather, *C^∞-stability of mappings.*

[M2] ———, II *Infinitesimal stability implies stability*, Ann. Math., 89 (1969), pp. 254-291.

[M3] ———, III *Finitely determined map germs*, Inst. Hautes Etudes Sci. Publ. Math., 35 (1968), pp. 127-156.

[M4] ———, IV *The classification of stable germs by $\mathbb{R}$-algebras*, Inst. Hautes Etudes Sci. Publ. Math., (1969), pp. 223-248.

[M5] ———, V *Transversality*, Adv. in Math., 4 (1970), pp. 301-336.

[M6] ———, VI *The nice dimensions*, Liverpool Singularities Symposium, Lecture Notes in Math., 192, Springer-Verlag, Berlin-New York, 1970, pp. 207-253.

[R] F. Ronga, *A new look at Faá de Bruno's formula for higher derivatives of composite functions and the expression of some intrinsic derivatives*, Proc. Sympos. Pure Math., 40 (1983), pp. 423-432.

[S1] S. Shahshahani, *Periodic solutions of polynomial first order differential equations*, Nonlinear Anal., 5 (1981), pp. 157-165.

[S2] ———, *Some examples of dynamical systems*, preprint.

[T] F. Takens, *Unfoldings of certain singularities of vector fields: Generalized Hopf bifurcation*, J. Differential Equations, 14 (1973), pp. 476-493.

[GL] M. Golubitsky and W. F. Langford, *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41 (1981), pp. 375-415.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

008

THE CONVERSE OF PÓLYA'S MEAN VALUE THEOREM*

JAMES S. MULDOWNEY†

Abstract. Suppose L is a real linear scalar ordinary differential operator of order n with continuous coefficients which is disconjugate on an interval I. Pólya showed that if v is any n times differentiable function on I which has $n+1$ zeros, then $Lv(p)=0$ for some point p intermediate to the zeros of v. It is shown that an operator L has this mean value property with respect to functions v on an interval I if and only if L is disconjugate on the interior of I.

Key words. linear equations, zeros, disconjugacy

AMS(MOS) subject classification. Primary 34C10

1. Introduction. We consider linear differential operators of the form

$$Lu = u^{(n)} + a_1(t)u^{(n-1)} + \cdots + a_n(t)u \tag{1.1}$$

where a_i, $i=1,\cdots,n$ are continuous real valued functions on a fixed open interval J. Such an operator is said to be *disconjugate* on an interval $I \subset J$ if the only solution of $Lu=0$ which, counting multiplicities, has n or more zeros in I is the zero solution.

In the paper [4], Pólya shows that disconjugate operators have the following mean value property. Let v be any real valued function such that $v^{(n)}$ exists on I and v has $n+1$ or more zeros in I. Then, if L is disconjugate on I, there exists a point p intermediate to the zeros of v such that $Lv(p)=0$. The result is a generalization of Rolle's Theorem and, in fact, Pólya's proof is an induction on the order of L based on Rolle's Theorem.

Pólya also shows in [4, Thm. V] that, if v is a function which has n zeros in an interval I on which L is disconjugate and Lv is of one sign, then v satisfies a certain sign restriction on I. This result is often referred to as Čaplygin's Inequality since the case of the n zeros of v all occurring at a single point was investigated by Čaplygin (cf. [1]).

In this paper we consider the question of whether the converses of the Pólya Mean Value Theorem and Čaplygin's Inequality hold. For example, if L has the Pólya mean value property on an interval I, is L necessarily disconjugate on I? We find that this is in fact the case if the interval I is not closed. If I is a closed interval on which L has the Pólya mean value property, we show that L is necessarily disconjugate on the interior of I.

2. Results. The following notation will be used throughout this paper.

DEFINITION 2.1. (a) Let $t_i \in J$, $i=1,\cdots,m$, $t_1 \leqq t_2 \leqq \cdots \leqq t_m$ and consider the m-tuple $\tau = (t_1,\cdots,t_m)$. If $t\in J$, then $\tau^t = (s_1,\cdots,s_{m+1})$ is the $(m+1)$-tuple obtained by inserting an extra entry t in τ in such a way that $s_1 \leqq s_2 \leqq \cdots \leqq s_{m+1}$. If $s \in \tau$, then τ_s denotes the $(m-1)$-tuple obtained by deleting one occurrence of the entry s from τ. In particular τ_s^t denotes the m-tuple obtained from τ by deleting an entry s from τ and adding an entry t in such a way that the ordering is maintained.

* Received by the editors March 18, 1985; accepted for publication May 22, 1986. This work was supported by Natural Sciences and Engineering Research Council of Canada grant 7197.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

(b) A function v has m zeros at $\tau=(t_1,\cdots,t_m)$ if $v(t_i)=v'(t_i)=\cdots=v^{(j-1)}(t_i)=0$ for each distinct entry $t_i\in\tau$, where j is the number of times the entry t_i occurs in τ.

(c) If $u_1,\cdots,u_m$ are real functions on J and τ is as in (a), $\mathcal{W}(u_1,\cdots,u_m)(\tau)$ denotes the determinant of the $m\times m$ matrix whose kth column is $\operatorname{col}(u_k(t_1),\cdots,u_k(t_m))$, $k=1,\cdots,m$, when the entries t_i in τ are all distinct; the terms $u_k(t_i), u_k(t_{i+1}),\cdots,u_k(t_{i+j-1})$ are replaced by $u_k(t_i), u_k'(t_i),\cdots,u_k^{(j-1)}(t_i)$ if the entry t_i occurs j times in τ. In particular, the Wronskian determinant $W(u_1,\cdots,u_m)(t)=\mathcal{W}(u_1,\cdots,u_m)(t,\cdots,t)$.

(d) For τ as in (a) $\operatorname{co}(\tau)=[t_1,t_m]$.

Definition 2.2. Let $I\subset J$ be an interval and let int I, cl I denote the interior and closure in J, respectively, of I.

(a) $C'_{n-1}(I)$ denotes the set of real valued functions v on I such that $v^{(n-1)}$ exists and is continuous on I and $v^{(n)}$ exists on int I.

(b) $\mathcal{D}_n(I)$ is the set of operators L of the form (1.1) which are disconjugate on I as defined in the first paragraph of the introduction.

(c) $\mathcal{P}_n(I)$ is the set of operators L of the form (1.1) with the property that if $v\in C'_{n-1}(I)$ has $n+1$ zeros at $(t_1,\cdots,t_{n+1})$, $t_i\in I$, $t_1\leqq t_2\leqq\cdots\leqq t_{n+1}$, $t_1<t_{n+1}$, then $Lv(p)=0$ for some $p\in(t_1,t_{n+1})$. Thus $\mathcal{P}_n(I)$ is the set of operators with the Pólya mean value property on I.

(d) $\mathcal{Q}_n(I)$ denotes the set of operators L of the form (1.1) with the property that if $t\in I$ and $v\in C'_{n-1}(I)$ has n zeros at $\tau=(t_1,\cdots,t_n)$, $t_i\in I$, $t\neq t_i$ and $Lv\geqq 0$ on int I with $Lv(p)>0$ for some $p\in\operatorname{co}(\tau^t)$, then

$$\prod_{i=1}^{n}(t-t_i)v(t)>0. \tag{2.1}$$

$\mathcal{Q}_n(I)$ is the set of operators for which Čaplygin's inequality holds on I.

We will prove the following theorem:

Theorem 2.3. (a) *For any subinterval I of J,*

$$\mathcal{D}_n(\operatorname{cl} I)\subset\mathcal{D}_n(I),\quad \mathcal{P}_n(\operatorname{cl} I)=\mathcal{P}_n(I),\quad \mathcal{Q}_n(\operatorname{cl} I)=\mathcal{Q}_n(I);$$

(b) *For any subinterval I of J such that I is not closed in* $\mathbf{R}$,

$$\mathcal{D}_n(I)=\mathcal{P}_n(I)=\mathcal{Q}_n(I).$$

It is clear that $\mathcal{D}_n$, $\mathcal{P}_n$, $\mathcal{Q}_n$ are nonincreasing with respect to set inclusion: $I_1\subset I_2$ implies

$$\mathcal{D}_n(I_2)\subset\mathcal{D}_n(I_1),\quad \mathcal{P}_n(I_2)\subset\mathcal{P}_n(I_1),\quad \mathcal{Q}_n(I_2)\subset\mathcal{Q}_n(I_1).$$

In particular,

$$\mathcal{D}_n(\operatorname{cl} I)\subset\mathcal{D}_n(I),\quad \mathcal{P}_n(\operatorname{cl} I)\subset\mathcal{P}_n(I),\quad \mathcal{Q}_n(\operatorname{cl} I)\subset\mathcal{Q}_n(I). \tag{2.2}$$

We will prove Theorem 2.3 by establishing the following three propositions and using (2.2).

Proposition 2.4. *If I is any subinterval of J, then*

$$\mathcal{Q}_n(I)\subset\mathcal{P}_n(I).$$

Proposition 2.5. *If I is any subinterval of J, then*

$$\mathcal{D}_n(I)\subset\mathcal{Q}_n(\operatorname{cl} I).$$

PROPOSITION 2.6. *If I is any subinterval of J such that I is not closed in* $\mathbf{R}$, *then*

$$\mathcal{P}_n(I) \subset \mathcal{D}_n(I).$$

Since Theorem 2.3(a) is trivial if I is closed, we may suppose that I is not closed. Then the three propositions and (2.2) imply

$$\mathcal{Q}_n(\mathrm{cl}\, I) \subset \mathcal{Q}_n(I) \subset \mathcal{P}_n(I) \subset \mathcal{D}_n(I) \subset \mathcal{Q}_n(\mathrm{cl}\, I),$$

$$\mathcal{Q}_n(\mathrm{cl}\, I) \subset \mathcal{P}_n(\mathrm{cl}\, I) \subset \mathcal{P}_n(I) \subset \mathcal{D}_n(I) \subset \mathcal{Q}_n(\mathrm{cl}\, I)$$

which proves Theorem 2.3(a),(b).

Proof of Proposition 2.4. To prove Proposition 2.4, suppose $L \in \mathcal{Q}_n(I)$, $v \in C'_{n-1}(I)$ and v has $n+1$ zeros at $(t_1, \cdots, t_{n+1})$, $t_i \in I$, $t_1 \leqq t_2 \leqq \cdots \leqq t_{n+1}$, $t_1 < t_{n+1}$. If $Lv > 0$ on (t_1, t_{n+1}), then

$$\prod_{i=1}^{n} (t-t_i)v(t) > 0, \qquad \prod_{i=2}^{n+1} (t-t_i)v(t) > 0$$

if $t \in (t_1, t_{n+1})$, $t \neq t_i$, from (2.1), and these two inequalities contradict each other. Similarly, the assumption $Lv < 0$ on (t_1, t_{n+1}) leads to a contradiction. Thus $Lv(p) = 0$ for some $p \in (t_1, t_{n+1})$ and we conclude that $L \in \mathcal{P}_n(I)$, which proves Proposition 2.4.

Pólya's Mean Value Theorem states that $\mathcal{D}_n(I) \subset \mathcal{P}_n(I)$ and Čaplygin's Inequality, which was also first proved in generality by Pólya [4], states that $\mathcal{D}_n(I) \subset \mathcal{Q}_n(I)$ for any interval I. In view of (2.2) and Proposition 2.4, it follows that Proposition 2.5 is a slightly stronger statement than both of these. To prove Proposition 2.5, we shall need the following lemma which is essentially Theorem V of [4] and is Pólya's statement of Čaplygin's inequality. A detailed proof may be found in [3, pp. 376-378].

LEMMA 2.7. *Suppose* $u_1, \cdots, u_{n+1} \in C'_{n-1}(I)$,

$$u_1 > 0, \qquad W(u_1, u_2) > 0, \cdots, W(u_1, \cdots, u_n) > 0 \tag{2.3}$$

and $W(u_1, \cdots, u_{n+1}) \geqq 0$. *Then, if* $\sigma = (t_1, \cdots, t_{n+1})$, $t_i \in I$, $t_1 \leqq t_{n+1}$,

$$\mathcal{W}(u_1, \cdots, u_{n+1})(\sigma) \geqq 0 \tag{2.4}$$

and the inequality (2.4) *is strict if* $W(u_1, \cdots, u_{n+1})(p) > 0$ *for some* $p \in [t_1, t_{n+1}]$.

Proof of Proposition 2.5. We first show that Lemma 2.7 implies Proposition 2.5 for subintervals I of J which are closed relative to J. Suppose $I = \mathrm{cl}\, I$, $L \in \mathcal{D}_n(I)$, $v \in C'_{n-1}(I)$ has n zeros at $\tau = (t_1, \cdots, t_n)$, $t_i \in I$, $Lv \geqq 0$ and $t \in I$ is such that $Lv(p) \geqq 0$ for some $p \in \mathrm{co}\,(\tau^t)$. Since $L \in \mathcal{D}_n(\mathrm{co}\,(\tau^t))$, solutions $u_1, \cdots, u_n$ of $Lu = 0$ may be chosen so that (2.3) is satisfied on $\mathrm{co}\,(\tau^t)$ (cf. [2, p. 94]). Also $Lv = W(u_1, \cdots, u_n, v)/W(u_1, \cdots, u_n)$ so that Lemma 2.7 implies $\mathcal{W}(u_1, \cdots, u_n)(\tau) > 0$, $\mathcal{W}(u_1, \cdots, u_n, v)(\tau^t) > 0$, and therefore

$$\mathcal{W}(u_1, \cdots, u_n, v)(\tau^t)/\mathcal{W}(u_1, \cdots, u_n)(\tau) > 0. \tag{2.5}$$

Now since

$$\mathcal{W}(u_1, \cdots, u_n, v)(\tau^t) = v(t)\mathcal{W}(u_1, \cdots, u_n)(\tau)\ \mathrm{sgn} \prod_{i=1}^{n} (t-t_i), \tag{2.6}$$

we conclude that v satisfies (2.1) and $L \in \mathcal{Q}_n(I)$. Thus Proposition 2.5 holds if $I = \mathrm{cl}\, I$.

The discussion of the preceding paragraph also establishes Proposition 2.5 for subintervals I of J which have at least one endpoint in common with J, since (cf. [2, p. 102])

$$\mathcal{D}_n(I) = \mathcal{D}_n(\text{int } I) \quad \text{if } i \text{ is not closed.} \tag{2.7}$$

Also (2.7) shows that, to complete the proof, it suffices to consider open intervals $I = (a, b)$ which have no endpoint in common with J. It is assumed that L and v satisfy the same conditions as before while the conditions on t_i, $i = 1, \cdots, n$ and t are relaxed to t_i, $t \in \text{cl } I$. The discussion of the preceding paragraph applied to closed subintervals of I and continuity considerations imply that if $Lu_i = 0$, $i = 1, \cdots, n$ and $W(u_1, \cdots, u_n) > 0$, then

$$\mathcal{W}(u_1, \cdots, u_n)(\tau) \geqq 0, \qquad \mathcal{W}(u_1, \cdots, u_n, v)(\tau^t) \geqq 0. \tag{2.8}$$

We assert that the existence of the function v as described implies that both of these inequalities are strict, and the result follows from (2.5), (2.6) as when I is closed.

To prove the last assertion, suppose first that $\mathcal{W}(u_1, \cdots, u_n)(\tau) = 0$. From (2.7), both endpoints a, b of I must occur in τ since $\mathcal{W}(u_1, \cdots, u_n)(\tau) = 0$ implies the existence of a nontrivial solution u of $Lu = 0$ which has n zeros at τ. We suppose further that, of all τ, v satisfying our assumptions, the ones under consideration are such that the number of occurrences of endpoints of I in τ is a minimum. The constant c may be chosen so that $w = v + cu$ has a zero at $s \in I$, $s \notin \tau$. Therefore w has $n+1$ zeros at $\tau^s = (s_1, \cdots, s_{n+1})$ and, in particular, w has n zeros at each of $\sigma = \tau_b^s = (s_1, \cdots, s_n)$, $\rho = \tau_a^s = (s_2, \cdots, s_{n+1})$. The number of occurrences of endpoints of I in each of σ, ρ is one fewer than in τ. Since $Lw = Lv \geqq 0$ and $Lw(p) = Lv(p) > 0$, and from the minimal character of τ with respect to endpoints, we conclude that $\mathcal{W}(u_1, \cdots, u_n)(\sigma) > 0$, $\mathcal{W}(u_1, \cdots, u_n)(\rho) > 0$. The second inequality (2.8) and (2.6) hold with v, τ, τ^t replaced by w, σ, σ^t respectively and by w, ρ, ρ^t, respectively. Therefore

$$\prod_{i=1}^{n} (t - s_i) w(t) \geqq 0, \qquad \prod_{i=2}^{n+1} (t - s_i) w(t) \geqq 0$$

for each $t \in I$, and $w(t) = 0$ for all $t \in I$ contradicting $Lw(p) > 0$; we conclude that either $\mathcal{W}(u_1, \cdots, u_u)(\tau) > 0$ or v satisfying our hypothesis does not exist. It remains to show that $\mathcal{W}(u_1, \cdots, u_n, v)(\tau^t) > 0$ for each $t \in \text{cl } I$; if not then (2.6), (2.8) imply that v has $(n+1)$ zeros at $\tau^s = (r_1, \cdots, r_{n+1})$ for some $s \in \text{cl } I$ and hence v has n zeros at each of $\delta = \tau_a^s$, $\gamma = \tau_b^s$. We have shown that the hypotheses on v imply $\mathcal{W}(u_1, \cdots, u_n)(\delta) > 0$, $\mathcal{W}(u_1, \cdots, u_n)(\gamma) > 0$ and, as before,

$$\prod_{i=1}^{n} (t - r_i) v(t) \geqq 0, \qquad \prod_{i=2}^{n+1} (t - r_i) v(t) \geqq 0$$

implying $v = 0$, which contradicts $Lv(p) > 0$.

Before proving Proposition 2.6 we recall some facts about Green's functions. The function $G(t, s)$, $t, s \in J$ is Green's function associated with the boundary value problem

$$Lv = f, \quad v \text{ has } n \text{ zeros at } \tau, \tag{2.9}$$

if it has the following properties:

(i) For each $s \in J$, $u(t) = G(t, s)$ is a solution of $Lu = 0$ for $t < s$ and $t > s$ and u has n zeros at τ, if $s \notin \tau$.

(ii) For each $s \in J$, $u(t)$ and its first $n-2$ derivatives are continuous at s and $u^{(n-1)}(s+) - u^{(n-1)}(s-) = 1$.

Green's function exists provided $\mathcal{W}(u_1, \cdots, u_n)(\tau) \neq 0$, where $u_1, \cdots, u_n$ is a fundamental solution set for $Lu = 0$ [2, p. 106] and has the property that, if $f \in C(J)$, the solution of (2.9) is

$$v(t)=\int_J f(s)G(t,s)\,ds. \tag{2.10}$$

In the special case that all n points in τ equal t_1, $G(t,s)$ is called the Cauchy function at t_1 for L. Thus the Cauchy function at t_1 is defined by: $G(t,s)=u(t)$ is a solution of $Lu=0$, $u(s)=\cdots=u^{(n-2)}(s)=0$, $u^{(n-1)}(s)=\operatorname{sgn}(t-s)$, if $|s-t_1|\leqq|t-t_1|$ and $G(t,s)=0$ otherwise.

LEMMA 2.8. *Suppose $L\in\mathcal{D}_n[b,c]$ and $G(t,s)$ is the Cauchy function for L at b. Then*

$$\begin{aligned} G(t,s)&=0, \qquad b\leqq t\leqq s\leqq c,\\ G(t,s)&>0, \qquad b\leqq s<t\leqq c. \end{aligned} \tag{2.11}$$

Proof. Since $G(t,s)=0$, $b\leqq t\leqq s$ and $G(t,s)=u(t)$, $b\leqq s\leqq t$, is a solution of $Lu=0$ which has $n-1$ zeros at s, it follows from $L\in\mathcal{D}_n[b,c]$ that u cannot have a zero in $(s,c]$ and thus $u^{(n-1)}(s+)=1$ implies $u(t)>0$, if $t\in(s,c]$.

Proof of Proposition 2.6. Suppose the proposition fails to hold for some interval $I\subset J$. Since I is not closed, it may be assumed that I is open, from (2.7). Our hypothesis implies the existence of $L\in\mathcal{P}_n(I)$, $L\notin\mathcal{D}_n(I)$. Thus I contains a subinterval $[a,b]$ such that b is the first right conjugate point of a with respect to L:

$$L\in\mathcal{D}_n[a,b),\qquad L\notin\mathcal{D}_n[a,b]. \tag{2.12}$$

We will show that this leads to a contradiction. First (2.12) implies that a fundamental solution set $u_1,\cdots,u_n$ of $Lu=0$ satifies $\mathcal{W}(u_1,\cdots,u_n)(\sigma)=0$ for some σ, $\operatorname{co}(\sigma)=[a,b]$ (both endpoints a, b occur in σ). We suppose that σ is such that the number $m+1$, $m\geqq 0$, of occurrences of b in σ is a minimum. There is a nontrivial solution u_m of $Lu=0$ such that u_m has n zeros at σ and there is a neighbourhood of b in which the only zeros of u_m are the $m+1$ zeros at b. Choose $c\in I$ such that $c>b$, u_m has no zeros in $(b,c]$ and

$$L\in\mathcal{D}_n[b,c]. \tag{2.13}$$

Now let $\tau=\sigma_b^a$, the n-tuple obtained by deleting one of the entries b from σ and adding an extra entry a. Note that b occurs $n\geqq 0$ times in τ. From the minimal character of σ with respect to b, it follows that $\mathcal{W}(u_1,\cdots,u_n)(\tau)\neq 0$ and Green's function $G(t,s)$ for (2.9) exists. For $f\in C(J)$, the solution v of (2.9) is given by (2.10). Moreover, if $f>0$, v satisfies (2.1), from (2.12) and Proposition 2.5, since

$$L\in\mathcal{D}_n[a,b)\subset\mathcal{Q}_n[a,b].$$

From (2.1) we conclude that $v^{(m)}(b)\geqq 0$. In fact this inequality is strict since, if $v^{(m)}(b)=0$, v would have at least $n+1$ zeros in $[a,c]\subset I$ and $Lv=f>0$ contradicting $L\in\mathcal{P}_n(I)$. We conclude that, if $f>0$, the solution v of (2.9) satisfies

$$\begin{aligned} &v(b)=\cdots=v^{(m-1)}(b)=0, \qquad v^{(m)}(b)>0 \quad \text{if } m>0,\\ &v(b)>0 \quad \text{if } m=0. \end{aligned} \tag{2.14}$$

Note that if $s>b$ and $t\leqq s$, then $G(t,s)=0$ since $u(t)=G(t,s)$ is a solution of $Lu=0$ with n zeros at τ, which implies $u=0$, because $\mathcal{W}(u_1,\cdots,u_n)(\tau)\neq 0$. Thus, if $s,t\in(b,c)$, $G(t,s)$ is the Cauchy function for L at b and, from (2.13), satisfies (2.11). Similarly $G(t,s)=0$, if $s<a$ and $t>s$.

Consider $K(t, s) = G(t, s) - G(c, s)u_m(t)/u_m(c)$. Observe that $u(t) = K(t, s)$ is a solution of $Lu = 0$ if $t \neq s$, u and its first $n-2$ derivatives are continuous at s, $u^{(n-1)}(s+) - u^{(u-1)}(s-) = 1$ and u has $n-1$ zeros at the points common to τ and σ as well as a zero at c. Thus $K(t, s)$ is Green's function for the problem

$$Lw = f, \quad w \text{ has } n \text{ zeros at } \tau_a^c. \tag{2.15}$$

The point b occurs m times in τ_a^c. Let $t_0 \in (b, c)$ and let $f \in C(J), f > 0$, be chosen so that

$$\left| \int_a^{t_0} f(s)[G(t_0, s) - G(c, s)u_m(t_0)/u_m(c)]\, ds \right| < \int_{t_0}^c f(s)G(c, s)u_m(t_0)/u_m(c)\, ds. \tag{2.16}$$

This may be accomplished by choosing f sufficiently large on (t_0, c) since, as we have seen in the preceding paragraph, (2.11) implies $G(c, s) > 0$ for $s \in (b, c)$ and u_m has no zeros in $(b, c]$.

The solution of (2.15) is

$$w(t) = \int_J f(s)K(t, s)\, ds \tag{2.17}$$

and this function has m zeros at b, with

$$w^{(m)}(b) = \int_J f(s)\frac{\partial^m}{\partial t^m}K(b, s)\, ds = \int_J f(s)\frac{\partial^m}{\partial t^m}G(b, s)\, ds \tag{2.18}$$

since $u^{(m)}(b) = 0$, because u_m has $m+1$ zeros at b. Therefore

$$w^{(m)}(b) = v^{(m)}(b) > 0 \quad \text{from (2.10), (2.14), (2.18),}$$

so that for some $\varepsilon > 0$

$$w(t) > 0, \qquad t \in (b, b+\varepsilon). \tag{2.19}$$

Now from (2.17)

$$\begin{aligned} w(t_0) = \int_J f(s)K(t_0, s) &= \int_J f(s)[G(t_0, s) - G(c, s)u_m(t_0)/u_m(c)]\, ds \\ &= \int_a^{t_0} f(s)[G(t_0, s) - G(c, s)u_m(t_0)/u_m(c)]\, ds \\ &\quad - \int_{t_0}^c f(s)G(c, s)u_m(t_0)/u_m(c)\, ds, \end{aligned}$$

since $G(t_0, s) = G(c, s) = 0$, $s < a$ and $G(t_0, s) = 0$, $s > t_0$ and $G(c, s) = 0$, $s > c$. It follows from (2.16) that $w(t_0) < 0$ which, with (2.19), implies w has a zero in (b, c). Since w has n zeros at τ_a^c and a zero in (b, c), w has $n+1$ zeros in $[a, c] \subset I$. But then $Lw = f > 0$ contradicts $L \in \mathcal{P}_n(I)$.

We have now shown that the assumption $L \in \mathcal{P}_n(I)$, $L \notin \mathcal{D}_n(I)$, where I is a subinterval of J which is not closed, leads to a contradiction. This establishes Proposition 2.6.

REFERENCES

[1] S. A. Čaplygin, *New methods in the approximate integration of differential equations*, Gosudartsv, Izdat. Tech-Teoret. lit., Moscow, 1950. (In Russian.)

[2] W. A. Coppel, *Disconjugacy*, Lecture Notes in Mathematics 220, Springer-Verlag, New York, 1971.

[3] S. Karlin and W. J. Studden, *Tchebycheff Systems with Applications in Analysis and Statistics*, Interscience, New York, 1966.

[4] G. Pólya, *On the mean-value theorem corresponding to a given linear homogeneous differential equation*, Trans. Amer. Math. Soc., 24 (1922), pp. 312-324.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

009

SUBSPACES OF STABLE AND UNSTABLE SOLUTIONS OF A FUNCTIONAL DIFFERENTIAL EQUATION IN A FADING MEMORY SPACE: THE CRITICAL CASE*

G. S. JORDAN†, OLOF J. STAFFANS‡ AND ROBERT L. WHEELER§

Abstract. We study the asymptotic behavior of the linear infinite delay, autonomous system of functional differential equations

$$(*)\qquad \begin{aligned} &x'(t)+\mu * x(t)=f(t) \qquad (t\geqq 0),\\ &x(t)=\phi(t) \qquad (t\leqq 0).\end{aligned}$$

Here μ is an n-dimensional matrix-valued measure supported on $[0,\infty)$, finite with respect to a weight function, and f, ϕ and x are $\mathbf{C}^n$-valued continuous or locally integrable functions bounded with respect to a fading memory norm. We find conditions that ensure that the state space of (*) can be written as a direct sum of a stable subspace, which is characterized by the fact that solutions are small at infinity, a finite dimensional central-stable subspace in which solutions are neither small nor large at infinity, and a finite dimensional exponentially unstable subspace consisting of exponentially growing solutions. We give estimates for the rate of decay at infinity of solutions belonging to the stable subspace. Our results extend earlier work of Staffans [10], [11] since we analyze the critical case in which the components of the solutions are not exponentially separated, as well as the noncritical case.

Key words. functional differential equations, convolution, infinite delay, state space decomposition

AMS(MOS) subject classifications. Primary 34K25, 45D05, 45E10, 45F05, 45M05

1. Introduction.

We study the asymptotic behavior of the solutions of the linear, infinite delay, autonomous system of functional differential equations

$$(1.1)\qquad \begin{aligned} &x'(t)+\mu * x(t)=f(t), \qquad t\in\mathbf{R}^+,\\ &x(t)=\phi(t), \qquad t\in\mathbf{R}^-.\end{aligned}$$

Here $\mathbf{R}^+=[0,\infty)$, $\mathbf{R}^-=(-\infty,0]$, μ is an n by n matrix-valued measure supported on $\mathbf{R}^+$ which is finite with respect to a weight function, and x, f and ϕ are $\mathbf{C}^n$-valued functions. The initial function ϕ and the forcing function f belong to certain weighted function spaces compatible with the weighted measure space containing μ. As usual, $\mu * x$ denotes the convolution

$$(\mu * x)(t)=\int_{-\infty}^{\infty} d\mu(s)x(t-s).$$

We find conditions that ensure that the solution subspace of (1.1) can be decomposed into a direct sum of a *stable subspace* $\mathcal{S}$, which is characterized by the fact that the solutions in $\mathcal{S}$ are small at infinity, a finite dimensional *central-stable subspace* $\mathcal{C}$ in which solutions do not decay, but also do not grow at exponential rates, and finally a finite dimensional *unstable subspace* $\mathcal{U}$ consisting of exponentially growing solutions.

* Received by the editors July 3, 1986; accepted for publication November 10, 1986.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300.

‡ Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

§ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-4097. The work of this author was partially supported by the National Science Foundation under grants MCS 8300559 and DMS-850097.

The question that we consider here has been discussed before in a very similar setting in [10] and [11], and our present results contain those of [10] and [11]. The chief improvement of our results over those in [10] and [11] is that we now treat the critical case in which the growth rates of the components of the solution are not exponentially separated, as well as the noncritical case previously studied.

We do not employ semigroup theory, but one could give a semigroup formulation of our results. In this formulation, the noncritical case is the one where the stable subspace and its complementary subspace correspond to disjoint, separated parts of the spectrum of the generator. This is in contrast to the critical case where the spectrum corresponding to the central subspace lies on the boundary of the spectrum corresponding to the stable subspace.

We also give an estimate for the rate of decay at infinity of solutions belonging to the stable subspace (Corollary 6.1). In the case of systems of integrodifferential equations, a similar decay estimate was obtained by Jordan and Wheeler [5]. For other results on the structure of solutions of systems of integrodifferential equations, see [6] and [8]. A similar structure theorem for an integrodifferential equation in Hilbert space is given in [9].

In order to treat the critical case, we have been forced first to refine the description given in [11] of the null space of the convolution operator appearing on the left side of (1.1), since the description in [11] applies only to the noncritical case. We have separated this part of the theory into the paper [4], and the present paper can be regarded as a continuation of [4]. In [4] we describe the null space of this operator in terms of the Jordan chains at its eigenvalues, and to do this we develop a Smith factorization theorem for locally analytic matrix-valued functions. A similar description of the null space in the case of finite delay equations is given by Kappel and Wimmer [7]. We expect the reader to be familiar with the concepts and results of [4].

2. The functional differential equation. The setting in which we study the functional differential equation (1.1) is very similar to the setting used in § 7 of [11]. It is more general in the sense that the critical case is also included. The equation in [1] is neutral rather than retarded, but to keep the technical details as simple as possible we here require the equation to be retarded.

As we already mentioned above, we expect the reader to be familiar with [4]. In spite of this, let us recall the most basic concepts from [4].

We call a continuous positive function ρ a *dominating function* or a *weight function* if ρ is submultiplicative, i.e., ρ satisfies

$$\rho(s+t) \leqq \rho(s)\rho(t), \qquad s, t \in \mathbf{R}, \tag{2.1}$$

and $\rho(0)=1$. A continuous positive function η is called an *influence function* dominated by ρ if $\eta(0)=1$ and

$$\eta(s+t) \leqq \eta(s)\rho(t), \qquad t, s \in \mathbf{R}. \tag{2.2}$$

It follows from (2.1) and (2.2) that η always satisfies

$$[\rho(-t)]^{-1} \leqq \eta(t) \leqq \rho(t), \qquad t \in \mathbf{R}. \tag{2.3}$$

Throughout this paper we work in weighted spaces, where the weights are either dominating functions or influence functions. For example, if we let the set of n by n matrices be denoted by $\mathbf{C}^{n\times n}$, then $M(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$ is the set of matrix-valued measures μ on $\mathbf{R}$ for which $\int_{\mathbf{R}} \rho(s)\,d|\mu|(s) < \infty$. (In the scalar case $|\mu|$ is the total variation measure of μ, and in the matrix case we use a matrix norm when computing the total variation $|\mu|$ of μ.) The space $L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$ is defined analogously with measures replaced by measurable functions, and $V(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$ is the subspace of measures which one gets by adding point masses at zero to functions in $L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$. When we write $x \in L^p(\mathbf{R}; \mathbf{C}^n; \eta)$, we mean that x is defined on $\mathbf{R}$ with values in $\mathbf{C}^n$, and that ηx belongs to the ordinary nonweighted L^p-space on $\mathbf{R}$. A continuous function x belongs to $BUC(\mathbf{R}; \mathbf{C}^n; \eta)$ if $x \in L^\infty(\mathbf{R}; \mathbf{C}^n; \eta)$, and if $\tau_t x \to x$ in $L^\infty(\mathbf{R}; \mathbf{C}^n; \eta)$ as $t \to 0$, where τ_t is the translation operator $\tau_t x(s) = x(s+t)$, $s, t \in \mathbf{R}$. A special subclass of $BUC(\mathbf{R}; \mathbf{C}^n; \eta)$ is $BC_0(\mathbf{R}; \mathbf{C}^n; \eta)$ which consists of those functions x in $BUC(\mathbf{R}; \mathbf{C}^n; \eta)$ which satisfy $\lim_{|t|\to\infty} \eta(t)x(t) = 0$. We define $W^{m,p}(\mathbf{R}; \mathbf{C}^n; \eta)$, $BUC^m(\mathbf{R}; \mathbf{C}^n; \eta)$ and $BC_0^m(\mathbf{R}; \mathbf{C}^n; \eta)$ to be the set of functions which together with their first m derivatives belong to $L^p(\mathbf{R}; \mathbf{C}^n; \eta)$, $BUC(\mathbf{R}; \mathbf{C}^n; \eta)$ or $BC_0(\mathbf{R}; \mathbf{C}^n; \eta)$, respectively. In the preceding function space notation, when we replace $\mathbf{R}$ by $\mathbf{R}^+$ or $\mathbf{R}^-$, we mean the spaces of functions which one gets by restricting the functions in question to $\mathbf{R}^+$ or $\mathbf{R}^-$.

After these preliminaries, let us return to the original equation. Defining $\mathscr{L}$ to be the operator

$$\mathscr{L}x = x' + \mu * x, \tag{2.4}$$

where x' is the distribution derivative of x, we can write (1.1) in the following form:

$$\begin{aligned} \mathscr{L}x(t) &= f(t), && t \in \mathbf{R}^+, \\ x(t) &= \phi(t), && t \in \mathbf{R}^-. \end{aligned} \tag{2.5}$$

We let ρ be a given weight function on $\mathbf{R}$, and suppose that $\mu \in M(\mathbf{R}+; \mathbf{C}^{n\times n}; \rho)$, i.e., $\mu \in M(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$, and μ is supported on $\mathbf{R}^+$. We let η be an influence function dominated by ρ, let $\mathscr{B}$ be one of the spaces L^p, $1 \leqq p \leqq \infty$, BUC, or BC_0, and let $\mathscr{B}^{m+1}$ be one of the spaces $W^{m,p}$, BUC^m or BC_0^m. The functions ϕ and f in (2.5) are throughout assumed to satisfy $\phi \in \mathscr{B}^{m+1}(\mathbf{R}^-; \mathbf{C}^n; \eta)$ and $f \in \mathscr{B}^m(\mathbf{R}^+; \mathbf{C}^n; \eta)$ for some $m \geqq 0$.

We look for a solution of (2.5) which is locally in $\mathscr{B}^{m+1}$; in particular, we require x to be continuous at zero. A necessary condition for the existence of such a solution is that ϕ and f in (2.5) are compatible in the sense that if we define g by

$$g(t) = \begin{cases} \mathscr{L}\phi(t), & t < 0, \\ f(t), & t \geqq 0, \end{cases} \tag{2.6}$$

then g must belong to $\mathscr{B}^m$. This is true, because the operator $\mathscr{L}$ maps $\mathscr{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta)$ continuously into $\mathscr{B}^m(\mathbf{R}; \mathbf{C}^n; \eta)$. If $\mathscr{B} = L^p$, and $m = 0$, then this extra condition is automatically satisfied. If $\mathscr{B} = L^p$ and $m > 0$, then it is equivalent to

$$\mathscr{L}\phi^{(k)}(0) = f^{(k)}(0), \qquad 0 \leqq k \leqq m-1,$$

and if $\mathscr{B} = BUC$ or $\mathscr{B} = BC_0$, then it is equivalent to

$$\mathscr{L}\phi^{(k)}(0) = f^{(k)}(0), \qquad 0 \leqq k \leqq m.$$

From now on, we shall only consider data (ϕ, f) satisfying this extra condition, and choose our *state space* $\mathscr{D}$ to be

$$\mathscr{D} = \{(\phi, f) \in \mathscr{B}^{m+1}(\mathbf{R}^-; \mathbf{C}^n; \eta) \times \mathscr{B}^m(\mathbf{R}^+; \mathbf{C}^n; \eta) \,|\, \text{the function } g \text{ in (2.6) belongs to } \mathscr{B}^m(\mathbf{R}; \mathbf{C}^n; \eta)\}.$$

This is not the only possible choice. One could also restrict $\mathscr{D}$ further, and get a slightly different set of results. We shall return to that question elsewhere.

For each point $(\phi, f) \in \mathscr{D}$, we denote the solution of (2.5) which corresponds to this pair of functions by $x(\phi, f)$.

Let r be the fundamental solution of (2.5), i.e., the unique right continuous solution of the equation

$$r'+r*\mu=r'+\mu*r=\delta I \tag{2.7}$$

which vanishes on $(-\infty, 0)$. Here r' is the distribution derivative of r, δ is the unit point mass at zero, and I is the identity matrix. This function r is locally absolutely continuous with the exception of a jump discontinuity at zero, and it grows at most exponentially at infinity; see e.g. [10, Thm. 5.2].

The solution $x=x(\phi,f)$ of (2.5) can be written in the form

$$x(\phi,f)(t)=\begin{cases}\phi(t), & t\in\mathbf{R}^-,\\ r(t)\phi(0)+r*(f+M\phi)(t), & t\in\mathbf{R}^+,\end{cases} \tag{2.8}$$

where

$$M\phi(t)=-\int_{(t,\infty)} d\mu(s)\phi(t-s), \qquad t\in\mathbf{R}^+, \tag{2.9}$$

and we interpret f and $M\phi$ to be zero on $(-\infty, 0)$ in the definition of $r*(f+M\phi)$. See e.g. [12].

When one wants to know how the solutions of (2.5) behave asymptotically one has to look at eigenvalues of the formal Laplace transform

$$\hat{L}(z)=zI+\hat{\mu}(z) \tag{2.10}$$

of the operator $\mathscr{L}$ in (2.4). This function is well defined for $\Re z\geqq\omega$, where

$$\omega=-\lim_{t\to\infty} t^{-1}\log\rho(t)$$

is the exponential order of decay of ρ at infinity (cf. [4]). Throughout the sequel we assume that $\omega\leqq 0$, so that $\hat{L}(z)$ is defined in the closed right half plane. The importance of the eigenvalues of $\hat{L}$ becomes obvious once one realizes that the Laplace transform $\hat{r}$ of r in (2.7) is $\hat{r}(z)=[\hat{L}(z)]^{-1}$ for $\Re z$ sufficiently large. Roughly speaking, those eigenvalues of $\hat{L}(z)$ which belong to the open right half plane give rise to exponentially growing solutions of (1.1), and those on the imaginary axis give rise to solutions which neither grow exponentially, nor die out. If $\omega<0$, then it is possible that $\hat{L}(z)$ also has eigenvalues in the open left half plane, but these eigenvalues are less interesting in the sense that the solutions of (1.1) which they produce die out exponentially as $t\to\infty$. Accordingly, we shall more or less ignore the open left half plane, and we classify the eigenvalues of $\hat{L}(z)$ in the closed right half plane as *unstable* or *central* depending on whether they belong to the open right half plane or to the imaginary axis, respectively.

Even though the condition $\omega\leqq 0$ is sufficient to imply that $\hat{L}(z)$ is defined in the closed right half plane, it is not yet sufficient to imply that the solutions of (1.1) behave asymptotically in the way in which we would like. One also needs a growth condition at infinity on f and $M\phi$ in (2.8). To see this one can, e.g., transform (2.8) to get

$$\hat{x}(z)=\hat{r}(z)(\phi(0)+\hat{f}(z)+(M\phi)^{\wedge}(z)). \tag{2.11}$$

It is important to us that $\hat{f}(z)$ and $(M\phi)^{\wedge}(z)$ are analytic in the open right half plane. To insure this we shall impose a growth restriction on ρ at minus infinity, and suppose that $\alpha\leqq 0$, where $\alpha=-\lim_{t\to-\infty} t^{-1}\log\rho(t)$ is the exponential order of decay of ρ at minus infinity. It follows from (2.3) that the condition $\alpha\leqq 0$ will indeed imply that $\hat{f}(z)$ and $(M\phi)^{\wedge}(z)$ are analytic in the open right half plane. It is always true that $\omega\leqq\alpha$; hence the new assumption $\alpha\leqq 0$ implies the earlier one $\omega\leqq 0$.

Above we have divided the eigenvalues of $\hat{L}$ in the closed right half plane into two classes: The set of unstable eigenvalues and the set of central eigenvalues. From a stability point of view, this division is very appropriate, but from a technical point of view, the line $\Re z=0$ is no different from any other vertical line in the complex plane. In [10] and [11] all the decomposition arguments were made with the imaginary axis replaced by an arbitrary vertical line. A more technical division of the eigenvalues of $\hat{L}(z)$, which in spirit is more similar to the approach in [10] and in [11], would be to classify the eigenvalues as critical or noncritical depending on whether they belong to the critical region $\omega \leqq \Re z \leqq \alpha$ or to the noncritical region $\Re z > \alpha$. Roughly speaking, the eigenvalues on the line $\Re z = \omega$ are those which are most difficult to handle (none of $\hat{L}, \hat{f}$ and $(M\phi)^{\wedge}$ is analytic on this line), those in the strip $\omega < \Re z \leqq \alpha$ are somewhat easier to deal with (here $\hat{L}$ is analytic but not $\hat{f}$ and $(M\phi)^{\wedge}$), and eigenvalues in $\Re z > \alpha$ cause virtually no technical difficulties ($\hat{L}, \hat{f}$ and $(M\phi)^{\wedge}$ are all analytic). In [10] and [11] only noncritical eigenvalues are allowed, and the chief improvement here is that we are also able to treat critical eigenvalues. Going back to the original decomposition of the eigenvalues into central and unstable eigenvalues, this means that here we are able to deal with the case when $\alpha=0$ and $\hat{L}(z)$ has central eigenvalues, and even with the case when $\omega=\alpha=0$ and $\hat{L}(z)$ has central eigenvalues.

3. The central-stable and the unstable subspaces. As we observed earlier, the eigenvalues in the closed right half plane of $\hat{L}$ in (2.8) are of crucial importance for the asymptotic behavior of the solutions of (2.5). As $|z|\to\infty$, also $|\det(\hat{L}(z))|\to\infty$, so the set of all eigenvalues of $\hat{L}$ is bounded. Because of the analyticity of $\hat{L}$ in $\Re z>\omega$, the eigenvalues can only accumulate at the line $\Re z=\omega$. Throughout the sequel we shall suppose that the open right half plane contains only finitely many eigenvalues. It follows from the preceding discussion that this assumption is automatically satisfied if $\omega<0$. Later on we shall also assume that there are at most finitely many central eigenvalues, but for the moment this assumption is not needed.

In this section we intend to show that the state space $\mathscr{D}$ can be divided into an unstable subspace $\mathscr{U}$, and a central-stable subspace $\mathscr{CS}$. Here we use the same technique as is used in [10] and [11], and a reader familiar with [10] and [11] should have no difficulty in extending the results presented here to the neutral equation (and to allow infinitely many unstable eigenvalues of the same type as in [10] and [11]).

Let Z_U be the set of all eigenvalues z_l of $\hat{L}$ satisfying $\Re z_l>0$, and suppose that Z_U is a finite (possibly empty) set. If $Z_U=\emptyset$, then define $\lambda=\infty$, $\kappa=0$; otherwise, define

$$\lambda=\min\{\Re z_l | z_l\in Z_U\}, \qquad \kappa=\max\{\Re z_l | z_l \in Z_U\}. \tag{3.1}$$

DEFINITION 3.1. A point (ϕ,f) in $\mathscr{D}$ belongs to the *central-stable subspace* $\mathscr{CS}$, if the solution $x(\phi,f)$ of (2.5) satisfies $x(\phi,f)(t)=o(e^{\lambda t})$ as $t\to\infty$. A point $(\phi,0)$ in $\mathscr{D}$ belongs to the *unstable subspace* $\mathscr{U}$ if $\mathscr{L}\phi(t)=0$ for $t\in\mathbf{R}^-$, and $\phi(t)=O(e^{-|\varepsilon t|})$ as $t\to-\infty$, for some $\varepsilon>0$.

Here we consider the growth condition $x(\phi,f)=o(e^{\lambda t})$ as $t\to\infty$ to be vacuously satisfied if $\lambda=\infty$. Also, $\mathscr{L}\phi(t)$ is defined in the obvious way, namely $\mathscr{L}\phi(t)=\phi'(t)+\mu*\phi(t)$ for $t\in\mathbf{R}^-$ (as μ vanishes on $(-\infty,0)$, this definition makes sense).

As we already mentioned above, we want to show that $\mathscr{D}=\mathscr{CS}\oplus\mathscr{U}$. To do this we proceed as follows. Fix two constants γ and ξ, $0<\gamma<\lambda$, and $\xi>\kappa$. Define three new dominating functions ρ^C, ρ^U and ρ_U and a new influence function η_U by

$$\begin{aligned} &\rho^C(t)=e^{-\gamma t}, && t\in\mathbf{R},\\ &\rho^U(t)=e^{-\xi t}, && t\in\mathbf{R},\\ &\rho_U(t)=\max\{\rho^C(t),\rho^U(t)\}, && t\in\mathbf{R},\\ &\eta_U(t)=\min\{\rho^C(t),\rho^U(t)\}, && t\in\mathbf{R}. \end{aligned} \tag{3.2}$$

By Theorem 5.2 of [10], we have $r \in L^1(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^U)$. It follows from (2.7) and Lemma 2.1 of [10] that $r' - \delta I \in L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho^U)$; therefore $r \in W^{1,1}(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^U)$. By Lemma 3.7 of [10], $r \in BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^U)$. Also, by Theorem 5.1 of [10], there is a unique right continuous function $r_{CS} \in W^{1,1}((-\infty, 0); \mathbf{C}^{n\times n}; \rho^C) \cap W^{1,1}(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^C)$ (the comment above on the smoothness of r also applies here), which satisfies (2.7) with r replaced by r_{CS}. It follows again from Lemma 3.7 of [10] that $r_{CS} \in BC_0((-\infty, 0); \mathbf{C}^{n\times n}; \rho^C) \cap BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^C)$. Define $d_U = r - r_{CS}$. Then $d_U \in BC_0(\mathbf{R}; \mathbf{C}^{n\times n}; \eta_U)$ (the two jump discontinuities of r and r_{CS} at zero cancel each other). (A more explicit description of d_U is given in § 6.)

The function d_U defined above plays a crucial role in our decomposition of $\mathscr{D}$ into $\mathscr{U} \oplus \mathscr{CS}$, and it will be very important that r_{CS} and d_U can be convolved with the function $f + M\phi$ appearing in (2.8). That this is the case one can see as follows: We know that $f + M\phi$ in (2.8) belongs to $\mathscr{B}(\mathbf{R}^+; \mathbf{C}^n; \eta)$ when $\mathscr{B} = L^p$ for some p, $1 \leqq p \leqq \infty$, and to $L^\infty(\mathbf{R}^+; \mathbf{C}^n; \eta)$ when $\mathscr{B} = BUC$ or $\mathscr{B} = BC_0$. As $\gamma > 0 \geqq \alpha$, this implies that $f + M\phi \in L^1(\mathbf{R}^+; \mathbf{C}^n; \rho_U)$. If we define both f and $M\phi$ to vanish on $(-\infty, 0)$ then $f + M\phi \in L^1(\mathbf{R}; \mathbf{C}^n; \rho_U)$. Now, ρ^C and η_U are both influence functions dominated by ρ_U, and therefore, by Lemma A.1 in Appendix A, $r_{CS} * (f + M\phi)$ and $d_U * (f + M\phi)$ are well defined, $r_{CS} * (f + M\phi) \in BC_0(\mathbf{R}; \mathbf{C}^n; \rho^C)$, and $d_U * (f + M\phi) \in BC_0(\mathbf{R}; \mathbf{C}^n; \eta_U)$.

For each point $(\phi, f) \in \mathscr{D}$, we define $y(\phi, f)$ and $z(\phi, f)$ by

$$y(\phi,f)(t) = \begin{cases} \phi(t) + r_{CS}(t)\phi(0) + r_{CS} * (f + M\phi)(t), & t < 0, \\ r_{CS}(t)\phi(0) + r_{CS} * (f + M\phi)(t), & t \geqq 0, \end{cases} \tag{3.3}$$

$$z(\phi,f)(t) = d_U(t)\phi(0) + d_U * (f + M\phi)(t), \qquad t \in \mathbf{R}.$$

Then, by (2.8), $x(\phi, f) = y(\phi, f) + z(\phi, f)$.

THEOREM 3.1. *Given $(\phi, f) \in \mathscr{D}$, define $M\phi$ by (2.9), and $y(\phi, f)$ and $z(\phi, f)$ by (3.3), and let ψ and ζ be the restrictions of $y(\phi, f)$ and $z(\phi, f)$, respectively, to $\mathbf{R}^-$. Let P_{CS} be the operator which maps (ϕ, f) into (ψ, f), and let P_U be the operator which maps (ϕ, f) into $(\zeta, 0)$. Then P_{CS} and P_U are continuous projections in $\mathscr{D}$ with ranges $\mathscr{CS}$ and $\mathscr{U}$, respectively, and $P_{CS} + P_U = I$.*

The proof of Theorem 3.1 depends upon two lemmas. In the first the unstable subspace $\mathscr{U}$ is described in a different and much more explicit way. For each unstable eigenvalue $z_l \in Z_U$, we let $\mathscr{N}_l$ consist of the zero function together with all functions of the form

$$\sum_{i=0}^{p-1} \frac{t^i}{i!} r_{p-1-i}\, e^{z_l t}, \qquad t \in \mathbf{R}, \tag{3.4}$$

where $r_0 \neq 0$ and $r_0, \cdots, r_{p-1}$ is a right Jordan chain of $\hat{L}$ at z_l (see [4, Definition 4.1]).

LEMMA 3.1. *A point $(\phi, 0)$ in $\mathscr{D}$ belongs to $\mathscr{U}$ if and only if ϕ is the restriction to $\mathbf{R}^-$ of a function $x \in \oplus_{z_l \in Z_U} \mathscr{N}_l$, or equivalently, if and only if ϕ is the restriction to $\mathbf{R}^-$ of a function $x \in BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_U)$ satisfying $\mathscr{L}x = 0$. Moreover, $\mathscr{CS} \cap \mathscr{U} = \{(0, 0)\}$.*

Lemma 3.1 has one especially important consequence, namely, it implies that the dimension of $\mathscr{U}$ is finite, and equal to the sum of the algebraic orders of the unstable eigenvalues of $\hat{L}$. This is true because by the discussion in [4, § 4], this is the dimension of $\oplus_{z_l \in Z_U} \mathscr{N}_l$, and because the correspondence between the initial function ϕ and $x(\phi, 0)$ is one to one.

Proof. That a function $x \in BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_U)$ satisfies $\mathscr{L}x = 0$ if and only if $x \in \oplus_{z_l \in Z_U} \mathscr{N}_l$ follows from Theorem 5.1 of [10]. Thus, the two different characterizations of the functions x in Lemma 3.1 are equivalent.

It is clear that if ϕ is of the type mentioned in Lemma 3.1, then $(\phi, 0) \in \mathscr{U}$.

Conversely, suppose that $(\phi, 0) \in \mathscr{U}$. We know that $r \in BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \eta_U)$, that r vanishes on $(-\infty, 0)$, and that $f + M\phi \in L^1(\mathbf{R}; \mathbf{C}^n; \rho_U)$. Hence, by (2.8) and Lemma A.1 in Appendix A, we have $x(\phi, 0) \in BC_0(\mathbf{R}^+; \mathbf{C}^n; \eta_U)$. If we choose γ in (3.2) to be so small that $\phi(t) = O(e^{\varepsilon t})$, $t \to -\infty$, for some $\varepsilon > \gamma$, then $x(\phi, 0) \in BC_0(\mathbf{R}^-; \mathbf{C}^n; \eta_U)$. Thus, since $x(\phi, 0)$ is continuous at zero, $x(\phi, 0) \in BC_0(\mathbf{R}; \mathbf{C}^n; \eta_U)$. Moreover, as $\mathscr{L}\phi(t) = 0$ for $t \in \mathbf{R}^-$, and $x(\phi, 0)$ satisfies (2.5) with $f = 0$, we have $\mathscr{L}x(\phi, 0) = 0$. Explicitly, the condition $\mathscr{L}x(\phi, 0) = 0$ means that $(x(\phi, 0))' = -\mu * (\phi, 0)$, and this together with Lemma 2.1 of [10] gives $(x(\phi, 0))' \in BC_0(\mathbf{R}; \mathbf{C}^n; \eta_U)$. Thus, ϕ is the restriction to $\mathbf{R}^-$ of the function $x(\phi, 0)$, which belongs to $BC_0^1(\mathbf{R}; \mathbf{C}^n; \eta_U)$ and satisfies $\mathscr{L}x(\phi, 0) = 0$.

The fact that $\mathscr{CS} \cap \mathscr{U} = \{(0, 0)\}$ follows immediately from the growth rate at infinity imposed on $x(\phi, f)$ in the case when $(\phi, f) \in \mathscr{CS}$ together with the fact that $x(\phi, 0) \in \oplus_{z_l \in Z_U} \mathscr{N}_l$ when $(\phi, 0) \in \mathscr{U}$. □

LEMMA 3.2. *The range of P_{CS} is contained in $\mathscr{CS}$, and the range of P_U is contained in $\mathscr{U}$.*

Proof. Define ψ and ζ as in Theorem 3.1. We have to show that $(\psi, f) \in \mathscr{CS}$, and that $(\zeta, 0) \in \mathscr{U}$.

Let us begin with the claim that $(\zeta, 0) \in \mathscr{U}$. It follows from the discussion in the paragraph containing (3.2) that $d_U \in BC_0(\mathbf{R}; \mathbf{C}^{n\times n}; \eta_U)$. Moreover, as both r and r_{CS} satisfy (2.7), and $d_U = r - r_{CS}$, we have $\mathscr{L}d_U = 0$. In the same way as in the preceding proof we conclude that $d_U \in BC_0^1(\mathbf{R}; \mathbf{C}^{n\times n}; \eta_U)$. Recall that $f + M\phi \in L^1(\mathbf{R}; \mathbf{C}^n; \rho_U)$, and use (3.3) and Lemma 3.5 of [10] to get

$$\begin{aligned}\mathscr{L}z(\phi, f) &= \mathscr{L}(d_U\phi(0)) + \mathscr{L}(d_U * (f + M\phi)) \\ &= (\mathscr{L}d_U)\phi(0) + (\mathscr{L}d_U) * (f + M\phi) = 0.\end{aligned}$$

It follows from (3.3) and Lemma 3.5 of [10] that $z(\phi, f) \in BC_0^1(\mathbf{R}; \mathbf{C}^n; \eta_U)$. By Lemma 3.1, $(\zeta, 0) \in \mathscr{U}$.

It remains to show that $(\psi, f) \in \mathscr{CS}$. We know from the preceding argument that $z(\phi, f) \in \mathscr{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta_U)$, $\mathscr{L}z(\phi, f) = 0$, and $y(\phi, f) = x(\phi, f) - z(\phi, f)$, so $y(\phi, f)$ belongs locally to $\mathscr{B}^{m+1}$, and $y(\phi, f)$ is the solution of (2.5) with ϕ replaced by ψ (in other words, $y(\phi, f) = x(\psi, f)$). Therefore, to show that $(\psi, f) \in \mathscr{CS}$, it suffices to prove that $y(\phi, f)(t) = o(e^{\lambda t})$ as $t \to \infty$. This, however, follows directly from (3.3), because we already know that $r_{CS} \in BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^C)$ and $r_{CS} * (f + M\phi) \in BC_0(\mathbf{R}^+; \mathbf{C}^n; \rho^C)$; hence, $y(\phi, f)(t) = o(e^{\gamma t})$ as $t \to \infty$, which is a stronger statement than the needed $y(\phi, f)(t) = o(e^{\lambda t})$ as $t \to \infty$. □

Let us formally record the fact which became apparent at the end of the preceding proof:

COROLLARY 3.1. *If $(\phi, f) \in \mathscr{CS}$, then $x(\phi, f)(t) = o(e^{\varepsilon t})$ as $t \to \infty$ for every $\varepsilon > 0$.*

If one strengthens the assumptions on μ, f and ϕ, then this growth estimate can be further improved. See §§ 4 and 6.

Proof of Theorem 3.1. By now the proof of Theorem 3.1 is trivial, i.e., Theorem 3.1 is a direct consequence of Lemmas 3.1 and 3.2 (if $P_{CS} + P_U = I$, the range of P_{CS} is contained in $\mathscr{CS}$, the range of P_U is contained in $\mathscr{U}$, and $\mathscr{CS} \cap \mathscr{U} = \{(0, 0)\}$, then necessarily P_{CS} and P_U are projections, and their ranges are the claimed ones). □

4. The stable and the central subspaces in the noncritical case. In the preceding section we subtracted off an exponentially growing part from the solution, and got a remainder in $\mathscr{CS}$ which grows slower than exponentially as $t \to \infty$. If $\alpha < 0$, then one can use exactly the same technique to subtract off another part of the solution, which neither grows exponentially, nor tends to zero as $t \to \infty$. The new remainder tends to zero exponentially as $t \to \infty$. The technique is exactly the same as in § 3, and therefore we shall state the results, but leave the proofs to the reader.

Throughout this section we assume that $\alpha < 0$ (except in Proposition 4.1). Let Z_S be the (possibly empty, possibly infinite) set of eigenvalues z_l of $\hat{L}$ satisfying $\alpha < \Re z_l < 0$, and let Z_C be the set of eigenvalues of $\hat{L}$ on the imaginary axis. Define λ as in § 3. If $Z_S = \emptyset$, then define $\iota = \alpha$; otherwise define

$$\iota = \max \{\Re z_l | z_l \in Z_S\}. \tag{4.1}$$

Definition 4.1. A point (ϕ, f) in $\mathscr{D}$ belongs to the stable subspace $\mathscr{S}$, if the solution $x(\phi, f)$ of (2.5) satisfies $x(\phi, f)(t) \to 0$ as $t \to \infty$. A point $(\phi, 0)$ in $\mathscr{D}$ belongs to the *central subspace* $\mathscr{C}$ if $\mathscr{L}\phi(t) = 0$ for $t \in \mathbf{R}^-$, $\phi(t) = O(e^{(\iota+\varepsilon)t})$ as $t \to -\infty$, for some $\varepsilon > 0$, and $x(\phi, 0)(t) = o(e^{\lambda t})$ as $t \to \infty$.

Clearly, both $\mathscr{S}$ and $\mathscr{C}$ are subspaces of $\mathscr{C}\mathscr{S}$.

We want to show that $\mathscr{C}\mathscr{S} = \mathscr{S} \oplus \mathscr{C}$. To do this we proceed as in § 3. We fix two constants ν and γ, satisfying $\iota < \nu < 0$ and $0 < \gamma < \lambda$. Define three dominating functions ρ^S, ρ^C and ρ_C, and a new influence function η_C by

$$\begin{aligned} \rho^S(t) &= e^{-\iota t}, && t \in \mathbf{R}, \\ \rho^C(t) &= e^{-\gamma t}, && t \in \mathbf{R}, \\ \rho_C(t) &= \max \{\rho^S(t), \rho^C(t)\}, && t \in \mathbf{R}, \\ \eta_C(t) &= \min \{\rho^S(t), \rho^C(t)\}, && t \in \mathbf{R}. \end{aligned} \tag{4.2}$$

By Theorem 5.1 of [10], we can find a unique right continuous solution $r_S \in W^{1,1}((-\infty, 0); \mathbf{C}^{n \times n}; \rho^S) \cap W^{1,1}(\mathbf{R}^+; \mathbf{C}^{n \times n}; \rho^S)$ to the resolvent equation (2.7). Define $d_C = r_{CS} - r_S$. Then $d_C \in BC_0(\mathbf{R}; \mathbf{C}^{n \times n}; \eta_C)$.

For each point $(\phi, f) \in \mathscr{D}$, we define $v(\phi, f)$ and $w(\phi, f)$ by

$$\begin{aligned} v(\phi, f)(t) &= \begin{cases} \phi(t) + r_S(t)\phi(0) + r_S * (f + M\phi)(t), & t \leqq 0, \\ r_S(t)\phi(0) + r_S * (f + M\phi)(t), & t \geqq 0, \end{cases} \\ w(\phi, f)(t) &= d_C(t)\phi(0) + d_C * (f + M\phi)(t), \qquad t \in \mathbf{R}. \end{aligned} \tag{4.3}$$

Then, by (3.3), $v(\phi, f) + w(\phi, f) = y(\phi, f)$.

Theorem 4.1. *Given $(\phi, f) \in \mathscr{D}$, define $v(\phi, f)$ and $w(\phi, f)$ by* (4.3), *and let θ and χ be the restrictions of $v(\phi, f)$ and $w(\phi, f)$, respectively, to $\mathbf{R}^-$. Let P_S be the operator which maps (ϕ, f) into (θ, f), and let P_C be the operator which maps (ϕ, f) into $(\chi, 0)$. Then P_S and P_C are continuous projections in $\mathscr{D}$ with ranges $\mathscr{S}$ and $\mathscr{C}$, respectively, and $P_S + P_C = P_{CS}$.*

For each $z_l \in Z_C$ we define $\mathscr{N}_l$ in the same way as in the paragraph containing (3.4). Analogously to Lemma 3.1 we have Lemma 4.1.

Lemma 4.1. *A point $(\phi, 0)$ in $\mathscr{D}$ belongs to $\mathscr{C}$ if and only if ϕ is the restriction to $\mathbf{R}^-$ of a function $x \in \oplus_{z_l \in Z_C} \mathscr{N}_l$, or equivalently, if and only if ϕ is the restriction to $\mathbf{R}^-$ of a function $x \in BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_C)$ satisfying $\mathscr{L}x = 0$.*

Corollary 3.1 has the following analogue:

Corollary 4.1. *If $(\phi, f) \in \mathscr{S}$, then $x(\phi, f)(t) = O(e^{(\iota+\varepsilon)t})$ as $t \to \infty$ for every $\varepsilon > 0$.*

If $\alpha < 0$ and $\hat{L}$ has no central eigenvalues, then by Lemma 4.1, $\mathscr{C} = \{(0, 0)\}$, and for every $(\phi, f) \in \mathscr{C}\mathscr{S}$, the solution $x(\phi, f)$ of (2.5) tends exponentially to zero as $t \to \infty$. A similar result is true also when $\alpha = 0$ and $\hat{L}$ has no critical eigenvalues:

Proposition 4.1. *Suppose that $\alpha \leqq 0$ and that $\hat{L}$ has no eigenvalues in the region $\omega \leqq \Re z \leqq 0$. Then, for each $(\phi, f) \in \mathscr{C}\mathscr{S}$, the solution $x(\phi, f)$ of* (2.5) *satisfies $x \in \mathscr{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta)$.*

Proposition 4.1 is essentially contained in Theorems 6.5 and 6.6 of [10]. For completeness we include a proof.

Proof. Fix $\gamma, 0<\gamma<\lambda$, and define

$$\rho_S(t)=\max\{\rho(t), e^{-\gamma t}\}, \qquad t\in \mathbf{R},$$

$$\eta_S(t)=[\rho_S(-t)]^{-1}, \qquad t\in \mathbf{R}.$$

Then η_S is an influence function dominated by ρ_S. It follows from Theorem 5.1 of [10] that $r_{CS}\in L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho_S)$ and $r'_{CS}\in M(\mathbf{R}; \mathbf{C}^{n\times n}; \rho_S)$.

Let $(\phi, f)\in \mathscr{CS}$. Then, by (2.3), (2.5) and Corollary 3.1, $x(\phi, f)\in \mathscr{B}(\mathbf{R}; \mathbf{C}^n; \eta_S)$. Moreover, $x(\phi, f)$ satisfies $\mathscr{L}x(\phi, f)=g$, where $g\in \mathscr{B}^m(\mathbf{R}; \mathbf{C}^n; \eta)$ is the function defined in (2.6). If we write this equation in the form $(x(\phi, f))'=g-\mu * x(\phi, f)$, and use (2.3) and Lemma 2.1 of [10], we get $(x(\phi, f))'\in \mathscr{B}(\mathbf{R}; \mathbf{C}^n; \eta_S)$; i.e., $x\in \mathscr{B}^1(\mathbf{R}; \mathbf{C}^n; \eta_S)$. Convolving the equation $\mathscr{L}x(\phi, f)=g$ with r_{CS}, and using (2.7), we obtain $x(\phi, f)=r_{CS} * g$. Here $r_{CS}\in L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho_S)\subset L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$ and $r'_{CS}\in M(\mathbf{R}; \mathbf{C}^{n\times n}; \rho_S)\subset M(\mathbf{R}; \mathbf{C}^{n\times n}; \rho)$, and $g\in \mathscr{B}^m(\mathbf{R}; \mathbf{C}^n; \eta)$; therefore, by Lemmas 3.5 and 3.6 of [10], $x(\phi, f)\in \mathscr{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta)$. □

5. The singular part expansion. In the critical case, i.e., in the case when $\alpha=0$, the argument in the preceding section must be modified. Two problems arise. First, the method used to construct r_S and d_C is no longer available, unless $\omega<0$. Second, even if $\omega<0$, it need no longer be true that $r_S * (f+M\phi)$ and $d_C * (f+M\phi)$ are well defined.

In the scalar case one can define r_S and d_C by an alternative method which is based on the L^1-remainder theorem for locally analytic functions [3, Thm. 3.6]. The same approach can also be used in the vector case provided that we have an analogue of Theorem 3.6 in [3] for matrix-valued locally analytic functions. We devote this section to the development of such a result.

Our treatment is general in the sense that we assume that ρ is any weight function on $\mathbf{R}$ with exponential orders of growth ω and α at plus and minus infinity, respectively. Throughout this section we assume that the reader has our recent paper [4] at hand. We refer to [4] for the significance of the region $\Pi=\{x|\omega\leqq \Re z\leqq \alpha\}$, the definitions of a zero, a locally analytic zero, smoothness of order m, a local Smith factorization and partial multiplicity, and for the factorization theory for locally analytic matrix functions developed in that paper.

DEFINITION 5.1. Let M be an n by n matrix-valued function, which is locally analytic at $z_0\in\Pi$. If M^{-1} has the form

$$M^{-1}(z)=\sum_{i=1}^{p-1} K_i(z-z_0)^{i-p}+\zeta(z), \qquad z\in U\backslash\{z_0\}, \tag{5.1}$$

where each K_i is an n by n matrix, $K_0\neq 0, \zeta$ is locally analytic at z_0, and U is a neighborhood (relative to Π) of z_0, then we say that M^{-1} has a singular part expansion at z_0 of length p, and call $\sum_{i=1}^{p-1} K_i(z-z_0)^{i-p}$ the singular part of M^{-1}.

As one would expect from the scalar theory, (cf. [3, Thms. 3.4, 3.6]), the property that M^{-1} has a singular part expansion at z_0 is stronger than the property that M has a local Smith factorization at z_0:

PROPOSITION 5.1. *Let the n by n matrix function M be locally analytic at $z_0\in\Pi$, and let M^{-1} have a singular part expansion at z_0 of length p. Then M has a local Smith factorization at z_0, and the maximal partial multiplicity σ of M at z_0 is equal to p.*

Proof. By Theorem 4.4 of [1], one can construct a matrix function T that is analytic at z_0 such that T^{-1} and M^{-1} have the same singular part at z_0. In other words, if M^{-1} satisfies (5.1), then

$$T^{-1}(z)=\sum_{i=0}^{p-1} K_i(z-z_0)^{i-p}+\xi(z), \qquad z\in U\backslash\{z_0\}, \tag{5.2}$$

with ξ analytic at z_0. Clearly,

$$M^{-1}(z)=T^{-1}(z)-\xi(z)+\zeta(z), \qquad z\in U\backslash\{z_0\},$$

so that the equations

$$T^{-1}(z)M(z)=I-(\zeta(z)-\xi(z))M(z)$$

and

$$M^{-1}(z)T(z)=I+(\zeta(z)-\xi(z))T(z)$$

both hold for all $z\in U\backslash\{z_0\}$. The right-hand sides of the last two equalities are locally analytic at z_0, and they are inverses of each other for $z\neq z_0$. Thus, if we define

$$S(z)=I-(\zeta(z)-\xi(z))M(z), \qquad z\in U,$$

then S is locally analytic and invertible at z_0, and

$$M(z)=T(z)S(z), \qquad z\in U. \tag{5.3}$$

By, e.g., Theorem 3.1 of [4], the analytic matrix function T has a left local Smith factorization PDR at z_0, and, since $S(z_0)$ is invertible, $PDRS$ is a left local Smith factorization of M at z_0. The existence of a right local Smith factorization is proved in a completely analogous way.

The claim that $p=\sigma$ follows directly from the corresponding property in the analytic case, the fact that (by the construction of T) T^{-1} and M^{-1} have the same singular part expansion at z_0, and the fact that M and T have the same maximal partial multiplicity at z_0 (by Lemma 3.1 of [4] since S in (5.3) is invertible at z_0). □

Proposition 5.1 tells us that a necessary condition for M^{-1} to have a singular part expansion at z_0 is that M has a local Smith factorization at z_0. If we require the locally analytic factor in this factorization to be sufficiently smooth, then we get a sufficient condition for the existence of a singular part expansion of M^{-1} at z_0.

THEOREM 5.1. *Let the n by n matrix function M have a left local Smith factorization at $z_0\in\Pi$ with maximal partial multiplicity $\sigma>0$. In addition, assume that each factor in this factorization is smooth of order σ at z_0. Then M^{-1} has a singular part expansion at z_0.*

Of course, the analogue of Theorem 5.1 in which left factorization is replaced by right factorization is also valid.

Proof. By the hypothesis, M can be written as $M(z)=P(z)D(z)R(z)$ in a neighborhood of z_0, where P is a unimodular quasipolynomial,

$$D(z)=\operatorname{diag}\left\{\left(\frac{z-z_0}{z-c}\right)^{\kappa_1},\dots,\left(\frac{z-z_0}{z-c}\right)^{\kappa_n}\right\}$$

with $0\leqq\kappa_1\leqq\cdots\leqq\kappa_n=\sigma$, and where R is locally analytic and smooth of order σ at z_0 with $\det R(z_0)\neq 0$. Here c is a fixed constant with $\Re z<\omega$ [4, Definition 2.1]. Since R is smooth of order σ, Lemma 2.1 of [4] yields that R has the form $R(z)=Q(z)+(z-z_0)^\sigma\chi(z)$, where Q is a polynomial of degree at most $\sigma-1$ with $\det Q(z_0)\neq 0$, and χ is locally analytic at z_0. It follows that in some neighborhood U of z_0,

$$\begin{aligned} R^{-1}(z)-Q^{-1}(z)&=-R^{-1}(z)(z-z_0)^\sigma\chi(z)Q^{-1}(z)\\ &=-(z-z_0)^\sigma R^{-1}(z)\chi(z)Q^{-1}(z), \qquad z\in U. \end{aligned}$$

Thus, M^{-1} can be written as

$$\begin{aligned} M^{-1}(z)&=R^{-1}(z)D^{-1}(z)P^{-1}(z)\\ &=Q^{-1}(z)D^{-1}(z)P^{-1}(z)\\ &\quad -R^{-1}(z)\chi(z)Q^{-1}(z)(z-z_0)^\sigma D^{-1}(z)P^{-1}(z), \qquad z\in U. \end{aligned} \tag{5.4}$$

Now PDQ is analytic at z_0, and therefore $Q^{-1}D^{-1}P^{-1}$ has a singular part expansion at z_0 (cf. [1, p. 98]). Since the second term on the right-hand side of (5.4) is clearly locally analytic at z_0, M^{-1} has a singular part expansion at z_0. □

As an immediaate consequence of Theorem 5.1 combined with the factorization Theorem 3.1 of [4] (the localized matrix version of Theorem 3.4 of [3]), we obtain for matrix functions the following localized analogue of the scalar L^1-remainder theorem [3, Thm. 3.6].

COROLLARY 5.1. *Let the n by n matrix function M be locally analytic at* $z_0\in\Pi$, *and assume that* det M *has a zero of integral order* $k\geqq 0$ *at* z_0. *If* $n>1$, *let* $\sigma=\sigma(M)$ *be the smallest nonnegative integer such that every minor* Δ *of M of order* $n-1$ *has a zero of order at least* $k-\sigma$ *at* z_0; *in the scalar case* $n=1$ *set* $\sigma=k$. *If M is smooth of order* 2σ *at* z_0, *then* M^{-1} *has a singular part expansion at* z_0.

Proof. By Theorem 3.1 of [4], M has a (left) local Smith factorization $M=PDR$ in a neighborhood U of z_0 with maximal partial multiplicity equal to σ. Since $R(z)=D^{-1}(z)P^{-1}(z)M(z)$ for $z\in U\backslash\{z_0\}$ and M is smooth of order 2σ, it follows from Lemma 2.2 of [4] that the locally analytic factor R is smooth of order σ at z_0. Corollary 5.1 then follows from Theorem 5.1. □

We remark that the smoothness assumption on M in Corollary 5.1 is not necessary for M^{-1} to have a singular part expansion (compare with the corresponding discussion following the factorization Theorem 3.1 in [4]). For example, let $M(z)=\operatorname{diag}(\phi_1(z),(z-z_0))$, where ϕ_1 is locally analytic at z_0, and $\phi_1(z_0)\neq 0$. Then

$$M^{-1}(z)=\operatorname{diag}(0,(z-z_0)^{-1})+\operatorname{diag}(\phi_1^{-1}(z),0)$$

has a singular part expansion at z_0, but M is not smooth of order 2 at z_0 unless ϕ_1 is smooth of order 2 at z_0.

We conclude this section by showing that the singular part $\sum_{k=0}^{p-1}K_i(z-z_0)^{i-p}$ of M^{-1} is closely related to the left and right Jordan chains of M at z_0. In fact, the sequence $K_0, K_1, \cdots, K_{p-1}$ itself turns out to be both a left and a right Jordan chain of M at z_0 provided we define the notion of a matrix Jordan chain in the obvious way:

DEFINITION 5.2. Let the n by n matrix function M be locally analytic at $z_0\in\Pi$. If $K_0\neq 0$ and $M(z)\sum_{k=0}^{p-1}K_i(z-z_0)^i$ has a zero of order at least p at z_0, then we call $K_0, K_1, \cdots, K_{p-1}$ *a right* (*matrix*) *Jordan chain* of M at z_0. Similarly, if $K_0\neq 0$ and $\sum_{k=0}^{p-1}K_i(z-z_0)^iM(z)$ has a zero of order at least p at z_0, then we call $K_0, K_1, \cdots, K_{p-1}$ *a left* (*matrix*) *Jordan chain* of M at z_0.

Apart from the different setting with matrices instead of vectors, this definition is equivalent to Definition 4.1 in [4].

PROPOSITION 5.2. *Let the n by n matrix function M be locally analytic at* $z_0\in\Pi$, *and suppose that* M^{-1} *has a singular part expansion* $\sum_{k=0}^{p-1}K_i(z-z_0)^{i-p}$ *at* z_0. *Then* $K_0, K_1, \cdots, K_{p-1}$ *is both a left and a right Jordan chain of M at* z_0.

Proof. Define

$$R(z)=\sum_{k=0}^{p-1} K_i(z-z_0)^i.$$

Then, by the hypothesis, M^{-1} has the form

$$M^{-1}(z)=(z-z_0)^{-p}R(z)+\zeta(z), \qquad z\in U\backslash\{z_0\},$$

where U is a neighborhood of z_0, and ζ is locally analytic at z_0. Multiplying this equation from the left by $(z-z_0)^pM(z)$ we get

$$M(z)R(z)=(z-z_0)^p(I-M(z)\zeta(z)), \qquad z\in U.$$

By definition, this means that $K_0, K_1, \cdots, K_{p-1}$ is a right Jordan chain of M at z_0. The proof of the fact that $K_0, K_1, \cdots, K_{p-1}$ is also a left Jordan chain of M is completely analogous. □

6. The stable and central subspaces in the critical case. In this section we decompose the central-stable subspace $\mathscr{C}\mathscr{S}$ into a central subspace $\mathscr{C}$ and a stable subspace $\mathscr{S}$ in the critical case when $\alpha=0$. The crucial properties of the functions r_S and d_C in § 4 are that they have the appropriate growth rates at plus and minus infinity, and that $\mathscr{L}d_C=0$. In § 4 we first construct r_S, and then define d_C to be $r_{CS}-r_S$. Here we proceed in the opposite way, i.e., we first construct d_C, and then define $r_S=r_{CS}-d_C$. Our construction of d_C is based on the singular part expansion of $(\hat{L})^{-1}$ at the central eigenvalues of $\hat{L}$ in (2.7). In addition to being more general than the construction in § 3, it has another advantage, since it gives a much more explicit description of d_C (and also of the function d_U in § 3).

We begin with the following lemma.

LEMMA 6.1. *Let z_0 be an eigenvalue of $\hat{L}$, and suppose that $(\hat{L})^{-1}$ has a singular part expansion with singular part*

$$SP_{z_0}(z)=\sum_{i=0}^{p-1} K_i(z-z_0)^{i-p} \tag{6.1}$$

at z_0, and that

$$\int_{\mathbf{R}^+} t^{p-1}e^{-\gamma t}d|\mu|(t)<\infty, \tag{6.2}$$

where $\gamma=\Re z_0$. Then the function $d_{z_0}(t)$ defined by

$$d_{z_0}(t)=\sum_{i=0}^{p-1}\frac{t^i}{i!}K_{p-1-i}\,e^{z_0t}, \qquad t\in\mathbf{R}, \tag{6.3}$$

satisfies

$$d'_{z_0}+\mu * d_{z_0}=0=d'_{z_0}+d_{z_0}*\mu. \tag{6.4}$$

We remark that assumption (6.2) is automatically satisfied whenever $\gamma>\omega$.

Proof. We prove the left-hand equality; the proof of the right-hand equality is completely analogous.

By Proposition 5.2, $K_0, \cdots, K_{p-1}$ is a right Jordan chain of length p of $\hat{L}$ at z_0. Since (6.2) holds, $\hat{L}(z)$ is smooth of order $p-1$ at z_0 with respect to the dominating function $e^{-\gamma t}$, $t\in\mathbf{R}$, and a standard multiplication of series (cf. [4, (4.2)]) shows that

$$\sum_{i=0}^{j}\frac{\hat{L}^{(i)}(z_0)}{i!}K_{j-i}=0, \qquad 0\leqq j\leqq p-1. \tag{6.5}$$

Next, an elementary calculation that uses the binomial theorem and that is given in the proof of Lemma 5.2 of [4] yields that

$$\mu * d_{z_0}(t) = e^{z_0 t} \sum_{j=0}^{p-1} \frac{t^{p-1-j}}{(p-1-j)!} \sum_{i=0}^{j} \frac{\hat{\mu}^{(i)}(z_0)}{i!} K_{j-i}.$$

This calculation is valid since (6.2) holds. Using the last line and (6.5), we get after another easy calculation that

$$d'_{z_0}(t) + \mu * d_{z_0}(t) = e^{z_0 t} \sum_{j=0}^{p-1} \frac{t^{p-1-j}}{(p-1-j)!} \sum_{i=0}^{j} \frac{\hat{L}^{(i)}(z_0)}{i!} K_{j-i} = 0. \qquad \square$$

We now proceed with our construction of the desired function r_S. We let ρ^S be a positive, continuous, nondecreasing and submultiplicative function on $\mathbf{R}^+$ satisfying $\rho^S(0)=1$ and $\rho^S(t) \leqq \rho(t)$ for $t \in \mathbf{R}^+$ (this function will determine the rate of convergence to zero of the solutions corresponding to initial data in the stable subspace). If we extend ρ^S to all of $\mathbf{R}$ by defining $\rho^S(t)=1$ for $t<0$, then ρ^S is a weight function on $\mathbf{R}$. We assume that $\hat{L}(z)$ has no eigenvalues in $\omega^S \leqq \Re z < 0$, where $\omega^S = -\lim_{t\to\infty} t^{-1} \log \rho^S(t)$ (if $\omega^S = 0$, then this condition is vacuously satisfied). We also assume that $\hat{L}(z)$ has only a finite number of eigenvalues z_l in the closed right half plane, and let the sets

$$Z_C = \{z_l | \Re z_l = 0\}, \qquad Z_U = \{z_l | \Re z_l > 0\},$$

consist of the central and unstable eigenvalues, respectively. Let $Z = Z_C \cup Z_U$. Assume that $\hat{L}(z)$ has a singular part expansion *with respect to* ρ^S with singular part $SP_{z_l}(z) = \sum_{i=0}^{p_l-1} K_{l,i}(z-z_l)^{i-p_l}$ at each eigenvalue $z_l \in Z$. Of course, this assumption is automatically satisfied at all the unstable eigenvalues, and also at all the central eigenvalues if $\omega < 0$. Theorem 5.1 gives sufficient smoothness assumptions on $\hat{L}$ for this hypothesis to hold at central eigenvalues in the case when $\omega = 0$.

For each eigenvalue $z_l \in Z$, set

$$d_{z_l}(t) = \sum_{i=0}^{p_l-1} \frac{t^i}{i!} K_{l,p_l-1-i}\, e^{z_l t}, \qquad t \in \mathbf{R}.$$

Define the matrix-valued functions d_U, d_C and d by

$$d_U(t) = \sum_{z_l \in Z_U} d_{z_l}(t), \qquad d_C(t) = \sum_{z_l \in Z_C} d_{z_l}(t),$$

and $d = d_C + d_U$.

Before proceeding we first observe that the function d_U as defined above is the same as the function d_U previously defined in § 3, and, in the noncritical case, d_C as defined above is the same as the function d_C defined in § 4. The verifications of these two claims are completely similar, so let us only give the argument for d_U. Define ρ^U, r_{CS} and d_U in the same way as in § 3. We know that $r \in W^{1,1}(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^U)$, that r vanishes on $(-\infty, 0)$ and that $\hat{r}(z) = [\hat{L}(z)]^{-1}$ for $\Re z = \xi$. Therefore, for each $t \neq 0$ we can express $r(t)$ as an inverse Laplace integral

$$r(t) = \lim_{T\to\infty} \frac{1}{2\pi i} \int_{\xi-iT}^{\xi+iT} e^{zt} [\hat{L}(z)]^{-1}\, dz.$$

In the same way, one gets for each $t \neq 0$,

$$r_{CS}(t) = \lim_{T\to\infty} \frac{1}{2\pi i} \int_{\gamma-iT}^{\gamma+iT} e^{zt} [\hat{L}(z)]^{-1}\, dz,$$

where γ is the constant used in the definition of ρ^C in (3.2). We defined d_U to be $r-r_{CS}$; hence, for each $t\neq 0$,

$$d_U(t)=\lim_{T\to\infty}\frac{1}{2\pi i}\left(\int_{\xi-iT}^{\xi+iT}-\int_{\gamma-iT}^{\gamma+iT}\right)e^{zt}[\hat{L}(z)]^{-1}\,dz.$$

The function $[\hat{L}(z)]^{-1}$ tends to zero, uniformly in the strip $\gamma\leqq\Re z\leqq\xi$ as $\Im z\to\infty$. Therefore, if we define Γ to be the closed curve in the complex plane which is composed of four straight lines, and joins the points $\gamma-iT$, $\xi-iT$, $\xi+iT$, $\gamma+iT$ in a counterclockwise way, then, for $t\neq 0$,

$$d_U(t)=\lim_{T\to\infty}\frac{1}{2\pi i}\int_\Gamma e^{zt}[\hat{L}(z)]^{-1}\,dz.$$

By the residue theorem, the integral on the right-hand side is the sum of the residues of the meromorphic function inside the integral. Expanding e^{zt} into a power series around each unstable eigenvalue z_l of $\hat{L}$, and using the singular part expansion of $[\hat{L}(z)]^{-1}$ at z_l, one easily gets the desired formula.

Continuing with our construction of r_S, we let $r_{CS}=r-d_U$ (as in §§ 3 and 4), and define $r_S=r-d$. Let

$$\sigma=\max\{p_l|z_l\in Z_C\}$$

be the maximal partial multiplicity of the central eigenvalues of $\hat{L}$. At this stage we assume that $Z_C\neq\varnothing$; the case $Z_C=\varnothing$ will be discussed later.

Fix some constant γ, $0<\gamma<\lambda$, with λ defined as in § 3, and define

$$\begin{aligned}
\rho_S(t)&=\begin{cases}1, & t\in\mathbf{R}^-,\\ (1+t)^{\sigma-1}\rho^S(t), & t\in\mathbf{R}^+,\end{cases}\\
\eta_S(t)&=\begin{cases}(1+|t|)^{1-\sigma}, & t\in\mathbf{R}^-,\\ \rho^S(t), & t\in\mathbf{R}^+,\end{cases}\\
\rho^C(t)&=e^{-\gamma t},\qquad t\in\mathbf{R},\\
\rho_C(t)&=\max\{\rho^C(t),(1+|t|)^{\sigma-1}\},\qquad t\in\mathbf{R},\\
\eta_C(t)&=[\rho_C(-t)]^{-1},\qquad t\in\mathbf{R}.
\end{aligned}\tag{6.6}$$

LEMMA 6.2. *The function r_S satisfies $r_S\in BUC((-\infty,0);\mathbf{C}^{n\times n};\eta_S)\cap BC_0(\mathbf{R}^+;\mathbf{C}^{n\times n};\eta_S)$. In particular, $r_S(t)\to 0$ as $t\to\infty$.*

Proof. For $t<0$ we have $r_S(t)=-d(t)$, and the claim $r_S\in BUC((-\infty,0);\mathbf{C}^{n\times n};\eta_S)$ follows immediately. It remains to show that $r_S\in BC_0(\mathbf{R}^+;\mathbf{C}^{n\times n};\eta_S)$.

Define $d^+(t)=d(t)$ for $t\geqq 0$, $d^+(t)=0$ for $t<0$, and likewise $r_S^+(t)=r_S(t)$ for $t\geqq 0$, $r_S^+(t)=0$ for $t<0$. Then $r_S^+=r-d^+$, and it follows from elementary Laplace transform theory that for $\Re z$ sufficiently large,

$$(r_S^+)\hat{}\,(z)=\hat{r}(z)-(d^+)\hat{}\,(z)=[\hat{L}(z)]^{-1}-\sum_{z_l\in Z}SP_{z_l}(z).\tag{6.7}$$

The expression on the right-hand side is locally analytic with respect to $V(\mathbf{R}^+;\mathbf{C}^{n\times n};\rho^S)$ on $\{z|\Re z\geqq\omega^S)\cup\{\infty\}$ and it vanishes at infinity; hence by Proposition 2.3 of [3], $r_S^+\in L^1(\mathbf{R}^+;\mathbf{C}^{n\times n};\rho^S)$. Multiplying (6.7) by z, we get

$$[(r_S^+)']\hat{}\,(z)=z\left\{[\hat{L}(z)]^{-1}-\sum_{z_l\in Z}SP_{z_l}(z)\right\},\qquad \Re z\geqq\omega^S.$$

Again the expression on the right-hand side is easily seen to be locally analytic with respect to $V(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^S)$ in $\{z|\Re z \geqq \omega^S\} \cup \{\infty\}$; hence, $(r_S^+)' \in V(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^S)$. We conclude that $r_S^+ \in W^{1,1}(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^S)$. By Lemma 3.7 of [10], $r_S^+ \in BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \rho^S)$, or equivently, $r_S \in BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \eta_S)$. □

We decompose $\mathscr{C}\mathscr{S}$ into the two subspaces $\mathscr{S}$ and $\mathscr{C}$ and construct projections of $\mathscr{D}$ onto $\mathscr{S}$ and $\mathscr{C}$ in a manner which parallels our earlier decomposition in § 4. To do this we must now assume that the influence function η is sufficiently small at minus infinity and sufficiently large at plus infinity. More precisely, henceforth we assume that

$$\rho(t) \geqq (1+t)^{\sigma-1}, \qquad t \in \mathbf{R}^+, \tag{6.8}$$

$$BUC(\mathbf{R}^-; \mathbf{C}^n; \eta_S) \subset \mathscr{B}(\mathbf{R}^-; \mathbf{C}^n; \eta), \tag{6.9}$$

$$\mathscr{B}(\mathbf{R}^+; \mathbf{C}^n; \eta) \subset L^1(\mathbf{R}^+; \mathbf{C}^n; \rho_S). \tag{6.10}$$

We discuss assumptions (6.8)-(6.10) after we have stated and proved our results. Note that the closed graph theorem implies that the inclusions in (6.9) and (6.10) are continuous.

DEFINITION 6.1. A point (ϕ, f) in $\mathscr{D}$ belongs to the stable subspace $\mathscr{S}$ if the solution $x(\phi, f)$ of (2.5) satisfies $x(\phi, f)(t) \to 0$ as $t \to \infty$. A point $(\phi, 0)$ in $\mathscr{D}$ belongs to the central subspace $\mathscr{C}$ if $\mathscr{L}\phi(t) = 0$ for $t \in \mathbf{R}^-$, $\phi(t) = O(|t|^{\sigma-1})$ as $t \to -\infty$, and $x(\phi, 0)(t) = o(e^{\lambda t})$ as $t \to \infty$.

THEOREM 6.1. *Assume that* (6.8)-(6.10) *hold. Given* $(\phi, f) \in \mathscr{D}$, *define* $v(\phi, f)$ *and* $w(\phi, f)$ *by* (4.3), *and let* θ *and* χ *be the restrictions of* $v(\phi, f)$ *and* $w(\phi, f)$, *respectively, to* $\mathbf{R}^-$. *Let* P_S *be the operator which maps* (ϕ, f) *into* (θ, f), *and let* P_C *be the operator which maps* (ϕ, f) *into* $(\chi, 0)$. *Then* P_S *and* P_C *are continuous projections in* $\mathscr{D}$ *with ranges* $\mathscr{S}$ *and* $\mathscr{C}$, *respectively, and* $P_S + P_C = P_{CS}$.

As usual, one can reduce the proof of Theorem 6.1 to the proof of the following two lemmas.

LEMMA 6.3. *A point* $(\phi, 0)$ *in* $\mathscr{D}$ *belongs to* $\mathscr{C}$ *if and only if* ϕ *is the restriction to* $\mathbf{R}^-$ *of a function* $x \in \oplus_{z_l \in Z_C} \mathscr{N}_l$, *or equivalently, if and only if* ϕ *is the restriction to* $\mathbf{R}^-$ *of a function* $x \in BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_C)$ *satisfying* $\mathscr{L}x = 0$. *Moreover,* $\mathscr{S} \cap \mathscr{C} = \{(0, 0)\}$.

LEMMA 6.4. *The range of* P_S *is contained in* $\mathscr{S}$, *and the range of* P_C *is contained in* $\mathscr{C}$.

Proof of Lemma 6.3. That the two different characterizations of the functions x in Lemma 3.1 are equivalent follows as usual from Theorem 5.1 of [4]. It is also clear that if ϕ is of the type mentioned in Lemma 6.3, then $(\phi, 0) \in \mathscr{C}$.

Conversely, suppose that $(\phi, 0) \in \mathscr{C}$. Then $(\phi, 0) \in \mathscr{C}\mathscr{S}$; hence, by Corollary 3.1, $x(\phi, 0)(t) = o(e^{\gamma t})$ as $t \to \infty$. Therefore, $x(\phi, 0) \in L^\infty(\mathbf{R}; \mathbf{C}^n; \eta_C)$. As $\mathscr{L}\phi(t) = 0$ for $t \in \mathbf{R}^-$, and $x(\phi, 0)$ satisfies (2.5) with $f = 0$, we have $\mathscr{L}x(\phi, 0) = 0$. As usual, this makes it possible for us first to conclude that $x \in W^{1,\infty}(\mathbf{R}; \mathbf{C}^n; \eta_C)$, so that by Lemma 3.7 of [10], $x \in BUC(\mathbf{R}; \mathbf{C}^n; \eta_C)$, and then to use the fact that $\mathscr{L}x(\phi, 0) = 0$ once more to get $x \in BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_C)$. Thus, ϕ is the restriction to $\mathbf{R}^-$ of the function $x(\phi, 0)$, which belongs to $BUC^1(\mathbf{R}; \mathbf{C}^n; \eta_C)$ and satisfies $\mathscr{L}x(\phi, 0) = 0$.

The fact that $S \cap C = \{(0, 0)\}$ follows immediately from the growth rate at infinity imposed on $x(\phi, f)$ in the case when $(\phi, f) \in \mathscr{S}$ together with the fact that $x(\phi, 0) \in \oplus_{z_l \in Z_C} \mathscr{N}_l$ when $(\phi, 0) \in \mathscr{C}$. □

Proof of Lemma 6.4. Define θ and χ as in Theorem 6.1. We have to show that $(\theta, f) \in \mathscr{S}$, and that $(\chi, 0) \in \mathscr{C}$.

It follows from the way in which d_C was defined that $d_C \in BUC^1(\mathbf{R}; \mathbf{C}^{n\times n}; \eta_C)$. By (6.6) and (6.10), $f + M\phi \in L^1(\mathbf{R}^+; \mathbf{C}^n; \rho_S) \subset L^1(\mathbf{R}^+; \mathbf{C}^n; \rho_C)$. We can now argue exactly in the same way as in the proof of Lemma 3.2 to show that $(\chi, 0) \in \mathscr{C}$.

It remains to show that $(\theta, f) \in \mathcal{S}$. As in the proof of Lemma 3.2 we conclude that $v(\phi, f) = x(\theta, f)$, so it suffices to show that $v(\phi, f)(t) \to 0$ as $t \to \infty$. To prove this, let us first observe that η_S is dominated by ρ_S. To see that this is the case it suffices to note that the influence function which is $(1+|t|)^{1-\sigma}$ for $t \leqq 0$ and 1 for $t > 0$ is dominated by the weight function which is 1 for $t \leqq 0$ and $(1+|t|)^{\sigma-1}$ for $t > 0$, and ρ^S, extended to $\mathbf{R}$ to be 1 on $(-\infty, 0)$, is dominated by itself. It follows trivially from (2.1) and (2.2) that if η_1 is dominated by ρ_1 and η_2 is dominated by ρ_2, then $\eta_1\eta_2$ is dominated by $\rho_1\rho_2$, and therefore our claim that η_S is dominated by ρ_S is correct. By Lemma 6.2, $r_S \in BUC((-\infty, 0); \mathbf{C}^{n\times n}; \eta_S) \cap BC_0(\mathbf{R}^+; \mathbf{C}^{n\times n}; \eta_S)$, and by (6.10), $f + M\phi \in L^1(\mathbf{R}^+; \mathbf{C}^n; \rho_S)$. It follows from (4.3) and Lemma A.1 in Appendix A that $v(\phi, f)(t) \in BC_0(\mathbf{R}^+; \mathbf{C}^n; \eta_S) = BC_0(\mathbf{R}^+; \mathbf{C}^n; \rho^S)$. As $\rho^S(t) \geqq 1$ for $t \in \mathbf{R}^+$ we get $v(\phi, f)(t) \to 0$ as $t \to \infty$. □

Again we observe that we obtained a better convergence rate in the stable subspace than the one used in Definition 6.1:

COROLLARY 6.1. *If* $(\phi, f) \in \mathcal{S}$, *then* $x(\phi, f)(t) = o([\rho^S(t)]^{-1}$ *as* $t \to \infty$.

Above, in the definition of σ, and in the definition of the weight and influence functions in (6.6), we assumed that Z_C is nonempty. If $Z_C = \emptyset$, then one can still proceed in essentially the same way. There is no longer a need to assume (6.8) and (6.9), so these two conditions can be dropped. In this case, since $d_C = 0$ and $r_S = r_{CS}$, we have $r_{CS} \in L^1(\mathbf{R}; \mathbf{C}^{n\times n}; \rho^S)$ and $r'_{CS} \in V(\mathbf{R}; \mathbf{C}^{n\times n}; \rho^S)$. If we still assume (6.10), but with ρ_S replaced by ρ^S, then for every $(\phi, f) \in \mathcal{CS}$, it follows from (3.3) that $y(\phi, f) = x(\phi, f)$, and from Lemma 2.1 of [10] that $x(\phi, f) \in W^{1,1}(\mathbf{R}^+; \mathbf{C}^n; \rho^S)$; hence, by Lemma 3.7 of [10], $x(\phi, f) = o([\rho^S(t)]^{-1})$ as $t \to \infty$. In other words, the conclusion of Corollary 6.1 still holds if we replace $\mathcal{S}$ by $\mathcal{CS}$ and assume (6.10) with ρ_S replaced by ρ^S, but drop (6.8) and (6.9).

One can also obtain a slightly different result in the case when $Z_C = \emptyset$ by appealing to Proposition 4.1 with ρ replaced by ρ^S. For this to be possible we have to replace (6.10) by the assumption that there exists an influence function η^S dominated by ρ^S, and satisfying

$$(6.11) \qquad \eta^S(t) \leqq \eta(t), \qquad t \in \mathbf{R}^+.$$

Clearly, when (6.11) holds, we have $\mathcal{B}(\mathbf{R}^+; \mathbf{C}^n; \eta) \subset \mathcal{B}(\mathbf{R}^+; \mathbf{C}^n; \eta^S)$. If we also knew that

$$(6.12) \qquad \eta^S(t) \leqq \eta(t), \qquad t \in \mathbf{R}^-,$$

then we could apply Proposition 4.1 to conclude that for every $(\phi, f) \in \mathcal{CS}$, $x(\phi, f) \in \mathcal{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta^S)$. However, in general (6.12) will not be satisfied. Still, it turns out that (6.11) alone implies that $x(\phi, f) \in \mathcal{B}^{m+1}(\mathbf{R}^+; \mathbf{C}^n; \eta^S)$ for each $(\phi, f) \in \mathcal{CS}$, a fact one can see as follows. Let ξ be an infinitely many times continuously differentiable "cutoff" function satisfying $\xi(t) = 1$ for $t \leqq -1$ and $\xi(t) = 0$ for $t \geqq 0$. Define $x_1(\phi, f)(t) = \xi(t)x(\phi, f)(t)$ and $x_2(\phi, f)(t) = (1-\xi(t))x(\phi, f)(t)$ for $t \in \mathbf{R}$. Then, if we let f_1 be the restriction of the function $\mathcal{L}x_1$ to $\mathbf{R}^+$ and $f_2 = f - f_1$, we find that $x_1(\phi, f) = x(\phi_1, f_1)$ and $x_2(\phi, f) = x(\phi_2, f_2)$, where $\phi_1(t) = \xi(t)\phi(t)$ and $\phi_2(t) = (1-\xi(t))\phi(t)$ for $t \in \mathbf{R}^-$. As $\mathcal{L}$ maps $\mathcal{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta)$ into $\mathcal{B}^m(\mathbf{R}; \mathbf{C}^n; \eta)$, both f_1 and f_2 belong to $\mathcal{B}^m(\mathbf{R}^+; \mathbf{C}^n; \eta) \subset \mathcal{B}^m(\mathbf{R}^+; \mathbf{C}^n; \eta^S)$. If follows from Definition 3.1 and the fact that $(\phi, f) \in \mathcal{CS}$ that both (ϕ_1, f_1) and (ϕ_2, f_2) belong to $\mathcal{CS}$. We can apply Proposition 4.1 with ρ replaced by ρ^S, η replaced by η^S, and (ϕ, f) replaced by (ϕ_2, f_2) to conclude that $x_2(\phi, f) = x(\phi_2, f_2) \in \mathcal{B}^{m+1}(\mathbf{R}; \mathbf{C}^n; \eta^S)$. But $x(\phi, f)(t) = x_2(\phi, f)(t)$ for $t \in \mathbf{R}^+$, and therefore also $x(\phi, f) \in \mathcal{B}^{m+1}(\mathbf{R}^+; \mathbf{C}^n; \eta^S)$, as claimed.

We conclude this section with a short discussion of the conditions (6.9) and (6.10), and with an example which shows that the decay rate in Corollary 6.1 is optimal.

The conditions (6.9) and (6.10) require that η be small and large at $-\infty$ and ∞, respectively. As one might expect, this essentially implies (6.8), as the following example shows:

Example 6.1. Let $p \geqq 0$, and set $\rho(t)=1$ for $t \leqq 0$, $\rho(t)=(1+t)^p$ for $t>0$. Then ρ is a dominating function with $\omega=\alpha=0$. Let q_- and q_+ be nonnegative. Then the function η defined by

$$\eta(t)=\begin{cases}(1+|t|)^{-q_-}, & t\in \mathbf{R}^-,\\ (1+|t|)^{q_+}, & t\in \mathbf{R}^+,\end{cases}$$

is dominated by ρ if and only if $q_-+q_+\leqq p$.

To show that the condition $p \geqq q_-+q_+$ is sufficient for ρ to dominate η one argues in the same way as in the proof of Lemma 6.4 (take $\rho^S(t)=\eta(t)$ for $t\geqq 0$). That this condition is necessary follows directly from (2.2), which for $s=-t/2$ gives $\rho(t)\geqq \eta(t/2)/\eta(-t/2)$.

That the decay rate in Corollary 6.1 is the best possible one can be shown with the following example.

Example 6.2. Let ρ and η be as in Example 6.1 with $q_->1$, define $\rho^S(t)=\eta(t)$ for $t\geqq 0$, and let $\mathscr{B}=L^1$ and $m=0$. Consider the scalar equation

$$\begin{aligned} x'(t)&=f(t), && t\in\mathbf{R}^+,\\ x(t)&=\phi(t), && t\in\mathbf{R}^-. \end{aligned} \tag{6.13}$$

The characteristic function $\hat{L}(z)=z$ has a simple central eigenvalue at $z=0$. For this example, $\mathscr{U}=\{(0,0)\}$, $\mathscr{C}=\{(k,0)|k\in\mathbf{C}\}$, $\rho_S(t)=\rho^S(t)$ for $t\geqq 0$, and

$$\mathscr{S}=\left\{(\phi,f)\in W^{1,1}(\mathbf{R}^-;\mathbf{C};\eta)\times L^1(\mathbf{R}^+;\mathbf{C};\eta)|\phi(0)+\int_0^\infty f(s)\,ds=0\right\}.$$

The solution x of (6.13) corresponding to $(\phi,f)\in\mathscr{S}$ is given by

$$x(\phi,f)(t)=-\int_t^\infty f(s)\,ds,\qquad t\in\mathbf{R}^+,$$

so $x(t)=o(t^{-q_+})$ as $t\to\infty$. Clearly, this estimate for the rate of decay of $x(t)$ to zero as $t\to\infty$ cannot be improved.

Appendix A. Above we needed the following variant of Lemma 2.3 of [10].

LEMMA A.1. *Let $a\in L^1(\rho)$ and $\phi\in L^\infty(\eta)$, where η is dominated by ρ. Then $a*\phi\in BUC(\eta)$. If, in addition,* $\lim_{t\to\infty}\operatorname{ess\,sup}_{s\geqq t}\eta(t)|\phi(t)|=0$, *then $\eta(t)(a*\phi)(t)\to 0$ as $t\to\infty$, and if* $\lim_{t\to-\infty}\operatorname{ess\,sup}_{s\leqq t}\eta(t)|\phi(t)|=0$, *then $\eta(t)(a*\phi)(t)\to 0$ as $t\to-\infty$.*

As in [10], this follows directly from the fact that the same statement is true when $\rho(t)\equiv\eta(t)\equiv 1$.

REFERENCES

[1] I. GOHBERG AND L. RODMAN, *Analytic matrix functions with prescribed local data*, J. d'Analyse Math., 40 (1981), pp. 90-128.

[2] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[3] G. S. JORDAN, O. J. STAFFANS AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749-782.

[4] G. S. Jordan, O. J. Staffans and R. L. Wheeler, *Convolution operators in a fading memory space: The critical case*, this Journal, 18 (1987), pp. 366-386.

[5] G. S. Jordan and R. L. Wheeler, *Asymptotic behavior of unbounded solutions of linear Volterra integral equations*, J. Math. Anal. Appl., 55 (1976), pp. 596-615.

[6] ———, *Structure of resolvents of Volterra integral and integrodifferential systems*, this Journal, 11 (1980), pp. 119-132.

[7] F. Kappel and H. K. Wimmer, *An elementary divisor theory for autonomous linear functional differential equations*, J. Differential Equations, 21 (1976), pp. 134-147.

[8] R. K. Miller, *Structure of solutions of unstable linear Volterra integrodifferential equations*, J. Differential Equations, 15 (1974), pp. 129-157.

[9] R. K. Miller and R. L. Wheeler, *Asymptotic behavior for a linear Volterra integral equation in Hilbert space*, J. Differential Equations, 23 (1977), pp. 270-284.

[10] O. J. Staffans, *On a neutral functional differential equation in a fading memory space*, J. Differential Equations, 50 (1983), pp. 183-217.

[11] ———, *The null space and the range of a convolution operator in a fading memory space*, Trans. Amer. Math. Soc., 281 (1984), pp. 361-388.

[12] ———, *Semigroups generated by a neutral functional differential equation*, this Journal, 17 (1986), pp. 46-57.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

LOCAL STABILITY RESULTS FOR THE ELASTIC BEAM EQUATION*

ALESSANDRA LUNARDI†

Abstract. We study classical solutions of a partial differential equation which arise as a model for the transverse deflection of an extensible beam with hinged ends in a viscous medium. In particular, we study the stability properties of all the stationary solutions and of small periodic orbits near stationary solutions.

Key words. stability, Hopf bifurcation, beam equation, abstract parabolic semilinear equations

AMS(MOS) subject classifications. 35K60, 58D25, 73H10

Introduction. In this paper we study the asymptotic behavior of the solutions of a partial differential equation

$$\begin{aligned} &u_{tt}(t,x)+\alpha u_{xxxx}(t,x)-\left(\beta+k\int_0^l (u_x(\xi,t))^2\,d\xi\right),\\ (0.1)\qquad &u_{xx}(t,x)+\gamma u_{xxxxt}(t,x)-\sigma\int_0^l u_x(t,\xi)u_{xt}(t,\xi)\,d\xi,\\ &u_{xx}(t,x)+\delta u_t(t,x)=0,\qquad t\geqq 0,\quad 0\leqq x\leqq l \end{aligned}$$

with initial and boundary conditions

$$(0.2)\qquad u(t,0)=u(t,l)=u_{xx}(t,0)=u_{xx}(t,l)=0,\qquad t\geqq 0,$$

$$(0.3)\qquad \begin{aligned} &u(0,x)=u_0(x),\qquad 0\leqq x\leqq l,\\ &u_t(0,x)=v_0(x),\qquad 0\leqq x\leqq l \end{aligned}$$

for initial data near stationary solutions. In particular, we study stability and instability of the stationary solutions and of small periodic orbits near stationary solutions.

Equation (0.1) arises as a model for the transverse deflection $u(t,x)$ of the centerline of an elastic beam in a viscous medium; the boundary conditions in (0.2) correspond to the case of hinged ends (see [9], [14]). The coefficients are given by

$$(0.4)\qquad \alpha=\frac{EI}{\rho},\quad \beta=\frac{EA\Delta}{\rho},\quad \gamma=\frac{\eta I}{\rho},\quad k=\frac{EA}{2l\rho},\quad \sigma=\frac{A\eta}{l\rho},$$

where E is the Young's modulus, I is the cross-sectional second moment of area, ρ is the mass per unit length, l is the length of the beam at stress-free state, Δ is the stretching, η is the viscosity and δ is the coefficient of external damping. Therefore $\alpha, k, \gamma, \sigma$ are positive whereas the sign of β and δ is unrestricted. The assumption $\gamma>0$ gives a parabolic character to (0.1). For the hyperbolic case $\gamma=0$ we refer to [8] and the references quoted there.

The initial value problem for (0.1) has been studied first in [2] and then in [7], [4], [5], [12], where results of existence in the large, uniqueness and continuous dependence of the solution on u_0 and v_0 in suitable norms are given. Moreover in [2], [4] and [5], sufficient conditions on the parameters are given which guarantee the convergence of the solution to zero or to some other stationary solution for any initial data. In [2] the Galerkin approximation method is used, and in [4], [5] the theory of

* Received by the editors March 5, 1986; accepted for publication (in revised form) August 14, 1986.

† Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56100 Pisa, Italy.

analytic semigroups in Hilbert space is used. In [7], [12] some qualitative methods (in particular, center manifold theory) are suggested for the study of the behavior of the solutions for small u_0 and v_0. All the mentioned authors work in a Hilbert space setting and obtain weak solutions which satisfy (0.1) in the variational sense ([2]), or strong solutions such that $u(t,\cdot)$, $u_t(t,\cdot)$ belong to $H^4(0,l)$, $u_{tt}(t,\cdot)$ belongs to $L^2(0,l)$ ([2], [4], [5]) or such that $u(t,\cdot)+\alpha u_t(t,\cdot)$ belongs to $H^4(0,l)$ and $u_t(t,\cdot)\in H^2(0,l)$, $u_{tt}\in L^2(0,l)$ ([7]).

We are interested here in classical solutions of (0.1) (that is, such that u_{tt}, u_{xxxx}, u_{xxxxt} exist and are continuous with respect to (t,x) up to $t=0$); therefore we consider only initial data belonging to $C^4([0,l])$ and satisfying the necessary compatibility conditions

$$u_0(0)=u_0(l)=u_0''(0)=u_0''(l)=v_0(0)=v_0(l)=v_0''(0)=v''(l)=0,$$
$$\alpha u_0^{iv}(0)+\gamma v_0^{iv}(0)=\alpha u_0^{iv}(l)+\gamma v_0^{iv}(l)=0.$$

Setting

$$E=\{\phi\in C([0,l];\mathbb{R});\ \phi(0)=\phi(l)=0\},$$
$$D=\{\phi\in C^4([0,l];\mathbb{R});\ \phi(0)=\phi(l)=\phi''(0)=\phi''(l)=0\},$$

we transform problem (0.1)–(0.3) into an abstract parabolic initial value problem in the Banach space $X=D\times E$:

$$w(t)=A(t)+g(w(t)),\quad t\geqq 0,\qquad w(0)=w_0, \tag{0.5}$$

where $w(t)=(u(t,\cdot),u_t(t,\cdot))$, $w_0=(u_0,v_0)$. The linear unbounded operator A: $D(A)=\{(\phi,\psi)\in D\times D;\ \alpha\phi^{iv}(0)+\gamma\phi^{iv}(0)=\alpha\phi^{iv}(l)+\gamma\psi^{iv}(l)=0\}\to X$, $A(\phi,\psi)=(\psi,-\alpha\phi^{iv}-\gamma\psi^{iv}+\beta\phi''-\delta\psi)$ generates an analytic semigroup in X, and the nonlinear function g maps $D(A)$ into an interpolation space $D_A(\theta,\infty)$. Now (0.5) is nothing but a semilinear abstract parabolic i.v.p., which is relatively easy to handle. In particular, it is not difficult to show local existence, uniqueness and continuous dependence (in the norm of $D(A)$) on w_0 of the solution and to give results of linearized stability and instability for any stationary solution $\bar{w}$ in the noncritical cases. Even in the cases of instability, we are able to give conditions on w_0 for existence in the large and convergence of the solution to $\bar{w}$ as $t\to+\infty$.

We study also the critical cases of stability, when the operator $L:D(A)\to X$, defined by $Lw=(A+f'(\bar{w}))w$, has some eigenvalues on the imaginary axis, and the remainder of the spectrum has negative real part. In our case, the purely imaginary eigenvalues of L are simple, and their corresponding eigenvectors span a subspace $X^+\subset D(A)$, whose dimension may be 1, 2, or 4. The dimension of X^+ is 1 only when $\bar{w}\equiv 0$ and $\delta\geqq-(\pi^4/l^4)\gamma$, $\beta=-(\pi^2/l^2)\alpha$. In this case, we have $f(D(L))\subset X^-$, where $X^-=(1-P)(X)$ and P is a projection on X^+. Thanks to this property it is not difficult to prove the stability of the null solution of (0.5).

In the other cases, the dimension of X^+ is 2 or 4, and we have $f(X^+)\subset X^+$: that is, X^+ is a center manifold, and it is shown to have the usual attractivity properties of the center manifolds. In particular, to study stability properties of the stationary solution $\bar{w}$ considered, and existence and attractivity of small periodic orbits near $\bar{w}$, we can reduce our problem to a system of ordinary differential equations in $\mathbb{R}^2$ or in $\mathbb{R}^4$.

The well-known techniques of Hopf bifurcation are employed in the two-dimensional system and we give necessary and sufficient conditions on the parameters to get asymptotic stability or instability. The study of the four-dimensional system is much more complicated because the results available in the literature are not as complete as in the two-dimensional case. Moreover, there are no relations between the stability properties of the stationary solutions and the stability properties of the periodic solutions given by Hopf's theorem. However, in our case, a nonresonance condition is satisfied and we find that $\bar{w}$ is asymptotically stable, whereas the periodic orbits are not stable.

The paper is organized as follows: in § 1, problem (0.5) is treated; in § 2, the results of § 1 are applied to problem (0.1)-(0.3) and the noncritical cases of stability are studied. In § 3 the critical cases of stability are considered. At the end of § 3 we give a summary scheme of all the results. Finally, the Appendix contains the proofs of the propositions of § 1.

1. Notation and preliminaries on abstract parabolic equations. Let X be a real Banach space with norm $\|\cdot\|$ and let $\tilde{X}=\{x+iy;\, x, y \in X\}$ be its complexification. If $L\colon D(L)\subset X\to X$ is a linear closed operator, we denote by $\tilde{L}\colon D(\tilde{L})\to\tilde{X}$ its complexification, defined by $\tilde{L}(x+iy)=Lx+iLy$, $x, y\in D(L)$. Throughout the paper we shall assume that $D(L)$ is dense in X and L generates an analytic semigroup e^{tL} in X, that is,

(1.1) there exist $\omega\in\mathbb{R}$, $M>0$, $\theta\in\,]\pi/2, \pi[$ such that the resolvent set $\rho(\tilde{L})$ of $\tilde{L}$ contains a sector $S=\{\lambda\in\mathbb{C};\, \lambda\neq\omega,\, |\arg(\lambda-\omega)|<\theta\}$ and $\|(\lambda-\tilde{L})^{-1}\|_{L(\tilde{X})}\leqq M|\lambda-\omega|^{-1}$ for $\lambda\in S$.

In this case, the interpolation spaces $D_L(\theta,\infty)$ $(0<\theta<1)$ are defined by

$$D_L(\theta,\infty)=\left\{x\in X;\, [x]_\theta=\sup_{0<t\leqq 1}\|t^{1-\theta}L\,e^{tL}x\|<+\infty\right\},\tag{1.2}$$
$$\|x\|_{D_L(\theta,\infty)}=\|x\|+[x]_\theta.$$

Then we have $D(L)\hookrightarrow D_L(\theta,\infty)\hookrightarrow D_L(\alpha,\infty)\hookrightarrow X$ for $0<\alpha<\theta<1$, where $D(L)$ is endowed with the graph norm. For other properties see [21, §§ 1.13, 1.14].

In the next section, problem (0.1), (0.2) will be reduced to an abstract semilinear initial value problem:

$$\dot{w}(t)=Lw(t)+g(w(t)),\quad t\geqq 0,\qquad w(0)=w_0,\tag{1.3}$$

where

$$L \text{ satisfies (1.1), } g \text{ belongs to } C^\infty(D(L);\, D_L(\theta,\infty))\tag{1.4}$$

for some $\theta\in\,]0, 1[$. For such a problem it is not difficult to give results of local existence and uniqueness of the solution, continuous dependence on w_0 and regularity (see the Appendix for a proof).

PROPOSITION 1.1. *Let* (1.4) *hold and let* $w_0\in D(L)$. *Then there exists a maximal time interval* $[0, \tau[(\tau=\tau(w_0)>0)$ *and a unique* $w\in C^1([0, \tau[;\, X)\cap C([0, \tau[;\, D(L))$ *which satisfies* (1.3).

Let $w_{0n}\in D(L)$ $(n\in\mathbb{N})$ *be such that* $\|w_{0n}-w_0\|_{D(L)}\to 0$ *as* $n\to+\infty$, *and denote by* $w_n\colon [0, \tau(w_n)[\to D(L)$ *the solution of* (1.3) *with initial value* w_{0n}. *Then for each* $\tau<\tau(w_0)$ *and* $\varepsilon>0$ *there exists* $\bar{n}$ *such that, for* $n\geqq\bar{n}$, $\tau(w_{0n})\geqq\tau$ *and* $\sup_{0\leqq t\leqq\tau}\|w_n(t)-w(t)\|_{D(L)}\leqq\varepsilon$. *In other words, the mapping* $w_0\to w(t)$ *is a local semiflow on* $D(L)$.

From now on, we shall assume

$$g(0)=0,\qquad g'(0)=0\tag{1.5}$$

and we shall study the stability properties of the null solution of (1.3) and of other invariant sets.

We recall that, if $S: \{(t, w_0);\ w_0 \in Y,\ t \in [0, \tau(w_0)[\} \to Y$ is a semiflow on a Banach space Y, a subset $\Omega \subset Y$ is said to be *invariant* if for each $w_0 \in \Omega$ we have $\tau(w_0) = +\infty$ and $S(t, w_0) \in \Omega$ for all $t \geqq 0$. Ω is said to be *stable* if for any $r > 0$ there exists $\varepsilon > 0$ such that for any $w_0 \in Y$ with $\operatorname{dist}(w_0, \Omega) \leqq \varepsilon$ we have $\tau(w_0) = +\infty$ and $\operatorname{dist}(S(t, w_0), \Omega) \leqq r$ for all $t \geqq 0$. Ω is said to be *asymptotically stable* if it is stable and moreover, for any $\varepsilon > 0$, there are $\delta > 0$, $T > 0$ such that if $w_0 \in Y$ and $\operatorname{dist}(w_0, \Omega) \leqq \delta$, then $\operatorname{dist}(S(t, w_0), \Omega) \leqq \varepsilon$ for all $t \geqq T$. Ω is said to be *unstable* if it is not stable.

Let now $Y = D(L)$, $S(t, w_0) = w(t)$, where w is the solution of (1.3). The null solution of (1.3) is said to be stable (resp. asymptotically stable, unstable) if the set $\Omega = \{0\}$ is stable (resp. asymptotically stable, unstable).

We begin with an exponential asymptotic stability result.

PROPOSITION 1.2. *Assume that*

$$\sup\{\operatorname{Re}\lambda;\ \lambda \in \sigma(L)\} \doteq -\bar{\omega} < 0. \tag{1.6}$$

Then the null solution of (1.3) *is asymptotically stable. In particular, for each* $\eta \in]0, \bar{\omega}[$ *there exists* r, $C > 0$ *such that for any* $w_0 \in D(L)$ *with* $\|w_0\|_{D(L)} < r$ *we have*

$$\begin{aligned} &\tau(w_0) = +\infty, \\ &\|w(t)\|_{D(L)} \leqq C e^{-\eta t} \|w_0\|_{D(L)} \quad \textit{for all } t \geqq 0. \end{aligned} \tag{1.7}$$

Assume now that the spectrum of $\tilde{L}$ has an element with positive real part. Then, if $\sigma^+(L) = \{\lambda \in \sigma(L);\ \operatorname{Re}\lambda > 0\}$ is closed, one can show (arguing as in [6, Thm. 5.1.3]) that the null solution is unstable. But, even in this case, it is often possible to prove the existence of an invariant stable manifold M, such that $0 \in M \subset D(L)$, M is homeomorphic to a ball of some subspace of $D(L)$, and for each $w_0 \in M$ the solution of (1.3) is defined in $[0, +\infty[$ and converges to 0 exponentially as $t \to +\infty$. Analogously, we can show the existence of an invariant unstable manifold W, such that $0 \in W \subset D(L)$, W is homeomorphic to a ball of another subspace of $D(L)$, and for each $w_0 \in W$ there exists a backward solution of (1.3), $w:]-\infty, 0] \to W$, with $\lim_{t \to -\infty} w(t) = 0$. To study critical cases of stability (namely, when $\sup\{\operatorname{Re}\lambda;\ \lambda \in \sigma(\tilde{L})\} = 0$), another invariant manifold—the so-called center manifold—may be introduced. The study of such manifolds and of their properties is, in general, rather lengthy and complicated. For the sake of brevity we do not consider here the most general case, but we assume on L and g some particular hypotheses which are satisfied in a sufficiently large class of equations, including the beam equation (0.1), (0.2). In particular, we set, for $\omega \in \mathbb{R}$:

$$\sigma^+_\omega(L) = \{\lambda \in \sigma(\tilde{L});\ \operatorname{Re}\lambda > \omega\}, \qquad \sigma^-_\omega(L) = \sigma(\tilde{L}) \backslash \sigma^+_\omega(L), \tag{1.8}$$

and we assume that there exists $\omega \in \mathbb{R}$ such that

(1.9) $\sigma^+_\omega(L)$ consists of a finite (nonzero) number of eigenvalues with finite algebraic multiplicity.

Let $\tilde{X}^+$ be the eigenspace of $\tilde{L}$ corresponding to $\sigma^+_\omega(L)$. If $\tilde{X}^+$ is spanned by $x_1 + iy_1, \cdots, x_n + iy_n$ $(n \in \mathbb{N})$, set

$$X^+ = \operatorname{span}_{\mathbb{R}}\{x_1, y_1, \cdots, x_n, y_n\}. \tag{1.10}$$

Let $P^+: X \to X^+$ be a projection that commutes with L, that is

$$\begin{aligned} &P^+ \in L(X), \quad P^+(X) \subset X^+, \quad P^+x = x \quad \forall x \in X^+, \\ &P^+Lx = LP^+x \quad \forall x \in D(L) \end{aligned} \tag{1.11}$$

(such a P^+ exists thanks to (1.9)), and set

$$P^-: X \to X, \quad P^-x = x - P^+x, \quad X^- = P^-(X). \tag{1.12}$$

In the following we shall often assume

$$(1.13) \qquad g(X^+) \subset X^+.$$

Condition (1.13) implies that X^+ is invariant; since, by (1.9), X^+ is finite-dimensional, all the stability results about ordinary differential equations hold for the semiflow in X^+. In particular, we easily get instability results as follows:

(1.14) If $\omega = 0$ satisfies (1.9), and (1.13) holds, then the null solution is unstable. In particular, there is $r > 0$ such that for each $w_0 \in X^+$ with $\|w_0\|_{D(L)} \leqq r$ there exists a backward solution w: $]-\infty, 0[\to X^+$ of (1.3), which converges exponentially to zero as $t \to -\infty$.

Moreover, as a corollary of Proposition 1.2, we get a stability result as follows:

(1.15) If some $\omega < 0$ satisfies (1.9) and $g(D(L) \cap X^-) \subset X^-$, then for each $\eta \in]0, -\sup \sigma_\omega^-(L)[$ there exist $\varepsilon = \varepsilon(\eta) > 0$, $C = C(\eta) > 0$ such that if $w_0 \in D(L) \cap X^-$ and $\|w_0\|_{D(L)} \leqq \varepsilon$, then (1.7) holds.

While (1.14) and (1.15) are quite obvious, the attractivity properties of X^+ stated in the next proposition are not trivial and will be proved in the Appendix.

PROPOSITION 1.3. *Let* (1.4), (1.5) *hold, and assume that* (1.9), (1.13) *are satisfied for some* $\omega < 0$. *Then*

(a) *for each* $\eta \in]0, -\sup \sigma_\omega^-(L)[$ *there exist* $\varepsilon_0 = \varepsilon_0(\eta) > 0$, $C_0 = C_0(\eta) > 0$ *such that if* $w_0 \in D(L)$ *and* $\|w_0\|_{D(L)} \leqq \varepsilon_0$, *then*

$$\|P^- w(t)\|_{D(L)} \leqq C_0 \, e^{-\eta t} \|P^- w_0\|_{D(L)} \quad \textit{for all } t \geqq 0;$$

(b) *there exists* $r > 0$ *such that if* $\Omega \subset X^+$ *is compact and asymptotically stable in* X^+, *and* $\operatorname{diam} \Omega = \max \{\|x\|_{D(L)};\ x \in \Omega\} \leqq r$, *then* Ω *is asymptotically stable in* $D(L)$.

Statement (a) is nothing but a local exponential attractivity property of X^+. Statement (b) gives the possibility of reducing local stability problems in Banach space to finite-dimensional ones; it will be widely used in § 3.

We finally consider a critical case of stability, which happens for some value of the parameters in the beam equation.

PROPOSITION 1.4. *Let* (1.4), (1.5) *hold. Assume that L has a finite number of semisimple eigenvalues on the imaginary axis, and* $\sup \{\operatorname{Re} \lambda;\ \lambda \in \sigma(L), \operatorname{Re} \lambda \neq 0\} < 0$, *so that* (1.9) *is satisfied for suitable* $\omega < 0$. *Assume also* $g(D(L)) \subset X^-$. *Then the zero solution of* (1.13) *is stable but not asymptotically stable.*

2. Reduction to an abstract problem, stability and instability in the noncritical cases. Setting $u_t = v$, problem (0.1)-(0.3) becomes

$$(2.1) \quad \begin{aligned} &\text{(i)} && u_t = v, \\ & && v_t = -\alpha u_{xxxx} + (\beta + k|u_x|^2 + \sigma\langle u, v\rangle) u_{xx} - \gamma v_{xxxx} - \delta v, \\ &\text{(ii)} && u(t, 0) = u(t, l) = u_{xx}(t, 0) = u_{xx}(t, l) = 0, \\ &\text{(iii)} && u(0, x) = u_0(x), \qquad v(0, x) = v_0(x), \end{aligned}$$

with the obvious notation

$$(2.2) \qquad \langle \phi, \psi \rangle = \int_0^l \phi(x)\psi(x)\, dx, \qquad |\phi| = \left(\int_0^l (\phi(x))^2 \, dx \right)^{1/2}.$$

Let E be the Banach space of all continuous functions $\phi:[0,l]\to\mathbb{R}$ such that $\phi(0)=\phi(l)=0$, endowed with the sup norm, and let

$$D=\{\phi\in C^4([0,1]);\ \phi(0)=\phi(l)=\phi''(0)=\phi''(l)=0\} \tag{2.3}$$

be endowed with the C^4-norm. Set

$$X=D\times E,$$

$$D(A)=\{(\phi,\psi)\in D\times D;\ \alpha\phi^{(iv)}(0)+\gamma\psi^{(iv)}(0)=\alpha\phi^{(iv)}(l)+\gamma\psi^{(iv)}(l)=0\} \tag{2.4}$$

(endowed with the product norms),

$$A(\phi,\psi)=(\psi,-\alpha\phi^{(iv)}-\gamma\psi^{(iv)}),\qquad (\phi,\psi)\in D(A). \tag{2.5}$$

Then the following proposition holds.

PROPOSITION 2.1. *$D(A)$ is dense in X and the operator $A: D(A)\subset X\to X$ satisfies* (1.1). *For $\theta\in\,]0,1[$ we have*[1]

$$D_A(\theta,\infty)=\begin{cases} D\times\{\phi\in C^{4\theta}([0,l]),\ \phi(0)=\phi(l)=0\} \\ \qquad\qquad\qquad\qquad \textit{if } 0<\theta<\frac{1}{2},\quad \theta\neq\frac{1}{4}, \\ D\times\{\phi\in C^{4\theta}([0,l]);\ \phi(0)=\phi(l)=\phi''(0)=\phi''(l)=0\} \\ \qquad\qquad\qquad\qquad \textit{if } \frac{1}{2}<\theta<1,\quad \theta\neq\frac{3}{4}, \end{cases} \tag{2.6}$$

with equivalence of the respective norms.

Proof. The density of $D(A)$ in X is obvious. Let $\Lambda: D\to C([0,l])$ be defined by $\Lambda u=-u^{iv}$, $u\in D$. It is easy to see that the spectrum of Λ consists of the simple eigenvalues $-\pi^4n^4/l^4$, $n\in\mathbb{N}$, and that Λ satisfies (1.1). Therefore Λ generates an analytic semigroup in $C([0,l])$ (for analytic semigroups with nondense domain, see [19]).

For $(\phi,\psi)\in\tilde{X}$ and $\lambda\in\mathbb{C}$, the equation $\lambda(u,v)-\tilde{A}(u,v)=(\phi,\psi)$ is equivalent to

$$\begin{gathered}(u,v)\in D(\tilde{A}),\qquad v=\lambda u-\phi,\\ \lambda^2u-(\alpha+\gamma\lambda)\Lambda u=\psi+\lambda\phi-\gamma\Lambda\phi.\end{gathered} \tag{2.7}$$

Hence the spectrum of $\tilde{A}$ consists of the point $\lambda_0=-\alpha/\gamma$ and of the eigenvalues

$$\lambda_n^{\pm}=\frac{\pi^2n^2}{2l^2}\left(-\frac{\pi^2n^2}{l^2}\gamma\pm\sqrt{\frac{\pi^4n^4}{l^4}\gamma^2-4\alpha}\right).$$

Moreover, for $\lambda\in\rho(\tilde{A})$ we have: $u=((\lambda^2/(\alpha+\gamma\lambda)-\Lambda)^{-1}(\psi+\lambda\phi-\gamma\Lambda\phi)/(\alpha+\gamma\lambda))$, so that (1.1) easily follows. Since the graph norm of A is equivalent to the norm of D on $D(A)$, then $D_A(\theta,\infty)=D\times D_\Lambda(\theta,\infty)$, with equivalence of the respective norms. To characterize $D_\Lambda(\theta,\infty)$ it is sufficient to observe that $\Lambda=-\Delta^2$, where Δ: $D(\Delta)\to E$, $D(\Delta)=\{\phi\in C^2([0,l]);\ \phi(0)=\phi(1)=0\}$, $\Delta\phi=\phi''$. Then we have $D_\Lambda(\theta,\infty)=D_\Delta(2\theta,\infty)$ for $\theta\neq\frac{1}{2}$. Theorem 2.10 of [11] implies now

$$D_\Delta(\eta,\infty)=\{\phi\in C^{2\eta}([0,l]),\ \phi(0)=\phi(l)=0\},\qquad 0<\eta<1,\quad \eta\neq\tfrac{1}{2}, \tag{2.8}$$

$$\begin{gathered}D_\Delta(\eta,\infty)=\{\phi\in C^{2\eta}([0,l]);\ \phi(0)=\phi(l)=\phi''(0)=\phi''(l)=0\},\\ 1<\eta<2,\quad \eta\neq\tfrac{3}{2},\end{gathered} \tag{2.9}$$

with equivalence of the respective norms, so that (2.6) follows. The results of general interpolation theory which we have just used (such as $D_{\Delta^2}(\theta,\infty)=D_\Delta(2\theta,\infty)$, etc.) are well known in the dense domain case (see, for instance, [21, §§ 1.13, 1.14]) and may be easily extended to our case (see [11]). □

[1] The cases $\theta=\frac{1}{4}$, $\theta=\frac{1}{2}$, $\theta=\frac{3}{4}$ would require the introduction of the Zygmund classes $\mathscr{C}^{1*}$, $\mathscr{C}^{2*}$, $\mathscr{C}^{3*}$ and will not be considered here.

Now set

$$f(\phi, \psi) = (0, k|\phi'|^2 + \sigma\langle\phi, \psi\rangle)\phi'', \qquad \phi \in C^2([0, l]), \quad \psi \in C^1([0, l]). \tag{2.10}$$

Then it is easy to see that f belongs to $C^\infty(D(A), D_A(\theta, \infty))$ for each $\theta < \frac{1}{2}$ and

$$\begin{aligned} f'(\bar\phi, \bar\psi)(\phi, \psi) = (0, 2k\langle\phi', \bar\psi'\rangle + \sigma\langle\bar\phi', \psi'\rangle + \sigma\langle\phi', \bar\psi'\rangle)\bar\phi'' \\ + (0, k|\bar\phi'|^2 + \sigma\langle\bar\phi', \bar\psi'\rangle)\phi''. \end{aligned} \tag{2.11}$$

Finally, define a linear operator $L: D(L) \to X$ by

$$D(L) = D(A), \quad L(\phi, \psi) = (\psi, -\alpha\phi^{iv} - \gamma\psi^{iv} + \beta\phi'' - \delta\psi), \quad (\phi, \psi) \in D(A). \tag{2.12}$$

Then $L = A + B$, where $B \in L(X)$, $B(\phi, \psi) = (0, \beta\phi'' - \delta\psi)$. Therefore L satisfies (1.1) and $D_L(\theta, \infty) = D_A(\theta, \infty)$ for each $\theta \in]0, 1[$, with equivalence of the norms.

Now, if we set $w = (u, v)$, problem (2.1) may be written in the abstract form (1.3), with

$$w_0 = (u_0, v_0), \qquad g(w) = f(u, v). \tag{2.13}$$

In fact, if $w(t) = (u(t), v(t))$ is a solution of (1.3) in some interval $[0, T]$ with X, $D(L)$, L defined by (2.4) and (2.12), then $(u(t, x) \doteq u(t)(x), v(t, x) \doteq v(t)(x))$ is a classical solution of (2.1), such that each term appearing in (2.1)(i) is continuous in $[0, T] \times [0, l]$. Conversely, if $(u(t, x), v(t, x))$ is a solution of (2.1) such that u, v, u_t, v_t, u_{xxxx}, v_{xxxx} are continuous in $[0, T] \times [0, l]$, then the compatibility conditions $v(t, 0) = v(t, l) = v_{xx}(t, 0) = v_{xx}(t, l) = \alpha u_{xxxx}(t, 0) + \gamma v_{xxxx}(t, 0) = \alpha u_{xxxx}(t, l) + \gamma v_{xxxx}(t, l) = 0$ $(0 \leqq t \leqq T)$ hold. Therefore $(u(t) \doteq u(t, \cdot), v(t) \doteq v(t, \cdot))$ belongs to $D(A)$ for $0 \leqq t \leqq T$ and $w(t) = (u(t), v(t))$ is a solution of (1.3) in $[0, T]$.

Then, applying Proposition 1.1, it is possible to obtain a local existence and uniqueness result for the solution of (2.1), together with continuous dependence on (u_0, v_0). More precisely, for each $u_0, v_0 \in C^4([0, l])$ such that the compatibility conditions

$$\begin{aligned} u_0(0) = u_0(l) = u_0''(0) = u_0''(l) = v_0(0) = v_0(l) \\ = v_0''(0) = v_0''(l) = 0, \\ \alpha u_0^{(iv)}(0) + \gamma v_0^{(iv)}(0) = \alpha u_0^{(iv)}(l) + \gamma v_0^{(iv)}(l) = 0 \end{aligned} \tag{2.14}$$

hold, there exist $\tau > 0$ and a unique solution $u: [0, \tau[\times [0, l] \to \mathbb{R}$ of (0.1) such that each derivative of u appearing in (0.1) is continuous in $[0, \tau[\times [0, l]$. Moreover, the C^4 norms of $u(t, \cdot)$ and $u_t(t, \cdot)$ depend continuously on the C^4 norms of u_0 and v_0 (in the sense of Proposition 1.1).

In fact, one could show existence, uniqueness and C^∞ regularity (for $t > 0$) for less regular initial data (u_0, v_0) and even existence in the large of the solution (that is, $\tau(u_0, v_0) = +\infty$). These are not purposes of the present paper and we shall limit ourselves to consider very regular initial data (u_0, v_0) near $(\bar\phi, 0)$, where $\bar\phi$ is any stationary solution of (0.1), (0.2).

We begin with the stationary solution $\bar\phi \equiv 0$. To apply the results of § 1 we have to compute the spectrum of $\tilde{L}$: it is easy to see that $\sigma(\tilde{L}) = \{-\alpha/\gamma\} \cup \{\lambda_{\pm h}, h \in \mathbb{N}\}$, where the simple eigenvalues $\lambda_{\pm h}$ are given by

$$\lambda_{\pm h} = -\frac{1}{2}\left(\delta + \frac{\pi^4}{l^4}\gamma h^4\right) \pm \frac{1}{2}\left[\left(\delta + \frac{\pi^4}{l^4}\gamma h^4\right)^2 - 4\frac{\pi^2 h^2}{l^2}\left(\beta + \frac{\pi^2}{l^2}\alpha h^2\right)\right]^{1/2}, \qquad h \in \mathbb{N}, \tag{2.15}$$

and the eigenspace corresponding to the eigenvalue $\lambda_{\pm h}$ is spanned by

$$\left(\sin\frac{\pi h}{l}x, \lambda_{\pm h}\sin\frac{\pi h}{l}x\right). \tag{2.16}$$

Using (1.14), (1.15) and Proposition 1.2 gives some results about the stability of the null solution.

THEOREM 2.2. *Let $\lambda_j, j\in\mathbb{Z}$, be defined by* (2.15). *Then*

(a) *If $\delta>-(\pi^4/l^4)\gamma$ and $\beta+(\pi^2/l^2)\alpha>0$, the null solution of* (0.1), (0.2) *is exponentially asymptotically stable in the C^4 norm; more precisely, for each $\eta\in\,]0,\min\{\alpha/\gamma,-\operatorname{Re}\lambda_1\}[$ there exist $r>0$, $C>0$ such that if $u_0, v_0\in C^4([0,l])$ satisfy the compatibility conditions* (2.14) *and $\|u_0\|_{C^4([0,l])}\leqq r$, $\|v_0\|_{C^4([0,l])}\leqq r$, then the solution u of* (0.1)-(0.3) *is defined for all positive t, and*

$$\begin{aligned}\|u(t,\cdot)\|_{C^4([0,l])}&+\|u_t(t,\cdot)\|_{C^4([0,l])}\\ &\leqq C e^{-\eta t}(\|u_0\|_{C^4([0,l])}+\|v_0\|_{C^4([0,l])})\quad \textit{for all } t\geqq 0.\end{aligned}\tag{2.17}$$

(b) *If $\delta<-(\pi^4/l^4)\gamma$ or $\beta+(\pi^2/l^2)\alpha<0$, the null solution is unstable. In particular, if $J\subset\mathbb{Z}$ is such that $\operatorname{Re}\lambda_j>0$ for each $j\in J$, then there exists $r>0$ such that for $a_j, b_j\in\mathbb{R}$, $|a_j|\leqq r$, $|b_j|\leqq r$ $(j\in J)$ and*

$$u_0(x)=\sum_{j\in J}a_j\sin\frac{\pi j}{l}x,\qquad v_0(x)=\sum_{j\in J}b_j\sin\frac{\pi j}{l}x,\quad 0\leqq x\leqq l,\tag{2.18}$$

there exists a backward solution $u:]-\infty,0]\times[0,l]\to\mathbb{R}$ of (0.1), (0.2), *$u(t,x)=\sum_{j\in J}a_j(t)\sin(\pi j/l)x$, and a_j, $\dot a_j$ decay exponentially to* 0 *as $t\to-\infty$.*

On the other hand,

(c) *let $\delta<-(\pi^4/l^4)\gamma$ or $\beta+(\pi^2/l^2)\alpha<0$, and let $J'=\{j\in\mathbb{Z};\operatorname{Re}\lambda_j<0\}$. Then for each $\eta\in\,]0,\min\{\alpha/\gamma,-\sup\{\operatorname{Re}\lambda_j;j\in J'\}\}[$ there exist $r>0$, $c>0$ such that if $u_0\in C^4([0,l])$, $v_0\in C^4([0,l])$ satisfy* (2.14),

$$\|u_0\|_{C^4([0,l])}\leqq r,\qquad \|v_0\|_{C^4([0,l])}\leqq r,$$

and

$$\left\langle u_0,\sin\frac{\pi j}{l}x\right\rangle=\left\langle v_0,\sin\frac{\pi j}{l}x\right\rangle=0\quad\forall j\in\mathbb{Z}\backslash J',\tag{2.19}$$

then the solution of (0.1) *is defined for all positive t and satisfies* (2.17).

The critical cases $\delta=-(\pi^4/l^4)\gamma$ and $\beta+(\pi^2/l^2)\alpha=0$ will be studied in the next section.

Let us consider now the other stationary solutions: it is easy to see that if $\beta+(\pi^2/l^2)\alpha\geqq0$ the unique stationary solution of (0.1), (0.2) is $\bar\phi\equiv0$, whereas if $\beta+(\pi^2n^2/l^2)\alpha<0$ for some $n\in\mathbb{N}$, then (0.1), (0.2) has exactly $2n$ nontrivial stationary solutions $\pm\bar\phi_j, j=1,\cdots,n$, given by

$$\begin{aligned}\bar\phi_j(x)=\frac{1}{\pi j}\left[-\frac{2}{kl}\left(\beta+\frac{\pi^2j^2}{l^2}\alpha\right)\right]^{1/2}\sin\frac{\pi j}{l}x,&\\ 0\leqq x\leqq l,\quad j=1,\cdots,n.&\end{aligned}\tag{2.20}$$

It is easy to see that a function u satisfies (0.1), (0.2) if and only if $-u$ does. Therefore for any $j=1,\cdots,n$, $\bar\phi_j$ and $-\bar\phi_j$ have the same stability properties, and, from now on, we shall consider only the $\bar\phi_j$'s.

To use the results of § 1 for the study of the stability of $\bar\phi_n$ we must define the linear operators $L_n:D(L_n)\to X$ and the functions $g_n:D(L_n)\to D_{L_n}(\theta,\infty)$:

$$D(L_n)=D(A),\qquad L_n(\phi,\psi)=L(\phi,\psi)+f'(\bar\phi_n,0)(\phi,\psi),\qquad(\phi,\psi)\in D(A),\tag{2.21}$$

$$g_n(\phi,\psi)=f(\bar\phi_n+\phi,\psi)-f'(\bar\phi_n,0)(\phi,\psi)-f(\bar\phi_n,0),\qquad(\phi,\psi)\in D(A),\tag{2.22}$$

where L,f,f' are given by (2.12), (2.20), (2.21), respectively.

Then L_n satisfies (1.1), $g_n \in C^\infty(D(L_n), D_{L_n}(\theta, \infty))$ for $\theta < \frac{1}{2}$, $g_n(0) = g_n'(0) = 0$, so that the results of the previous section may be applied to problem

$$\dot{w}_n(t) = L_n w_n(t) + g_n(w(t)), \quad t \geqq 0, \qquad w_n(0) = w_0,$$

with $w(t) = (u(t, \cdot) - \bar{\phi}_n, u_t(t, \cdot))$, $w_0 = (u_0 - \bar{\phi}_n, v_0)$.

The spectrum of $\tilde{L}_n$ consists of the point $\lambda_0 = -\alpha/\gamma$ and of the simple eigenvalues $\lambda_{\pm h}$, $h \in \mathbb{N}$, defined by

$$\lambda_{\pm h} = \begin{cases} -\dfrac{1}{2}\left(\delta + \dfrac{\pi^4}{l^4}\gamma h^4\right) \pm \dfrac{1}{2}\left[\left(\delta + \dfrac{\pi^4}{l^4}\gamma h^4\right)^2 - \dfrac{4\pi^4 h^2}{l^2}(h^2 - n^2)\alpha\right]^{1/2}, & \\ & h \in \mathbb{N}, \quad h \neq n, \\ -\dfrac{1}{2}(\delta + \mu(n)) \pm \dfrac{1}{2}\left[(\delta + \mu(n))^2 + \dfrac{8\pi^2 n^2}{l^2}\left(\beta + \dfrac{\pi^2 n^2}{l^2}\alpha\right)\right]^{1/2}, & h = n, \end{cases} \tag{2.23}$$

with $\mu(n) = (\pi^4 n^4/l^4)\gamma - (\pi^2 n^2/l^2)(\beta + (\pi^2 n^2/l^2)\alpha)\sigma/k$. The corresponding eigenspace is spanned by

$$\left(\sin \frac{\pi h}{l} x, \lambda_{\pm h} \sin \frac{\pi h}{l} x\right), \qquad 0 \leqq x \leqq l. \tag{2.24}$$

Applying (1.14), (1.15) and Proposition 1.2 gives results quite similar to those of Theorem 2.2.

THEOREM 2.3. *Let $\beta + (\pi^2 n^2/l^2)\alpha < 0$ for some $n \in \mathbb{N}$, and let $\bar{\phi}_j, j = 1, \cdots, n$ be defined by* (2.20). *The following are true*:

(a) *If $\delta > \max\{\delta_1, \delta_2\}$, where*

$$\delta_1 = -\frac{\pi^4}{l^4}\gamma + \frac{\pi^2}{l^2}\left(\beta + \frac{\pi^2}{l^2}\alpha\right)\frac{\sigma}{k}, \qquad \delta_2 = -16\frac{\pi^4}{l^4}\gamma, \tag{2.25}$$

then $\bar{\phi}_1$ is exponentially asymptotically stable: for each $n \in]0, \min\{-\alpha/\gamma, \operatorname{Re}\lambda_1, \operatorname{Re}\lambda_2\}[$ ($\lambda_1\lambda_2$ are given by (2.23) *with $n = 1$) there exist $r > 0$, $C > 0$ such that if $u_0, v_0 \in C^4$ ($[0, l]$) satisfy the compatibility conditions* (2.14), *and $\|u_0 - \bar{\phi}_1\|_{C^4([0,l])} \leqq r$, $\|v_0\|_{C^4([0,l])} \leqq r$, then the solution u of* (0.1)-(0.3) *is defined in $[0, +\infty[\times [0, l]$, and*

$$\begin{aligned} &\|u(t, \cdot) - \bar{\phi}_1\|_{C^4([0,l])} + \|u_t(t, \cdot)\|_{C^4([0,l])} \\ &\qquad \leqq C e^{-\eta t}(\|u_0 - \bar{\phi}_1\|_{C^4([0,l])} + \|v_0\|_{C^4([0,l])}) \quad \text{for all } t \geqq 0. \end{aligned} \tag{2.26}$$

(b) *If $\delta < \max\{\delta_1, \delta_2\}$, then $\bar{\phi}_1$ is unstable. In particular, if $J \in \mathbb{Z}$ is such that $\operatorname{Re}\lambda_j > 0$ for each $j \in J$ (the eigenvalues λ_j are defined in* (2.23) *with $n = 1$) then there exists $r > 0$ such that if $a_j, b_j \in \mathbb{R}$, $|a_j| \leqq r$, $|b_j| \leqq r$ for $j \in J$ and*

$$\begin{aligned} u_0(x) &= \bar{\phi}_1(x) + \sum_{j \in J} a_j \sin \frac{\pi j}{l} x, \qquad 0 \leqq x \leqq l, \\ v_0(x) &= \sum_{j \in J} b_j \sin \frac{\pi j}{l} x, \qquad 0 \leqq x \leqq l, \end{aligned} \tag{2.27}$$

then there exists a backward solution $u:]-\infty, 0] \times [0, l] \to \mathbb{R}$ of (0.1)-(0.3), *$u(t, x) = \bar{\phi}_1(x) + \sum_{j \in J} a_j(t) \sin(\pi j/l)x$ and a_j, $\dot{a}_j$ converge to* 0 *exponentially as $t \to -\infty$.*

(c) *The other stationary solutions $\bar{\phi}_2, \cdots, \bar{\phi}_n$ are unstable for each value of δ. The second statement of* (b) *holds for $\bar{\phi}_2, \cdots, \bar{\phi}_n$, with obvious modifications.*

(d) *Let* $\nu \in \{1, \cdots, n\}$ *and let* $\lambda_j, j \in \mathbb{Z}$, *be defined by* (2.23) *with n replaced by* ν. *Assume that* $\operatorname{Re} \lambda_\nu < 0$, *and let* $J^\nu \subset \mathbb{Z}$ *be the set of all integers j such that* $\operatorname{Re} \lambda_j < 0$. *Then for each* $\eta \in]0, \min\{\alpha/\gamma, -\sup\{\operatorname{Re} \lambda_j, j \in J^\nu\}\}[$ *there exist* $r > 0$, $C > 0$ *such that if* $u_0, v_0 \in C^4([0, l])$ *satisfy* (2.24), $\|u_0 - \bar{\phi}_\nu\|_{C^4([0,l])} \leqq r$, $\|v_0\|_{C^4([0,l])} \leqq r$, *and*

$$\left\langle u_0, \sin \frac{\pi j}{l} x \right\rangle = \left\langle v_0, \sin \frac{\pi j}{l} x \right\rangle = 0 \quad \forall j \in \mathbb{Z} \backslash J^\nu, \tag{2.28}$$

the solution u of (0.1)-(0.3) *is defined for all positive t, and*

$$\begin{aligned} &\|u(t, \cdot) - \bar{\phi}_\nu\|_{C^4([0,l])} + \|u_t(t, \cdot)\|_{C^4([0,l])} \\ &\quad \leqq C e^{-\eta t}(\|u_0 - \bar{\phi}_\nu\|_{C^4([0,l])} + \|v_0\|_{C^4([0,l])}) \quad \forall t \geqq 0. \end{aligned} \tag{2.29}$$

The critical case $\delta = \max\{\delta_1, \delta_2\}$ will be treated in the next section.

Let us remark that (d) differs from (c) of Theorem 2.2 because of the additional assumption $\operatorname{Re} \lambda_\nu < 0$. This is due to the fact that now we have (see (2.22))

$$\begin{aligned} g_\nu(\phi, \psi) &= (k|\phi'|^2 + 2k\langle \phi', \bar{\phi}'_\nu \rangle + \sigma\langle \bar{\phi}'; \psi' \rangle)\phi'' \\ &\quad + (k|\phi'|^2 + \sigma\langle \phi', \psi' \rangle)\bar{\phi}''_\nu, \qquad \nu = 1, \cdots, n \end{aligned} \tag{2.30}$$

so that, using the notation of § 1, there exists $\omega < 0$ such that $g_\nu(X^-) \subset X^-$ if and only if $\bar{\phi}_\nu \in X^-$, that is, if and only if $\operatorname{Re} \lambda_\nu < 0$. Then statement (d) follows from (1.15).

3. Critical cases of stability, bifurcation and stability of periodic orbits.

In this section we shall treat the problem of the stability of the stationary solutions $0, \bar{\phi}_1, -\bar{\phi}_1$, for the values of the parameters $\alpha, \beta, \gamma, \delta, k, \sigma$ which we did not consider in § 2: that is, when $\max\{\operatorname{Re} \lambda; \lambda \in \sigma(\tilde{L})\} = 0$ and when $\max\{\operatorname{Re} \lambda; \lambda \in \sigma(\tilde{L}_1)\} = 0$. In these cases, X^+ may be one-, two- or four-dimensional, so that we shall use classical results about ordinary differential equations in $\mathbb{R}^n$ to study the stability of $0, \bar{\phi}_1, -\bar{\phi}_1 \in X^+$ for the restriction of the semiflow to X^+. By Proposition 1.3(b), asymptotic stability of any stationary solution in X^+ implies its asymptotic stability in $D(L)$: therefore our problem is reduced to a finite-dimensional one. Proposition 1.3(b) holds not only for stationary solutions, but also for small compact sets (near stationary solutions): it will be used for studying the stability of periodic orbits which we shall prove to exist for suitable values of the parameters.

We shall use the following notion of stability, which involves not only u but also u_t.

Definition 3.1. A stationary solution $\bar{\phi}$ of (0.1) is said to be stable in $C^4([0, l])$ if

(i) for any $\varepsilon > 0$ there is $\delta_\varepsilon > 0$ such that if $u_0, v_0 \in C^4([0, l])$ satisfy the compatibility conditions (2.14) and $\|u_0 - \bar{\phi}\|_{C^4([0,l])} + \|v_0\|_{C^4([0,l])} \leqq \delta_\varepsilon$, then the solution u of (0.1)-(0.3) satisfies

$$\|u(t, \cdot) - \bar{\phi}\|_{C^4([0,l])} \leqq \varepsilon, \quad \|u_t(t, \cdot)\|_{C^4([0,l])} \leqq \varepsilon \quad \forall t \geqq 0.$$

It is said to be asymptotically stable in $C^4([0, l])$ if (i) holds and

(ii) there is $\bar{\delta} > 0$ such that for any $\varepsilon > 0$ there exists $T_\varepsilon > 0$ such that if $u_0, v_0 \in C^4([0, l])$ satisfy (2.14) and $\|u_0 - \bar{\phi}\|_{C^4([0,l])} + \|v_0\|_{C^4([0,l])} \leqq \bar{\delta}$, then

$$\|u(t, \cdot) - \bar{\phi}\|_{C^4([0,l])} + \|u_t(t, \cdot)\|_{C^4([0,l])} \leqq \varepsilon \quad \forall t \geqq T_\varepsilon.$$

It is said to be unstable if it is not stable. A time periodic solution $\phi(t, x)$ of (0.1) is said to be asymptotically stable in $C^4([0, l])$ if

(i′) for any $\varepsilon > 0$ there is $\delta_\varepsilon > 0$ such that if $u_0, v_0 \in C^4([0, l])$ satisfy (2.14) and

$$\|u_0 - \phi(t_0, \cdot)\|_{C^4([0,l])} + \|v_0 - \phi_t(t_0, \cdot)\|_{C^4([0,l])} \leqq \delta_\varepsilon$$

for some $t_0 \in \mathbb{R}$, then

$$\|u(t,\cdot)-\phi(t_0+t,\cdot)\|_{C^4([0,l])}+\|u_t(t,\cdot)-\phi_t(t_0+t,\cdot)\|_{C^4([0,l])} \leqq \varepsilon \quad \forall t \geqq 0;$$

(ii′) there is $\bar{\delta}>0$ such that for any $\varepsilon>0$ there exists $T_\varepsilon>0$ such that if $u_0, v_0 \in C^4([0, l])$ satisfy (2.14), and $\|u_0-\phi(t_0,\cdot)\|_{C^4([0,l])}+\|v_0-\phi_t(t_0,\cdot)\|_{C^4([0,l])} \leqq \bar{\delta}$ for some $t_0 \in \mathbb{R}$, then

$$\|u(t,\cdot)-\phi(t_0+t,\cdot)\|_{C^4([0,l])}+\|u_t(t,\cdot)-\phi_t(t_0+t,\cdot)\|_{C^4([0,l])} \leqq \varepsilon \quad \forall t \geqq T_\varepsilon.$$

We begin with the stationary solution $\bar{\phi} \equiv 0$.

3.1. Stability of the null solution and of small periodic orbits near 0. By (2.15), there are two possibilities for $\max\{\operatorname{Re}\lambda;\ \lambda \in \sigma(\tilde{L})\}$ to be zero: either $\delta \geqq \delta_0$, where

$$\delta_0 = -\frac{\pi^4}{l^4}\gamma \tag{3.1}$$

and $\beta+(\pi^2/l^2)\alpha=0$, so that the unique element of $\sigma(\tilde{L})$ on the imaginary axis is 0, or $\delta=\delta_0$ and $\beta+(\pi^2/l^2)\alpha>0$, so that $\sigma(\tilde{L})$ has exactly two purely imaginary conjugate eigenvalues.

Let us consider the first case.

PROPOSITION 3.2. *Let $\delta \geqq \delta_0$, $\beta+(\pi^2/l^2)\alpha=0$. Then $\bar{\phi}\equiv 0$ is stable* (*but not asymptotically stable*) *in $C^4([0, l])$. Moreover, for any $\eta \in]0, \min\{\alpha/\gamma, -\operatorname{Re}\lambda_2\}[$* (*the eigenvalues λ_j, $j \in \mathbb{Z}$ are defined in* (2.15)) *there exist r, $C>0$ such that if u_0, $v_0 \in C^4([0, l])$ satisfy the compatibility conditions* (2.14), *$\|u_0\|_{C^4([0,l])} \leqq r$, $\|v_0\|_{C^4([0,l])} \leqq r$, and*

$$\left\langle u_0, \sin\frac{\pi}{l}x\right\rangle = 0, \tag{3.2}$$

then (2.17) *holds.*

Proof. In this case, 0 is a simple eigenvalue of L and the corresponding eigenspace is spanned by $(\sin(\pi/l)x, 0)$; we have also $\max\{\operatorname{Re}\lambda;\ \lambda\in\sigma(\tilde{L}),\ \lambda\neq 0\} = \max\{-\alpha/\gamma, \operatorname{Re}\lambda_2\}$ (see (2.15), (2.16)). Therefore any $\omega \in [\max\{-\alpha/\gamma, \operatorname{Re}\lambda_2\}, 0[$ satisfies (1.10), and the projection

$$P^+(\phi,\psi)=\left(\left\langle \phi, \sin\frac{\pi}{l}x\right\rangle \sin\frac{\pi}{l}x, 0\right)$$

obviously satisfies (1.11). The statements follow now from Proposition 1.4. □

Let us consider now the case $\delta=\delta_0$, $\beta+(\pi^2/l^2)\alpha>0$.

PROPOSITION 3.3. *Let $\delta=\delta_0$, $\beta+(\pi^2/l^2)\alpha>0$. Then $\bar{\phi}\equiv 0$ is asymptotically stable in $C^4([0, l])$. Moreover, for any $\eta \in]0, \min\{\alpha/\gamma, -\operatorname{Re}\lambda_2\}[$ there exist r, $C>0$ such that if u_0, $v_0 \in C^4([0, l])$ satisfy* (2.14), *$\|u_0\|_{C^4([0,l])} \leqq r$, $\|v_0\|_{C^4([0,l])} \leqq r$, and*

$$\left\langle u_0, \sin\frac{\pi}{l}x\right\rangle = \left\langle v_0, \sin\frac{\pi}{l}x\right\rangle = 0, \tag{3.3}$$

then (2.17) *holds.*

By Theorem 2.2, when $\delta<\delta_0$, the null solution is unstable. But, for δ near δ_0, $\delta<\delta_0$, it can be proved that other solutions are stable (in fact, asymptotically stable). In particular, if $\beta+(\pi^2/l^2)\alpha<0$, the stationary solutions $\bar{\phi}_1$ and $-\bar{\phi}_1$ are stable (see further information later in the text), and if $\beta+(\pi^2/l^2)\alpha>0$, there exists a stable small periodic orbit, as the following proposition states.

PROPOSITION 3.4. *Let $\beta+(\pi^2/l^2)\alpha>0$. Then there exists $\varepsilon<0$ such that for each $\delta\in\,]\delta_0-\varepsilon,\delta_0[$, problem* (0.1), (0.2) *has a time periodic solution $\bar{u}_\delta(t,x)=\phi(\delta,t)\sin(\pi/l)x$, which is asymptotically stable in $C^4([0,l])$. Denoting by T_δ the period of $\bar{u}_\delta(\cdot,x)$, we have*

$$
(3.4)\qquad
\begin{aligned}
&\text{(i)}\quad \lim_{\delta\to\delta_0^-} T_\delta = 2l\left(\beta+\frac{\pi^2}{l^2}\alpha\right)^{-1/2},\\
&\text{(ii)}\quad \lim_{\delta\to\delta_0^-} \max\left\{\|\bar{u}_\delta(t,\cdot)\|_{C^4([0,l])}+\left\|\frac{\partial}{\partial t}\bar{u}_\delta(t,\cdot)\right\|_{C^4([0,l])};\ t\in\mathbb{R}\right\}=0.
\end{aligned}
$$

There exists $r_0>0$ such that $\bar{u}_\delta$ is the unique periodic solution (translations apart) of (0.1), (0.2) *in the class of the $u(t,x)$ such that $\|u(t,\cdot)\|_{C^4([0,l])}\leqq r_0$, $\|u_t(t,\cdot)\|_{C^4([0,l])}\leqq r_0$ for all $t\geqq 0$.*

The stability of the zero solution in Proposition 3.3 and the stability of the periodic orbits in Proposition 3.4 are closely related, so that the two propositions above will be proved together.

Proof of Propositions 3.3 *and* 3.4. Let $\eta>0$ be such that for $\delta\in\,]\delta_0-\eta,\delta_0+\eta[$ we have

$$
(3.5)\qquad
\begin{aligned}
&\text{(i)}\quad \operatorname{Re}\lambda_1=\operatorname{Re}\lambda_{-1}=-\tfrac{1}{2}(\delta-\delta_0),\\
&\text{(ii)}\quad -\tfrac{1}{2}(\delta-\delta_0)>\sup\{\operatorname{Re}\lambda;\ \lambda\in\sigma(L),\ \lambda\neq\lambda_{\pm1}\}=\max\{-\alpha/\gamma,\operatorname{Re}\lambda_2\}.
\end{aligned}
$$

Then, by (3.5)(ii), each $\omega\in[\max\{-\alpha/\gamma,\operatorname{Re}\lambda_2\},\operatorname{Re}\lambda_1[$ satisfies (1.9), and X^+ is spanned by $(\sin(\pi/l)x,0)$ and $(0,\sin(\pi/l)x)$. A projection satisfying (1.11) is, for instance,

$$
(3.6)\qquad P^+(\phi,\psi)=\left(\left\langle\phi,\sin\frac{\pi}{l}x\right\rangle\sin\frac{\pi}{l}x,\left\langle\psi,\sin\frac{\pi}{l}x\right\rangle\sin\frac{\pi}{l}x\right).
$$

For $\delta_0-\eta<\delta<\delta_0+\eta$ set

$$
(3.7)\qquad a(\delta)=-\frac{1}{2}(\delta-\delta_0),\qquad b(\delta)=\frac{1}{2}\left[\frac{4\pi^2}{l^2}\left(\beta+\frac{\pi^2}{l^2}\alpha\right)-(\delta-\delta_0)^2\right]^{1/2}.
$$

By (3.5)(ii), $b(\delta)\in\mathbb{R}$ and $\lambda_{\pm1}=a(\delta)\pm ib(\delta)$. Define a basis for X^+, setting

$$
(3.8)\qquad
\begin{aligned}
e_1&=\left(-\sin\frac{\pi}{l}x,-a(\delta)\sin\frac{\pi}{l}x\right),\\
e_2&=\left(0,b(\delta)\sin\frac{\pi}{l}x\right).
\end{aligned}
$$

With this choice of the basis in X^+ the operator $L_{|X^+}:X^+\to X^+$ may be represented with the matrix

$$
(3.9)\qquad A(\delta)=\begin{pmatrix} a(\delta) & -b(\delta)\\ b(\delta) & a(\delta)\end{pmatrix}.
$$

We have seen that if $(u_0,v_0)\in X^+$, then the solution $(u(t),v(t))$ of (2.1) belongs to X^+ for each t belonging to the interval of existence. Set

$$
(3.10)\qquad (u(t),v(t))=x(t)e_1+y(t)e_2
$$

so that

$$\begin{aligned} \dot{x}(t) &= a(\delta)x(t) - b(\delta)y(t), \\ \dot{y}(t) &= b(\delta)x(t) + a(\delta)y(t) + q(\delta, x(t), y(t)), \end{aligned} \tag{3.11}$$

where $q(\delta, x, y) = \langle f(xe_1 + ye_2), e_2\rangle$ is given by

$$q(\delta, x, y) = \frac{\pi^4}{2l^3 b(\delta)}((k + \sigma a(\delta))x^3 - \sigma b(\delta)x^2 y). \tag{3.12}$$

Moreover, by (3.7)

$$a(\delta_0) = 0, \qquad b(\delta_0) > 0, \qquad a'(\delta_0) = -\tfrac{1}{2} < 0 \tag{3.13}$$

so that the classical Hopf bifurcation assumptions are satisfied for system (3.11). Therefore (see for instance [13, Thm. 3.1] or [16, Thm. 2.1]) there exist $\varepsilon > 0$, $c_0 > 0$, a neighborhood U of 0 in $\mathbb{R}^2$ and a continuous function $[0, c_0[\to]\delta_0 - \varepsilon, \delta_0 + \varepsilon[$, $c \to \delta(c)$ (with $\delta(0) = \delta_0$), such that system (3.11) has a nontrivial periodic solution $(x(t), y(t))$ with $(x(t), y(t)) \in U$ for all $t \in \mathbb{R}$ if and only if there exists $c \in]0, c_0[$ with $\delta = \delta(c)$. Moreover any periodic solution $(x(t), y(t))$ of (3.11) with $(x(t), y(t)) \in U$ for all $t \in \mathbb{R}$ is one of the above (translations apart). Since the data are analytic, our periodic orbits have the same stability properties of the null solution of (3.11) for $\delta = \delta_0$: more precisely, exactly one of the following possibilities holds (see [16, Cor. 4.4]):

(3.14) (i) 0 is asymptotically stable for $\delta = \delta_0$, the periodic orbits given by Hopf's Theorem are asymptotically stable and occur only for $\delta < \delta_0$ (that is, $\delta(c) < \delta_0$ for any c).

(ii) 0 is unstable for $\delta = \delta_0$, and for any sufficiently small $(x_0, y_0) \in \mathbb{R}^2$ system (3.11) has a backward solution $(x(t), y(t))$ such that $x(0) = x_0$, $y(0) = y_0$ and $\lim_{t \to -\infty} x(t) = \lim_{t \to -\infty} y(t) = 0$; the periodic orbits are unstable and occur only for $\delta > \delta_0$.

(iii) 0 is stable but not asymptotically stable for $\delta = \delta_0$; the periodic orbits have the same property and occur only for $\delta = \delta_0$ (that is, $\delta(c) \equiv \delta_0$).

To see that, in our case, (3.14)(i) holds, we may, for instance, use the classical Poincaré procedure (see [16, § 3]) which allows us to construct a Lyapunov function for (3.13) when $\delta = \delta_0$. Using this procedure, we find that the polynomial

$$V(x, y) = x^2 + y^2 + \frac{\pi^4 k}{4l^3 b^2(\delta_0)} x^4 - \frac{\pi^4 \sigma}{8l^3 b(\delta_0)} x^3 y + \frac{\pi^4 \sigma}{8l^3 b(\delta_0)} xy^3 \tag{3.15}$$

is such that

$$\begin{aligned} \dot{V}(x, y) &\doteq -b(\delta_0)y \frac{\partial V}{\partial x}(x, y) + (b(\delta_0)x + q(\delta_0, x, y)) \frac{\partial V}{\partial x}(x, y) \\ &= -\frac{\pi^4}{8l^3}\sigma(x^2 + y^2)^2 + \text{higher order terms} \end{aligned} \tag{3.16}$$

so that $\dot{V}$ is negative for (x, y) near $(0, 0)$. Therefore $(0, 0)$ is asymptotically stable when $\delta = \delta_0$, so that (3.4)(i) holds. The statements of Propositions 3.2 and 3.3 now follow from (b) of Proposition 1.3 and (3.14)(i) (except for the second part of Proposition 3.2, which follows from Theorem 2.2(c), and (3.4)(i), which can be deduced from [16, § 2] or [13, Thm. 3.1]). □

Let us consider system (3.11) for $\delta = \delta_0$. Since $V(x, y)$ defined in (3.15) has degree 4, the origin is said to be 3-*asymptotically stable* in the notation of [16] and it is said to be a *vague attractor* in the notation of [13], [17].

Let us finally remark that the procedure above can be used to show the existence of small periodic solutions of (2.1) for δ near $h^4\delta_0$, for each $h \in \mathbb{N}$, $h > 1$.

3.2. Stability of $\bar{\boldsymbol{\phi}}_1$ and of small periodic orbits near $\bar{\boldsymbol{\phi}}_1$. We have seen in § 2 that if

$$\beta + \frac{\pi^2}{l^2}\alpha < 0 \tag{3.17}$$

then problem (0.1), (0.2) has (at least) two nontrivial stationary solutions $\pm\bar{\phi}_1$, given by (2.20). The spectrum of the corresponding linear operator L_1 consists of the point $\lambda_0 = -\alpha/\gamma$ and of the eigenvalues $\lambda_{\pm h}$, $h \in \mathbb{N}$, defined in (2.23) with $n = 1$. Therefore there are three critical cases of stability: namely, $\delta = \delta_1 > \delta_2$, $\delta = \delta_2 > \delta_1$ and $\delta = \delta_1 = \delta_2$, where δ_1 and δ_2 are defined by (2.25). In the first two cases, there are two eigenvalues on the imaginary axis and the situation is very similar to the one of Propositions 3.3 and 3.4, so the proofs of Propositions 3.5 and 3.6 below will be only outlined. Also in this section we shall consider only $\bar{\phi}_1$; all the results hold, with obvious modifications, for $-\bar{\phi}_1$ also.

PROPOSITION 3.5. *Let $\beta + (\pi^2/l^2)\alpha < 0$. Then, for $\delta = \delta_1 > \delta_2$, $\bar{\phi}_1$ is unstable. In particular, for*

$$\begin{aligned} u_0(x) &= \bar{\phi}_1(x) + a \sin\frac{\pi}{l}x, \qquad 0 \leqq x \leqq l, \\ v_0(x) &= b \sin\frac{\pi}{l}x, \qquad 0 \leqq x \leqq l \end{aligned} \tag{3.18}$$

with $a, b \in \mathbb{R}$, a, b sufficiently small, then there exists a backward solution $u:]-\infty, 0] \times [0, l] \to \mathbb{R}$ of (0.1)-(0.3), *$u(t, x) = \bar{\phi}_1(x) + a(t) \sin(\pi/l)x$ such that $\lim_{t\to-\infty} a(t) = \lim_{t\to-\infty} \dot{a}(t) = 0$. If $\delta = \delta_2 > \delta_1$, there $\bar{\phi}_1$ is asymptotically stable. Moreover, for any $n \in]0, \min\{a/\gamma, \operatorname{Re}\lambda_1, \operatorname{Re}\lambda_3\}[$ there exist $r, C > 0$ such that if $u_0, v_0 \in C^4([0, l])$ satisfy* (2.14), *$\|u_0 - \bar{\phi}_1\|_{C^4([0,l])} \leqq r$, $\|v_0\|_{C^4([0,l])} \leqq r$, and*

$$\left\langle u_0, \sin\frac{2\pi}{l}x\right\rangle = \left\langle v_0, \sin\frac{2\pi}{l}x\right\rangle = 0, \tag{3.19}$$

then (2.26) *holds.*

PROPOSITION 3.6. *Let $\beta + (\pi^2/l^2)\alpha < 0$ and $\delta_1 < \delta_2$ (resp. $\delta_1 > \delta_2$). Then there exists $\varepsilon > 0$ such that for any $\delta \in]\delta_2 - \varepsilon, \delta_2[$ (resp. $\delta \in]\delta_1, \delta_1 + \varepsilon[$), problem* (0.1), (0.2) *has a nonconstant time periodic solution $w_\delta(t, x) = \phi(\delta, t) \sin(2\pi/l)x$ (resp. $w_\delta(t, x) = \phi(\delta, t) \sin(\pi/l)x$), which is asymptotically stable (resp. unstable) in $C^4([0, l])$. Denoting by T_δ the period of $\phi(\delta, \cdot)$, we have*

$$\lim_{\delta\to\delta_2^-} T_\delta = \frac{l^2}{\pi}(3\alpha)^{-1/2}\left(\text{resp. } \lim_{\delta\to\delta_1^+} T_\delta = 2l\left[-2\left(\beta + \frac{\pi^2}{l^2}\alpha\right)\right]^{-1/2}\right),$$

$$\lim_{\delta\to\delta_2^-}\left(\text{resp. } \lim_{\delta\to\delta_1^+}\right) \max\left\{\|w_\delta(t, \cdot) - \bar{\phi}\|_{C^4([0,l])} + \left\|\frac{d}{dt}w_\delta(t, \cdot)\right\|_{C^4([0,l])}, \quad t \in \mathbb{R}\right\} = 0.$$

Moreover, there exists $r > 0$ such that w_δ is the unique periodic solution of (0.1), (0.2) *(translations apart) satisfying $\|w_\delta(t, \cdot) - \bar{\phi}_1\|_{C^4([0,l])} + \|(d/dt)w_\delta(t, \cdot)\|_{C^4([0,l])} \leqq r$ for all $t \in \mathbb{R}$.*

Sketch of the proof of Propositions 3.5 *and* 3.6. We have to use the procedure of Propositions 3.3 and 3.4, with obvious changes: δ_0 must be replaced by $\max\{\delta_1, \delta_2\}$, $u(t)$ by $u(t)-\bar{\phi}_1$, L by L_1 (see (2.21)). Now $a(\delta)$ and $b(\delta)$ are defined by

$$\text{(3.20)}\qquad \begin{aligned} a(\delta) &= -\frac{1}{2}(\delta - \max\{\delta_1, \delta_2\}),\\ b(\delta) &= \begin{cases} b_1(\delta) \doteq \dfrac{1}{2}\left[-\dfrac{8\pi^2}{l^2}\left(\beta + \dfrac{\pi^2}{l^2}\alpha\right) - (\delta - \delta_1)^2\right]^{1/2} & \text{if } \delta_1 > \delta_2,\\ b_2(\delta) \doteq \dfrac{1}{2}\left[48\dfrac{\pi^4}{l^4}\alpha - (\delta - \delta_2)^2\right]^{1/2} & \text{if } \delta_1 < \delta_2 \end{cases} \end{aligned}$$

and $q(\delta, x, y)$ in (3.11) is given by

$$\text{(3.21)}\qquad q(\delta, x, y) = \begin{cases} \text{(i)}\ \dfrac{\pi^3}{2l^2 b(\delta)}\left[-\dfrac{2}{kl}\left(\beta + \dfrac{\pi^2}{l^2}\alpha\right)\right]^{1/2} \\ \quad \cdot[(-3k - 2\sigma a(\delta))x^2 + 2\sigma b(\delta)xy] \\ \quad\quad + \dfrac{\pi^4}{2l^3 b(\delta)}[(k + \sigma a(\delta))x^3 - \sigma b(\delta)x^2 y] \\ \quad \doteq Ax^2 + Bxy + Cx^3 + Dx^2 y \quad \text{if } \delta_1 > \delta_2, \\ \text{(ii)}\ \dfrac{8\pi^4}{l^3 b(\delta)}[(k + \sigma a(\delta))x^3 - \sigma b(\delta)x^2 y] \\ \quad \doteq Ex^3 + Fx^2 y \quad \text{if } \delta_1 < \delta_2. \end{cases}$$

Then the polynomials

$$\begin{aligned} V_1(x, y) &= x^2 + y^2 + \frac{1}{b(\delta_1)}\left[\frac{2}{3}Ax^3 - \frac{2}{3}By^3 + \frac{1}{2}Cx^4 + \frac{1}{4}\left(D - \frac{AB}{b(\delta_1)}\right)x^3 y\right.\\ &\qquad\qquad \left. - \frac{1}{4}\left(D - \frac{AB}{b(\delta_1)}\right)xy^3 + \frac{1}{2b(\delta_1)}By^4\right] \quad \text{if } \delta_1 > \delta_2,\\ V_2(x, y) &= x^2 + y^2 + \frac{1}{2b(\delta_1)}\left[Ex^4 + \frac{1}{2}F(x^3 y - xy^3)\right] \quad \text{if } \delta_1 < \delta_2 \end{aligned}$$

(constructed by the Poincaré procedure) satisfy

$$\begin{aligned} \dot{V}_1(x, y) &= \frac{1}{4}\left(D - \frac{AB}{b(\delta_1)}\right)(x^2 + y^2)^2 + \text{higher order terms}\\ &= \frac{\pi^4\sigma}{4l^3}(x^2 + y^2)^2 + \text{h.o.t.},\\ \dot{V}_2(x, y) &= \frac{F}{4}(x^2 + y^2)^2 + \text{h.o.t.} = -\frac{2\pi^4\sigma}{l^3}(x^2 + y^2)^2 + \text{h.o.t.} \end{aligned}$$

Therefore, if $\delta_1 > \delta_2$, then (3.14)(ii) holds and the instability statements of Propositions 3.5 and 3.6 follow easily. If $\delta_1 > \delta_2$, then (3.17)(i) holds and the stability results of Propositions 3.5 and 3.6 follow by (1.15) and Proposition 1.3(b). □

The case $\delta = \delta_1 = \delta_2$ is much more complicated because it requires the study of a four-dimensional system. The problem of the stability (in the critical cases) of the stationary solutions of o.d.e. in $\mathbb{R}^4$ has not yet been completely solved, nor under nonresonance assumptions. However, in our case, we will obtain an equation which can be treated again by the Poincaré procedure, since (0.4) implies the nonresonance condition

$$\delta_1=\delta_2 \Rightarrow b_1(\delta_1)/b_2(\delta_1)\in\mathbb{R}\backslash\mathbb{Q} \tag{3.22}$$

($b_1(\delta)$ and $b_2(\delta)$ are defined in (3.20)).

Proposition 3.7. *Let $\delta=\delta_1=\delta_2$. Then $\bar{\phi}_1$ is asymptotically stable in $C^4([0,l])$. Moreover, for any $\eta\in\,]0,\min\{\alpha/\gamma,-\operatorname{Re}\lambda_3\}[$ (λ_3 is defined in* (2.23) *with $n=1$) there exist $r, C>0$ such that if $u_0, v_0\in C^4([0,l])$ satisfy* (2.14), *$\|u_0-\bar{\phi}_1\|_{C^4([0,l])}\leqq r$, $\|v_0\|_{C^4([0,l])}<r$, and*

$$\left\langle u_0,\sin\frac{\pi}{l}x\right\rangle=\left\langle u_0,\sin\frac{2\pi}{l}x\right\rangle=\left\langle v_0,\sin\frac{\pi}{l}x\right\rangle=\left\langle v_0,\sin\frac{2\pi}{l}x\right\rangle=0,$$

then (2.26) *holds.*

Proof. A procedure similar to the one employed in Propositions 3.4–3.6 gives rise to the system

$$\begin{aligned}
\dot{x}_1(t)&=a(\delta)x_1(t)-b_1(\delta)y_1(t),\\
\dot{y}_1(t)&=b_1(\delta)x_1(t)+a(\delta)y_1(t)+q_1(\delta,x_1(t),y_1(t),x_2(t),y_2(t)),\\
\dot{x}_2(t)&=a(\delta)x_2(t)-b_2(\delta)y_2(t),\\
\dot{y}_2(t)&=b_2(\delta)x_2(t)+a(\delta)y_2(t)+q_2(\delta,x_1(t),y_1(t),x_2(t),y_2(t)),
\end{aligned} \tag{3.23}$$

where a, b_1, b_2 are defined in (3.20), and

$$\begin{aligned}
&q_1(\delta,x_1,y_1,x_2,y_2)\\
&\quad=-\frac{\pi^3}{2l^2b_1(\delta)}\sqrt{\frac{2}{kl}\left(-\beta-\frac{\pi^2}{l^2}\alpha\right)}\\
&\qquad\cdot[3kx_1^2+4kx_2^2-2\sigma x_1(-x_1a(\delta)+y_1b_1(\delta))-4\sigma x_2(-x_2a(\delta)+y_2b_2(\delta))]\\
&\qquad-\frac{\pi^4}{2l^3b_1(\delta)}[-kx_1^3-4kx_1x_2^2+\sigma x_1^2(-x_1a(\delta)+y_1b_1(\delta))\\
&\qquad\qquad+4\sigma x_1x_2(-x_2a(\delta)+y_2b_2(\delta))]\\
&\quad\doteq c(\delta)x_1^2+d(\delta)x_2^2+e(\delta)x_1y_1+f(\delta)x_2y_2+g(\delta)x_1^3\\
&\qquad+h(\delta)x_1x_2^2+l(\delta)x_1^2y_1+m(\delta)x_1x_2y_2,
\end{aligned} \tag{3.24}$$

$$\begin{aligned}
&q_2(\delta,x_1,y_1x_2,y_2)\\
&\quad=-\frac{2\pi^3}{l^2b_2(\delta)}\sqrt{\frac{2}{kl}\left(-\beta-\frac{\pi^2}{l^2}\alpha\right)}\\
&\qquad\cdot[2kx_1x_2-\sigma x_2(-x_1a(\delta)+y_1b_1(\delta))]-\frac{2\pi^4}{l^3b_2(\delta)}\\
&\qquad\cdot[-kx_1^2x_2-4kx_2^3+\sigma x_1x_2(-x_1a(\delta)+y_1b_1(\delta))\\
&\qquad\qquad+4\sigma x_2^2(-x_2a(\delta)+y_2b_2(\delta))]\\
&\quad\doteq p(\delta)x_1x_2+q(\delta)x_2y_1+r(\delta)x_1^2x_2+s(\delta)x_2^3+t(\delta)x_1x_2y_1+u(\delta)x_2^2y_2.
\end{aligned} \tag{3.25}$$

Using the generalization of the Poincaré method described in [18] we construct a Lyapunov function for system (3.23) when $\delta=\delta_1=\delta_2$:

$$\begin{aligned}
V(x_1,y_1,x_2,y_2)&=x_1^2+y_1^2+x_2^2+y_2^2+Ax_1^3+By_1^3+Cx_1x_2^2+Dx_2^2y_1\\
&\quad+Ex_1y_2^2+Fy_1y_2^2+Gx_1x_2y_2+Hx_2y_1y_2,
\end{aligned}$$

where

$$A=\frac{2c(\delta)}{3b_1(\delta)},\qquad B=-\frac{2e(\delta)}{3b_1(\delta)},\qquad C=\frac{b_2(\delta)(2b_1(\delta)p(\delta)-4b_2(\delta)d(\delta))}{b_1(\delta)(4b_2^2(\delta)-b_1^2(\delta))},$$

$$D=\frac{2b_2(\delta)(q(\delta)+f(\delta))}{4b_2^2(\delta)-b_1^2(\delta)},\qquad E=\frac{b_2(\delta)(4b_2(\delta)d(\delta)-2b_1(\delta)p(\delta))}{b_1(\delta)(4b_2^2(\delta)-b_1^2(\delta))},$$

$$F=\frac{-2b_2(\delta)(q(\delta)+f(\delta))}{4b_2^2(\delta)-b_1^2(\delta)},\qquad G=\frac{-2b_1(\delta)(q(\delta)+f(\delta))}{4b_2^2(\delta)-b_1^2(\delta)},$$

$$H=-\frac{4b_2(\delta)d(\delta)-2b_1(\delta)p(\delta)}{4b_2^2(\delta)-b_1^2(\delta)}.$$

Then (see [18, § 3]) $\dot V$ is negative definite near $(0,0,0,0)$ if, setting for $\delta=\delta_1=\delta_2$,

$$H=\frac{2}{7b_1(\delta)}(b_1(\delta)l(\delta)-e(\delta)),$$

$$K=\frac{2}{7}\left(u(\delta)+\frac{b_1(\delta)p(\delta)f(\delta)+2b_2(\delta)q(\delta)d(\delta)}{4b_2^2(\delta)-b_1^2(\delta)}\right),$$

$$S=\frac{b_1(\delta)q(\delta)p(\delta)}{4b_2^2(\delta)-b_1^2(\delta)},$$

$$Q=\frac{1}{2}\left(-\frac{e(\delta)}{b_1(\delta)}-\frac{2b_2(\delta)q(\delta)d(\delta)}{4b_2^2(\delta)-b_1^2(\delta)}-\frac{b_2(\delta)p(\delta)f(\delta)}{4b_2^2(\delta)-b_1^2(\delta)}\right),$$

we have $H<0$, $K<0$ and HQ or $KS\geqq 0$; $\dot V$ is positive definite if H or K is positive and HQ or $KS\geqq 0$. In our case (see (3.24), (3.25)) H is negative, K and S have the same sign of $b_1^2(\delta)-4b_2^2(\delta)$. Therefore the null solution of (3.23) is asymptotically stable if $4b_2^2(\delta)-b_1^2(\delta)=\beta+(25\pi^2/l^2)\alpha$ is positive, and it is unstable if $\beta+(25\pi^2/l^2)\alpha<0$. In our case (see (0.4)) we have $\beta+(25\pi^2/l^2)\alpha=(9\pi^2/l^2)(EI/\rho)>0$. The conclusions follow now from Proposition 1.3(b) and (1.15). □

We have seen before that the problem of existence and stability of small periodic orbits for system (3.11) is closely related to the problem of the stability of the null solution. The same thing does not happen, in general, for higher-dimensional systems. For studying small periodic solutions of (3.23) we shall use the method employed in [15].

PROPOSITION 3.8. *Let $\beta+(\pi^2/l^2)\alpha<0$, $\delta=\delta_1=\delta_2$ and assume* (3.23). *Then there exists $\varepsilon>0$ such that for any $\delta\in\,]\delta_1-\varepsilon,\ \delta_1+\varepsilon[$, $\delta\neq\delta_1$, problem* (0.1), (0.2) *has a nonconstant time periodic solution u_δ, given by $u(t,x)=a_1(\delta,t)\sin(\pi/l)x+a_2(\delta,t)\sin(2\pi/l)x$, $t\in\mathbb{R}$, $x\in[0,l]$, with $a_2(\delta,\cdot)\equiv 0$ for $\delta>\delta_1$. u_δ is unstable in $C^4([0,l])$ and, denoting by T_δ the period of u_δ, we have*

$$\lim_{\delta\to\delta_1^-}T_\delta=\frac{l^2}{\pi}(3\alpha)^{-1/2},\qquad \lim_{\delta\to\delta_1^+}T_\delta=2l\left[-2\left(\beta+\frac{\pi^2}{l^2}\alpha\right)\right]^{-1/2},\tag{3.26}$$

$$\lim_{\delta\to\delta_1}\max\left\{\|u_\delta(t,\cdot)-\bar\phi_1\|_{C^4([0,l])}+\left\|\frac{d}{dt}u_\delta(t,\cdot)\right\|_{C^4([0,l])};\ t\in\mathbb{R}\right\}=0.\tag{3.27}$$

Moreover there exists $r>0$ such that if $v(t,x):\mathbb{R}\times[0,l]\to\mathbb{R}$ is a T-periodic solution of (0.1), (0.2) *with*

$$\sup\left\{\|v(t,\cdot)-\bar\phi\|_{C^4([0,l])}+\left\|\frac{d}{dt}v(t,\cdot)\right\|_{C^4([0,l])}\leqq r\right\}$$

and

$$\left|T-\frac{l}{\pi}(3\alpha)^{-1/2}\right|\leqq r\left(\textit{resp.}\left|T-2l\left[-2\left(\beta+\left(\frac{\pi^2}{l^2}\right)\alpha\right)\right]^{-1/2}\right|\leqq r\right),$$

then $v(t,\cdot)=v_\delta(t+r,\cdot)$ *for some* $\tau\in\mathbb{R}$, $\delta\in\,]\delta_1-\varepsilon,\delta_1[$ (*resp.* $\delta\in\,]\delta_1,\delta_1+\varepsilon[$).

Proof. Let us consider again system (3.23). Existence and uniqueness of small periodic solutions satisfying (3.26) may be proved by applying Hopf's Theorem twice (see [13, Thm. 3.15], since the transversality condition $a'(\delta_1)\neq 0$ holds. We find that the periodic orbits corresponding to the eigenvalues $a(\delta)\pm ib_1(\delta)$ (resp. $a(\delta)\pm ib_2(\delta)$) occur for $\delta>\delta_1$ (resp. $\delta<\delta_1$).

Let us consider first the couple of eigenvalues $a(\delta)\pm ib_1(\delta)$, where $a(\delta)$, $b_1(\delta)$ are defined in (3.20). The manifold $x_2=y_2=0$ is invariant for system (3.23). In particular, setting $x_2=y_2=0$, the (x_1,y_1)-part of (3.23) takes the form

$$x_1=-b_1(\delta)y_1,\qquad y_1=b_1(\delta)x_1+q(\delta,x_1,y_1), \tag{3.28}$$

where $q(\delta,x,y)$ is defined in (3.21)(i). Therefore (see the proof of Proposition 3.6) the small periodic solutions of (3.28) occurring for $\delta>\delta_1$ are unstable. Setting now $u_\delta(t,x)=\bar{x}_1(t)\sin(\pi/l)x$, u_δ satisfies all the statements of Proposition 3.8 (the uniqueness follows from Proposition 1.3(a)).

Let us consider now the second couple of eigenvalues $a(\delta)\pm ib_2(\delta)$. The existence of another invariant manifold $(x_1y_1)=\phi(x_2,y_2)$ (which could allow us to reduce again (3.23) to a two-dimensional system) is not evident. To study the stability properties of the periodic solutions of (3.23) given by Hopf's Theorem we shall use the procedure of [15]. To use the notation of [15], we replace t by $\tau=t/b_2(\delta_1)$. Setting $x(t)=x_2(t/b_2(\delta_1))$, $y(t)=y_2(t/b_2(\delta_1))$, $z_1(t)=x_1(t/b_2(\delta_1))$, $z_2(t)=y_1(t/b_2(\delta_1))$, and taking $\delta=\delta_1$, (3.23) becomes

$$\begin{aligned}\dot{x}&=-y, & \dot{y}&=x+\frac{1}{b_1(\delta_1)}q_2(\delta_1,z_1,z_2,x,y),\\ \dot{z}_1&=-\omega z_2, & \dot{z}_2&=\omega z_1+\frac{1}{b_2(\delta_1)}q_1(\delta_1,z_1,z_2,x,y),\end{aligned} \tag{3.29}$$

where $\omega=b_1(\delta_1)/b_2(\delta_1)$ and q_1, q_2 are defined in (3.24), (3.25). Thanks to the nonresonance condition (3.22), for each $h\in\mathbb{N}$ it is possible to find a polynomial $\Phi^{(h)}:\mathbb{R}^2\to\mathbb{R}^2$ of degree h, such that

$$\frac{d}{dt}(z-\Phi^{(h)}(x,y))|_{z=\Phi^{(h)}(x,y)}=o(x^2+y^2)^{h/2},$$

where d/dt is evaluated along the solutions of (3.29). In particular, for $h=3$, we get

$$\begin{aligned}\Phi^{(3)}(x,y)&=(A_1x^2+B_1xy+C_1y^2,\ A_2x^2+B_2xy+C_2y^2)\\&=(\Phi_1^{(3)}(x,y),\Phi_2^{(3)}(x,y)),\end{aligned}$$

where

$$A_1=\frac{d(\delta_1)}{\omega b_2(\delta_1)}\frac{2-\omega^2}{\omega^2-4},\qquad B_1=-\frac{\omega f(\delta_1)}{b_2(\delta_1)(\omega^2-4)},\qquad C_1=\frac{2d(\delta_1)}{\omega b_2(\delta_1)(\omega^2-4)},$$

$$A_2=\frac{f(\delta_1)}{b_2(\delta_1)(\omega^2-4)},\qquad B_2=-\frac{2d(\delta_1)}{b_2(\delta_1)(\omega^2-4)},\qquad C_2=-\frac{f(\delta_1)}{b_2(\delta_1)(\omega^2-4)}$$

and the functions d, f are defined in (3.24) and (3.25). When we set $z = \Phi^{(3)}(x, y)$ in (3.29), the (x, y)-part becomes

$$\begin{aligned} \dot{x} &= -y, \\ \dot{y} &= x + \frac{p(\delta_1)}{b_2(\delta_1)} x\Phi_1^{(3)}(x, y) + \frac{q(\delta_1)}{b_2(\delta_1)} x\Phi_2^{(3)}(x, y) \\ &\quad + \frac{s(\delta_1)}{b_2(\delta_1)} x^3 + \frac{u(\delta_1)}{b_2(\delta_1)} x^2 y + \frac{r(\delta_1)}{b_2(\delta_1)} (\Phi_1^{(3)}(x, y))^2 \\ &\quad + \frac{t(\delta_1)}{b_2(\delta_1)} \Phi_1^{(3)}(x, y)(\Phi_1^{(3)}(x, y))^2. \end{aligned} \tag{3.30}$$

By the Poincaré procedure we find that the polynomial

$$\begin{aligned} V(x, y) &= x^2 + y^2 + \frac{c(\delta_1)}{2} x^4 + \frac{1}{4}\left(\frac{p(\delta_1)}{b_2(\delta_1)} B_1 + \frac{q(\delta_1)}{b_2(\delta_1)} B_2\right) \\ &\quad \cdot (x^3 y + xy^3) - \frac{e(\delta_1)}{2} y^4 \end{aligned}$$

satisfies

$$\begin{aligned} \dot{V}(x, y) &= \frac{1}{4}\left(\frac{p(\delta_1)}{b_2(\delta_1)} B_1 + \frac{q(\delta_1)}{b_2(\delta_1)} B_2\right)(x^2 + y^2) + \text{h.o.t.} \\ &\doteq G_3(x^2 + y^2)^2 + \text{h.o.t.}, \end{aligned} \tag{3.31}$$

so that V is negative definite for small (x, y) (since $\delta_1 = \delta_2$ implies $\omega^2 - 4 > 0$; see (0.4)). The definiteness of V implies (see [15, Thms. 2.2, 3.1, 4.2]) that the stability properties of the bifurcating orbits may be recognized by studying the auxiliary system (3.30)-(3.32) and its normal form, where

$$\begin{aligned} \dot{\chi}_1 &= -\frac{G_3}{2}(x^2 + y^2)\chi_1 - \omega\chi_2, \\ \dot{\chi}_2 &= \omega\chi_1 - \frac{G_3}{2}(x^2 + y^2)\chi_2 + \frac{1}{b_2(\delta_1)} \\ &\quad \cdot [2c(\delta_1)\Phi_1^{(3)}(x, y) + e(\delta_1)\Phi_1^{(3)}(x, y) + h(\delta_1)x^2 + m(\delta_1)xy]\chi_1 \\ &\quad + \frac{e(\delta_1)}{b_2(\delta_1)} \Phi_1^{(3)}(x, y)\chi_2. \end{aligned} \tag{3.32}$$

The normal form of (3.32) (see [1], [20]) may be obtained by a change of variables $\chi = \xi + L(x, y)\xi$, where $L(x, y)\xi = (\phi_{11}(x, y)\xi_1 + \phi_{12}(x, y)\xi_2,\ \phi_{21}(x, y)\xi_1 + \phi_{22}(x, y)\xi_2)$ and $\phi_{ij}(j = 1, 2)$ are polynomials of degree 3 vanishing at the origin, such that (3.32) becomes

$$\begin{aligned} \dot{\xi}_1 &= P(x^2 + y^2)\xi_1 - (\omega + Q(x^2 + y^2))\xi_2 + \psi_1(x, y)\xi, \\ \dot{\xi}_2 &= (\omega + Q(x^2 + y^2))\xi_1 + P(x^2 + y^2)\xi_2 + \psi_2(x, y)\xi, \end{aligned} \tag{3.33}$$

where $P, Q \in \mathbb{R}$ and $\psi_j(x, y) = o(x^2 + y^2)^{3/2}$. Then ([15, Thm. 4.3]) the bifurcating orbits given by Hopf's Theorem are stable if both G_3 and P are negative (G_3 is given by

(3.31)), they are unstable if one of them is positive. In our case we get

$$P=-\frac{G_3}{2}-\frac{d(\delta_1)}{b_1(\delta_2)}$$

so that the orbits are unstable. Now it is sufficient to set $u_\delta(t,x)=-\bar{x}_1(t)\sin(\pi/l)x-\bar{x}_2(t)\sin(2\pi/l)x$ (where $(\bar{x}_1(\cdot), \bar{y}_1(\cdot), \bar{x}_2(\cdot), \bar{y}_2(\cdot))$ is the periodic solution of (3.32)), and the proof is finished. □

In Tables 1 and 2 we summarize the stability properties of the stationary solutions $0, \bar{\phi}_1, -\bar{\phi}_1$.

TABLE 1
Stability of the null solution.

	$\beta+\frac{\pi^2}{l^2}\alpha<0$	$\beta+\frac{\pi^2}{l^2}\alpha=0$	$\beta+\frac{\pi^2}{l^2}\alpha>0$
$\delta<\delta_0$	unstable Theorem 2.2(b)	unstable Theorem 2.2(b)	unstable Theorem 2.2(b)
$\delta=\delta_0$	unstable Theorem 2.2(b)	stable Proposition 3.2	asymptotically stable Proposition 3.3
$\delta>\delta_0$	unstable Theorem 2.2(b)	stable Proposition 3.2	asymptotically stable Theorem 2.2(a)

TABLE 2
Stability of $\bar{\phi}_1$ and $-\bar{\phi}_1$ $(\beta<-(\pi^2/l^2)\alpha)$.

$\delta<\max\{\delta_1,\delta_2\}$	unstable	Theorem 2.3(b)
$\delta=\delta_1>\delta_2$	unstable	Proposition 3.5
$\delta=\delta_1=\delta_2$	asymptotically stable	Proposition 3.7
$\delta=\delta_2>\delta_1$	asymptotically stable	Proposition 3.5
$\delta>\max\{\delta_1,\delta_2\}$	asymptotically stable	Theorem 2.3(a)

$$\delta_0=-\frac{\pi^4}{l^4}\gamma,\qquad \delta_1=-\frac{\pi^4}{l^4}\gamma+\frac{\pi^2}{l^2}\left(\beta+\frac{\pi^2}{l^2}\alpha\right)\frac{\sigma}{k},\qquad \delta_2=-\frac{16\pi^4}{l^4}\gamma.$$

Appendix.

Proofs of the propositions of § 1.

Proof of Proposition 1.1. The proof is straightforward, and it will be only outlined.

Let $\bar{w}\in D(L)$, and let $r>0$ be such that g is Lipschitz continuous on the closed ball centered at $\bar{w}$ with radius r.

Let $0\leqq t_0<t_1$, $w_0\in D(L)$, $\|w_0-\bar{w}\|_{D(L)}\leqq r/2$. Consider the problem

$$\text{(A.1)}\qquad \dot{w}=Lw+g(w),\quad t_0\leqq t\leqq t_1,\qquad w(t_0)=w_0.$$

A function $w\in C^1([t_0,t_1];X)\cap C([t_0,t_1];D(L))$ is a solution of (A.1) if and only if

$$\text{(A.2)}\qquad w(t)=e^{(t-t_0)L}w_0+\int_{t_0}^{t}e^{(t-s)L}g(w(s))\,ds,\qquad t_0\leqq t\leqq t_1.$$

Set

$$Y=\{w\in C([t_0,t_1];D(L));\ \|w(\cdot)-\bar{w}\|_{C([t_0,t_1];D(L))}\leqq r\}$$

$$\Gamma\colon Y\to C([t_0,t_1];D(L)),\qquad \Gamma(w)(t)=e^{(t-t_0)L}w_0+\int_{t_0}^{t}e^{(t-s)L}g(w(s))\,ds.$$

In fact, $\Gamma(Y)\subset C([t_0,t_1];D(L))$ because for each $w\in Y, g(w(\cdot))$ belongs to $C([t_0,t_1];D_L(\theta,\infty))$ and $w_0\in D(L)$, $Lw_0\in X=\overline{D(L)}$ (see [19, Thm. 5.5]). Moreover, using the estimates (see [19, § 1])

$$\|e^{sL}\|_{\mathscr{L}(D(L))}\leqq K,\qquad \|s^{1-\theta}\,e^{sL}\|_{\mathscr{L}(D_L(\theta,\infty);D(L))}\leqq K,\qquad 0\leqq s\leqq t_1-t_0,$$

we can easily see that if t_1-t_0 and r are sufficiently small, then $\Gamma(Y)\subset Y$ and Γ is a contraction with constant $\alpha<1$, not depending on w_0, so that Γ has a unique fixed point w in Y. It is also easy to show that in fact w is the unique solution of (A.1), and that if $w_{0n}\in D(L)(n\in\mathbb{N})$ are such that $\lim_{n\to\infty}\|w_0-w_{0n}\|_{D(L)}=0$, then, denoting by w_n the solution of (A.1) with initial value w_{0n}, we have $\lim_{n\to+\infty}\|w-w_n\|_{C([t_0,\,t_1];D(L))}=0$. Consider again the solution of (A.1): we have $w(t_1)\in D(L)$, so that w can be continued to some interval $[t_1,t_2]$. Set

$$\tau=\sup\{T>0;\ (1.3)\text{ has a solution in }[0,T]\}.$$

Then $[0,\tau[$ is the maximal interval of existence of the solution of (1.3). The last statement may be proved as in [10, Thm. 4.11], with obvious modifications. □

Proof of Proposition 1.2. For $\eta\in]0,\bar{\omega}[$ let $M(\eta)>0$ be such that

$$\|t^{1-\theta}\,e^{tL}\|_{L(D_L(\theta,\infty),D(L))}\leqq M(\eta)\,e^{-\eta t}\quad\text{for all } t>0,$$

$$\|e^{tL}\|_{L(D(L))}\leqq M(\eta)\,e^{-\eta t}\quad\text{for all } t>0.$$

Fix $0<\eta<\eta_1<\bar{\omega}$ and let $R,K>0$ be such that $\|x\|_{D(L)}\leqq R\Rightarrow\|g(x)\|_{D_L(\theta,\infty)}\leqq K\|x\|^2_{D(L)}$. Let $r\leqq R$ be such that $(M(\eta_1)K\Gamma(\theta)/(\eta_1-\eta)^\theta)r^2\leqq r/3$ (Γ denotes the Euler Gamma Function). We shall show that, if $w_0\in D(L)$ and $\|w_0\|_{D(L)}<r/3M(\eta)$, then $\tau(w_0)=+\infty$ and (1.7) holds. Assume by contradiction that $\tau(w_0)<+\infty$. Then, setting

$$T=\sup\{t>0,\ \tau(w_0)>2t,\ \|w(s)\,e^{\eta s}\|_{D(L)}\leqq r, 0\leqq s\leqq t\},$$

we have $T<+\infty$, and $\|w(T)\,e^{\eta T}\|_{D(L)}=r$ because of the continuity of w. Therefore

$$r=\|w(T)\,e^{\eta T}\|_{D(L)}\leqq M(\eta)\|w_0\|_{D(L)}+\left\|\int_0^T e^{(T-s)L}g(w(s))\,ds\right\|_{D(L)}$$

$$\leqq M(\eta)\|w_0\|_{D(L)}+e^{\eta T}M(\eta_1)\int_0^T e^{-\eta_1(T-s)}(T-s)^{\theta-1}Kr^2\,e^{-2\eta s}\,ds$$

$$\leqq\frac{r}{3}+M(\eta_1)Kr^2\int_0^T e^{(-\eta_1+\eta)(T-s)}(T-s)^{\theta-1}\,ds\leqq\frac{2}{3}r.$$

Hence $\tau(w_0)=T=+\infty$, and (1.7) holds with $C(\eta)=(3/2)M(\eta)$. □

Proof of Proposition 1.3. The proof follows the same lines as in [3, Thms. 3.3, 3.4], but is simpler because of the invariance properties of X^+.

(a) Set $\omega_1 = \sup\{\operatorname{Re}\lambda;\, \lambda \in \sigma_\omega^-(L)\}$, $w_2 = \min\{\operatorname{Re}\lambda;\, \lambda \in \sigma_\omega^+(L)\}$, $\omega_3 = \max\{\operatorname{Re}\lambda;\, \lambda \in \sigma_\omega^+(L)\}$, and define the operators:

$$\begin{aligned} &L_1\colon D(L_1) = D(L)\cap X^- \to X^-, \qquad L_1x = Lx,\\ &L_2\colon D(L_2) = X^+ \to X^+, \qquad L_2x = Lx. \end{aligned} \tag{A.3}$$

Then $D_{L_1}(\theta,\infty) = D_L(\theta,\infty)\cap X^-$. For $\eta \in \,]0, -\omega_1[$ let $M_1(\eta) > 0$ be such that

$$\begin{aligned} &\|e^{tL_1}\|_{\mathscr{L}(D(L_1))} \leqq M_1(\eta)\, e^{-\eta t}, \ \|t^{1-\theta}\, e^{tL_1}\|_{\mathscr{L}(D_L(\theta,\infty);D(L_1))}\\ &\qquad \leqq M_1(\eta)\, e^{-\eta t} \quad \forall t > 0 \end{aligned} \tag{A.4}$$

and, for $\varepsilon > 0$, let $M_2(\varepsilon) > 0$ be such that

$$\begin{aligned} &\|e^{tL_2}\|_{\mathscr{L}(X^+)} \leqq M_2(\varepsilon)\, e^{(\omega_2-\varepsilon)t} \quad \forall t \leqq 0,\\ &\|e^{tL_2}\|_{\mathscr{L}(X^+)} \leqq M_3(\varepsilon)\, e^{(\omega_3+\varepsilon)t} \quad \forall t \geqq 0. \end{aligned} \tag{A.5}$$

For $\rho > 0$ (to be chosen later) consider the system

$$\begin{aligned} &\dot{x}(t) = L_1x(t) + g_1(x(t), y(t)), \qquad x(0) = x_0,\\ &\dot{y}(t) = L_2y(t) + g_2(x(t), y(t)), \qquad y(0) = y_0, \end{aligned} \tag{A.6}$$

where

$$\begin{aligned} &x_0 = P^-w_0, \qquad y_0 = P^+w_0,\\ &g_1\colon D(L_1)\times X^+ \to D_{L_1}(\theta,\infty), \qquad g_1(x,y) = P^-g(x + \psi(y/\rho)y),\\ &g_2\colon D(L_1)\times X^+ \to X^+, \qquad g_2(x,y) = P^+g(x+\psi(y/\rho)y) \end{aligned}$$

and $\psi\colon X^+ \to \mathbb{R}$ is a C^∞ function such that

$$\begin{aligned} &0 \leqq \psi(x) \leqq 1 \quad \forall x \in X^+,\\ &\psi(x) = 1 \quad \text{if } \|x\| \leqq \tfrac{1}{2}, \qquad \psi(x) = 0 \quad \text{if } \|x\| \geqq 1. \end{aligned}$$

System (A.6) is equivalent (setting $x(t) = P^-w(t)$, $y(t) = P^+w(t)$) to (3.1) when $\|w(t)\|_{D(L)}$ is sufficiently small. Proposition 1.1 may be applied to system (A.6) to get local existence and uniqueness of the solution. Let

$$\begin{aligned} K(\rho) = \max\Bigg\{ &\left\|\frac{\partial g_1}{\partial x}(x,y)\right\|_{\mathscr{L}(D(L_1),D_{L_1}(\theta,\infty))}, \left\|\frac{\partial g_1}{\partial y}(x,y)\right\|_{\mathscr{L}(X^+,D_{L_1}(\theta,\infty))},\\ &\left\|\frac{\partial g_2}{\partial x}(x,y)\right\|_{\mathscr{L}(D(L_1),X^+)}, \left\|\frac{\partial g_2}{\partial y}(x,y)\right\|_{\mathscr{L}(X^+)},\\ &\qquad\qquad y \in X^+, x \in D(L_1), \|x\|_{D(L_1)} \leqq \rho \Bigg\}. \end{aligned} \tag{A.7}$$

Then $\lim_{\rho\to 0} K(\rho) = 0$.

It is easy to see (arguing as in the proof of Proposition 1.2) that there exist $\bar{\rho}, \bar{c} > 0$ such that for $0 < \rho < \bar{\rho}$ and $\|x_0\|_{D(L_1)} \leqq \bar{c}\rho$ then $\|x(t)\|_{D(L_1)} \leqq \rho$ for each $t \in [0, \tau(x_0, y_0)[$. This a priori estimate implies easily $\tau(x_0, y_0) = +\infty$. Fixed such ρ and x_0, let $t \geqq 0$ and let $z(s,t)(s \in \mathbb{R})$ be the solution of the i.v.p.

$$\dot{z}(s) = L_2z(s) + g_2(0, z(s)), \qquad z(t) = y(t).$$

Then for $0 \leqq s \leqq t$ and $\varepsilon > 0$

$$\|y(s) - z(s, t)\| = \left\| \int_t^s e^{(s-\sigma)L_2}[g_2(x(\sigma), y(\sigma)) - g_2(0, z(\sigma, t))]\, d\sigma \right\|$$

$$\leqq M_2(\varepsilon) \int_s^t e^{(\omega_2 - \varepsilon)(s-\sigma)} K(\rho)(\|x(\sigma)\|_{D(L_1)} + \|y(\sigma) - z(\sigma, t)\|)\, d\sigma.$$

Therefore, by Gronwall's Lemma

$$\|y(s) - z(s, t)\| \leqq M_2(\varepsilon) K(\rho) \int_s^t e^{(-\omega_2 + \varepsilon)(\sigma - s) + M_2(\varepsilon)K(\rho)(t-s)}$$

$$\cdot \|x(\sigma)\|_{D(L_1)}\, d\sigma, \qquad 0 \leqq s \leqq t,$$

so that, since $g_1(0, w(s, t)) = 0$, we have, for $\eta \in]\max\{-\omega_2, 0\}, -\omega_1[$ and $\varepsilon \in]0, \omega_2 + \eta[$:

$$\|e^{\eta s} g_1(x(s), y(s))\|_{D_{L_1}(\theta, \infty)}$$

$$\leqq K(\rho)\|e^{\eta s} x(s)\|_{D(L_1)} + M_2(\varepsilon)(K(\rho))^2 \int_s^t e^{(-\omega_2 + \varepsilon - \eta)(\sigma - s)}\, d\sigma\, e^{M_2(\varepsilon)K(\rho)(t-s)}$$

$$\cdot \sup_{s \leqq \sigma \leqq t} \|x(\sigma)\, e^{\eta\sigma}\|_{D(L_1)}$$

$$\leqq K(\rho)\left(1 + \frac{M_2 K(\rho)}{\omega_2 + \eta - \varepsilon}\right) \sup_{0 \leqq \sigma \leqq t} \|x(\sigma)\, e^{\mu\sigma}\|_{D(L_1)}, \quad 0 \leqq s \leqq t$$

if ρ is so small that $\omega_2 + \eta - \varepsilon > M_2(\varepsilon)K(\rho)$. Then we have (since $\eta < -\omega_1$)

$$\|e^{\eta t} x(t)\|_{D(L_1)} = \|e^{\eta t}\left(e^{tL_1} x_0 + \int_0^t e^{(t-s)L_1} g_1(x(s), y(s))\right) ds\|_{D(L_1)}$$

$$\leqq M_1(\eta)\|x_0\|_{D(L_1)} + M_1\left(\frac{\eta - \omega_1}{2}\right) K(\rho)\left(1 + \frac{M_2 K(\rho)}{\omega_2 + \eta - \varepsilon}\right) \int_0^{+\infty} e^{-(\eta - \omega_1)_s/2}$$

$$\cdot s^{\theta - 1}\, ds \sup_{0 \leqq s \leqq t} \|e^{\eta s} x(s)\|_{D(L_1)}.$$

Choose ρ so small that

$$K(\rho) M_1\left(\frac{\eta - \omega_1}{2}\right)\left(1 + \frac{M_2 K(\rho)}{\eta + \omega_2 - \varepsilon}\right)\left(\frac{2}{\eta - \omega_1}\right)^{\theta} \Gamma(\theta) \leqq \frac{1}{2}.$$

Then $\|x(t)\|_{D(L_1)} = \|P^- w(t)\|_{D(L)} \leqq 2M_1(\eta)\, e^{-\eta t}\|x_0\|_{D(L_1)} = 2M_1(\eta)\, e^{-\eta t}\|P^- w_0\|_{D(L_1)}$ for $t \geqq 0$ and the proof of (a) is finished.

(b) Let $\eta \in]0, -\omega_1[$ and $r \in]0, \varepsilon_0(\eta)/2]$ ($\varepsilon_0(\eta)$ is given in (a)).

Let $\Omega \subset X^+$ be compact and asymptotically stable for the semiflow in X^+, and assume $\operatorname{diam} \Omega \leqq r$. For $\delta > 0$ define $I(\Omega, \delta) = \{y \in X^+;\ \operatorname{dist}(y, \Omega) \leqq \delta\}$. Then, if δ is sufficiently small, say $\delta \leqq \bar{\delta}$, there exists a Lyapunov function $V: I(\Omega, \delta) \to \mathbb{R}$ such that

$$|V(y) - V(\bar{y})| \leqq \|y - \bar{y}\| \quad \text{for any } y, \bar{y} \in I(\Omega, \delta),$$

$$V(w(t)) \leqq e^{-t} V(w(0)) \quad \text{if } w(0) \in I(\Omega, \delta),$$

$$\alpha(\operatorname{dist}(y, \Omega)) \leqq V(y) \leqq \operatorname{dist}(y, \Omega) \quad \text{for any } y \in I(\Omega, \delta),$$

where $\alpha: [0, +\infty[\to [0, +\infty[$ is increasing, continuous and $\alpha(0) = 0$.

Let us define a Lyapunov function for system (A.6):

$$W: D(L_1) \times I(\Omega; \delta) \to [0, +\infty[, \qquad W(x, y) = p\|x\| - V(y),$$

where $p > 1$ will be chosen later. We have

(A.8) $$(\operatorname{dist}((x, y), \{0\} \times \Omega)) \leqq W(x, y) \leqq p \operatorname{dist}((x, y), \{0\} \times \Omega),$$

where $\beta^{-1}(\xi)=\xi+\alpha^{-1}(\xi)$ for $\xi\geqq 0$, and the distance is in the norm $\|(x, y)\|=\|x\|_{D(L_1)}+\|y\|$, $x\in D(L_1)$, $y\in X^+$.

Let $z(t)$ $(t\in\mathbb{R})$ be the solution of

$$z'(t)=L_2 z(t)+g_2(0, z(t)), \qquad z(0)=y_0.$$

Then for $\|x_0\|_{D(L)}\leqq \varepsilon_0(\eta)/2$, $\|y_0\|_{D(L)}\leqq \varepsilon_0(\eta)/2$, $y_0\in I(\bar\delta, \Omega)$ and $\varepsilon>0$, we have (using Proposition 1.3(a) and Gronwall's Lemma)

$$\|z(t)-y(t)\|\leqq c_1\, e^{c_2 t}\|x_0\|_{D(L)} \quad \text{for all } t\geqq 0,$$

where

$$c_1=\frac{K(\rho)M_3(\varepsilon)C_0(\eta)}{\omega_3+\varepsilon+\eta}, \qquad c_2=K(\rho)M_3(\varepsilon)-\eta.$$

Since

$$W(x(t), y(t))=V(z(t))+[V(y(t))-V(z(t))]+p\|x(t)\|_{D(L)}, \qquad t\geqq 0,$$

for $T\geqq 0$ and $T\leqq t\leqq 2T$ we have

$$W(x(t), y(t))\leqq e^{-T}V(y_0)+[c_1\, e^{c_2 t}+C_0(\eta)p\, e^{-\eta T}]\|x_0\|_{D(L)}.$$

Let $T>0$ be such that

$$\text{(A.9)} \qquad e^{-T}\leqq \tfrac{1}{2}, \quad C_0(\eta)\, e^{-T}\leqq \tfrac{1}{4}, \quad \alpha^{-1}(e^{-T}\bar\delta)\leqq\bar\delta$$

and choose p so large that $c_1\, e^{2c_2 T}\leqq p/4$; then $W(x(t), y(t))\leqq\frac{1}{2}W(x_0, y_0)$ for all $t\in[T, 2T]$. Then for

$$\|x_0\|_{D(L)}\leqq\frac{\bar\delta}{2c_1}\, e^{-c_2 T}\wedge\frac{\varepsilon_0(\eta)}{2}$$

and $y_0\in I(\Omega, \bar\delta)$ with $\|y_0\|\leqq\varepsilon_0(\eta)/2$ we have, by (A.9) and Proposition 1.3(a), $\|x(T)\|_{D(L)}\leqq\frac{1}{4}\|x_0\|_{D(L)}$, and

$$\begin{aligned}\operatorname{dist}(y(T), \Omega)&\leqq\|y(T)-z(T)\|+\operatorname{dist}(z(T), \Omega)\\ &\leqq c_1\, e^{c_2 T}\|x_0\|_{D(L)}+\alpha^{-1}(V(z(T)))\\ &\leqq\frac{\bar\delta}{2}+\alpha^{-1}(e^{-T}\bar\delta)\leqq\bar\delta\end{aligned}$$

so that we can repeat the previous argument replacing (x_0, y_0) by $(x(T), y(T))$. We get

$$W(x(t), y(t))\leqq 2^{-n}W(x_0, y_0), \qquad nT\leqq t\leqq(n+1)T, \qquad n\in\mathbb{N}$$

so that $W(x(t), y(t))\leqq 2\, e^{-(t/T)\log 2}W(x_0, y_0)$ for all $t\geqq T$. This implies, together with (A.5), that W is a Lyapunov function for system (A.6) and the conclusion follows. □

Proof of Proposition 1.4. Consider again system (A.6). In our cases we have $g_1\equiv 0$. Since the purely imaginary eigenvalues of L are semisimple, there exists $M_4>0$ such that

$$\|e^{tL_2}\|_{L(X^+)}\leqq M_4 \quad \forall t\in\mathbb{R}.$$

Therefore we have

$$\|y(t)\|_{D(L)}\leqq cM_4\|P^+w_0\|_{D(L)},$$

where $c>0$ is such that $\|y\|_{D(L)}\leqq c\|y\|$ for all $y\in X^+$. For each $\alpha\in\,]0,-\sup\{\operatorname{Re}\lambda;\lambda\in\sigma_\omega^-(L)\}[$ and $t\in\,]0,\tau(w_0)[$ we have, by (A.4),

$$\begin{aligned}\|x(t)\|_{D(L)}&\leqq M_1(\alpha)\|P^-w_0\|_{D(L)}\\&\quad+M_1(\alpha)\frac{\Gamma(\theta)}{\alpha^\theta}\sup_{0\leqq s\leqq t}\|g_1(x(s)+y(s))\|_{D_L(\theta,\infty)}\\&\leqq M_1(\alpha)\|P^-w_0\|_{D(L)}+\frac{M_1(\alpha)\Gamma(\theta)}{\alpha^\theta}\\&\quad\cdot K(\rho)(\sup_{0\leqq s\leqq t}\|x(s)\|_{D(L)}+cM_4\|P^+w_0\|_{D(L)}).\end{aligned}$$

Therefore for $\|w_0\|_{D(L)}\leqq\rho$, where ρ is such that $(M_1(\alpha)\Gamma(\theta)/\alpha^\theta)K(\rho)\leqq\frac{1}{2}$, we have

$$\|x(t)\|_{D(L)}\leqq 2M_1(\alpha)\|P^-w_0\|_{D(L)}+CM_4\|P^+w_0\|_{D(L)},\qquad 0\leqq t<\tau,$$

and the statement follows easily. □

Evaluation of the eigenvalues of $\tilde{L}$. For any $\lambda\in\mathbb{C}$, the equation $\lambda(\phi,\psi)=\tilde{L}(\phi,\psi)$ is equivalent to

$$\text{(A.10)}\qquad\begin{aligned}&\text{(i)}\quad(\phi,\psi)\in D(L),\qquad\text{(ii)}\quad\lambda\phi=\psi,\\&\text{(iii)}\quad(\alpha+\gamma\lambda)\phi^{(iv)}-\beta\phi''+(\lambda^2+\delta\lambda)\phi=0.\end{aligned}$$

Each solution of (A.10) is a couple of C^∞ functions whose even order derivatives vanish at $x=0$, $x=1$. Therefore $\phi(x)=\text{const.}\sin(h\pi/l)x(0\leqq x\leqq l)$ for some $h\in\mathbb{N}$, and

$$\text{(A.11)}\qquad\lambda^2+\lambda\left(\delta+\frac{\gamma h^4\pi^4}{l^4}\right)+\frac{h^2\pi^2}{l^2}\left(\frac{\alpha h^2\pi^2}{l^2}+\beta\right)=0$$

so that the eigenvalues of $\tilde{L}$ are given by (2.15).

REFERENCES

[1] V. Arnold, *Chapitres supplémentaires de la théorie des equations differentielles ordinaires*, MIR, Moscow, 1980.

[2] J. M. Ball, *Stability theory for an extensible beam*, J. Differential Equations, 14 (1973), pp. 399-418.

[3] G. Da Prato and A. Lunardi, *Stability, instability and center manifold theorem for fully nonlinear parabolic equations*, submitted to Arch. Rational Mech. Anal., preprint S.N.S. Pisa (1985).

[4] W. E. Fitzgibbon, *Strongly damped quasilinear evolution equations*, J. Math. Anal. Appl., 79 (1981), pp. 536-550.

[5] ———, *Representation and asymptotic behavior of strongly damped evolution equations*, in Nonlinear Phenomena in Mathematical Sciences, Academic Press, New York, 1982, pp. 389-396.

[6] D. Henry, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Math., 840, Springer, Berlin-New York, 1981.

[7] P. Holmes and J. Marsden, *Bifurcation to divergence and flutter in flow-induced oscillations: an infinite dimensional analysis*, Automatica, 14 (1978), pp. 367-384.

[8] ———, *A partial differential equation with infinitely many periodic orbits: chaotic oscillations of a forced beam*, Arch. Rational Mech. Anal., 76 (1981), pp. 135-165.

[9] N. C. Huang and W. Nachbar, *Dynamic snap through of imperfect visco-elastic shallow arches*, Trans. ASME Ser. E, J. Appl. Mech., 35 (1968), pp. 289-297.

[10] A. Lunardi, *Analyticity of the maximal solution of an abstract nonlinear parabolic equation*, Nonlinear Anal., 6 (1982), pp. 503-521.

[11] ———, *Interpolation spaces between domains of elliptic operators and spaces of continuous functions with applications to nonlinear parabolic equations*, Math. Nachr., 121 (1985), pp. 295-318.

[12] J. E. Marsden, *Qualitative methods in bifurcation theory*, Bull. Amer. Math. Soc., 84 (1978), pp. 1125-1148.

[13] J. E. MARSDEN AND M. MCCRACKEN, *The Hopf bifurcation and its applications*, Appl. Math. Sci., 19 (1976).

[14] E. METTLER, *Dynamic buckling*, in Handbook of Engineering Mechanics, S. Flügge, ed., McGraw-Hill, New York, 1962.

[15] V. MOAURO AND P. NEGRINI, *Hopf bifurcation in* $\mathbb{R}^n$: *stability properties of the bifurcating orbits*, Boll. Un. Mat. Ital. B(6) 3 (1984), pp. 623-640.

[16] P. NEGRINI AND L. SALVADORI, *Attractivity and Hopf bifurcation*, Nonlinear Anal., 3 (1979), pp. 87-100.

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

[18] L. SALVADORI, *Sulla stabilità dell'equilibrio nei casi critici*, Ann. Mat. Pura Appl. (4), 69 (1965), pp. 1-33.

[19] E. SINESTRARI, *On the abstract Cauchy problem of parabolic type in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16-66.

[20] F. TAKENS, *Singularities of vector fields*, Publ. Inst. Hautes Etudes Sci. Publ. Math., 43 (1974), pp. 48-100.

[21] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

011

SOME MATHEMATICAL ASPECTS ON A PROBLEM OF THE OPTIMAL DESIGN OF A VIBRATING BEAM*

KJELL HOLMÅKER†

Abstract. An optimal design problem for a vibrating beam is considered. The problem is to maximize (with respect to a design function α) the smallest eigenvalue of a certain eigenvalue problem involving a fourth order differential equation. There is no positive lower bound prescribed for α. Instead it is found that the proper condition on α is that a certain integral should be convergent. It is shown that without this condition the eigenvalue problem need not have any solution. On the other hand it is shown that for functions α satisfying the condition there are eigenvalues and, furthermore, that there is an α_0 such that the smallest eigenvalue is as large as possible. Finally, some necessary conditions satisfied by α_0 are derived in a rigorous way.

Key words. optimal design, eigenvalue optimization, vibrating beam

AMS(MOS) subject classifications. 34B25, 49A40, 49B40, 73K12, 73K40

1. Introduction. Problems of optimal design of beams or plates have been much studied in the engineering literature during the last decades, see e.g. [9], [7], [10] and [11] ([11] is a review paper with many references). In a mathematical formulation it is often the question of maximizing (with respect to a "design variable") the least eigenvalue of a certain eigenvalue problem involving a differential equation of the second or fourth order. Some of the mathematical problems appearing are the questions of existence of an optimal solution and of a rigorous derivation of the necessary conditions (justifying the formal calculations in many papers). Existence proofs when the design variable is bounded away from zero can be found in some papers, e.g. [8] and [3]. In this paper, however, we will have no such restriction. Necessary conditions for optimality for certain problems are derived in [1] and [2]. For eigenvalue problems the question of differentiability of the least eigenvalue with respect to the design variable is important; this is discussed in [6].

In this paper we will take up the problem from [9] and discuss the questions of existence and necessary conditions. The problem in [9] is to find the optimal shape of a simply supported vibrating beam. The equation of motion gives, after separation of variables, the eigenvalue problem

$$\frac{d^2}{dx^2}\left(EI(x)\frac{d^2y}{dx^2}\right)=\omega^2\rho A(x)y, \qquad 0<x<L,$$

$$y(0)=y(L)=0, \qquad EI(x)y''(x)|_{x=0}=EI(x)y''(x)|_{x=L}=0$$

where y is the deflection, L is the length of the beam, E is Young's modulus, ρ is the density, $I(x)$ is the cross sectional moment of inertia, $A(x)$ is the cross sectional area and ω is a natural frequency of the beam. We consider the volume of the beam to be given

$$\int_0^L A(x)\,dx=V,$$

* Received by the editors November 25, 1985; accepted for publication (in revised form) August 20, 1986.

† Department of Mathematics, University of Queensland, St. Lucia, 4067, Queensland, Australia. Permanent address, Department of Mathematics, Chalmers University of Technology, and the University of Göteborg, S-412 96 Göteborg, Sweden.

and we want to choose the cross section in such a way that the least natural frequency becomes as large as possible. We will assume a relation between $I(x)$ and $A(x)$ of the form

$$I(x)=c[A(x)]^p$$

where $p=1,2,3$ are of particular interest (see [7]). Introduce the dimensionless variable $\xi=x/L$ and replace then ξ by x. The equations are then simplified to the following:

$$\frac{d^2}{dx^2}\left(\alpha^p(x)\frac{d^2y}{dx^2}\right)=\lambda\alpha(x)y,\qquad 0<x<1, \tag{1.1a}$$

$$y(0)=y(1)=0, \tag{1.1b}$$

$$\alpha^p(x)y''(x)|_{x=0}=\alpha^p(x)y''(x)|_{x=1}=0, \tag{1.1c}$$

$$\int_0^1 \alpha(x)\,dx=1 \tag{1.2}$$

where

$$\alpha(x)=\frac{A(xL)L}{V},\qquad \lambda=\frac{\omega^2\rho L^{p+3}}{cEV^{p-1}}.$$

The function α is $\geqq 0$, and we allow that $\alpha(x)=0$ at some points, so that (1.1a) might be singular. We shall first (in §§ 2 and 3) consider the eigenvalue problem (1.1) for a fixed α. It turns out that there exist eigenvalues if and only if

$$\int_0^1 x^2(1-x)^2\alpha^{-p}(x)\,dx<\infty, \tag{1.3}$$

provided that α satisfies (1.2) and α^{-p} is locally integrable in (0, 1).

In § 4 we prove that, in the class of α's satisfying (1.2) and (1.3), there exists an α_0 such that the corresponding smallest eigenvalue λ_1 is maximal. Finally, in § 5, we derive some necessary conditions for optimality.

2. The eigenvalue problem. Let $\alpha:[0,1]\to\mathbb{R}$ be a measurable function satisfying

$$\alpha(x)\geqq 0 \text{ a.e.},\qquad \int_0^1 \alpha(x)\,dx<\infty, \tag{2.1}$$

$$\int_0^1 x^2(1-x)^2\alpha^{-p}(x)\,dx<\infty, \tag{2.2}$$

and consider the eigenvalue problem (1.1).

In discussing this problem it is convenient to introduce some classes of functions.

DEFINITION 2.1. (i) Let D be the class of all functions $f:[0,1]\to\mathbb{R}$ such that f' exists on all [0, 1] and is absolutely continuous on [0, 1]. If $f\in D$, it follows that f is absolutely continuous on [0, 1], and f'' exists a.e. and belongs to $L_1(0,1)$.

(ii) Let D_{loc} be the class of all functions $f:[0,1]\to\mathbb{R}$ such that f is absolutely continuous on [0, 1], f' exists on (0, 1) and is absolutely continuous on closed subintervals of (0, 1). If $f\in D_{\text{loc}}$, it follows that $f'\in L_1(0,1)$ and $f''\in L_{1,\text{loc}}(0,1)$.

(iii) Let D_α be the class of all functions $f:[0,1]\to\mathbb{R}$ such that $f\in D_{\text{loc}}$, $f(0)=f(1)=0$, and $\alpha^{p/2}f''\in L_2(0,1)$.

The eigenvalue problem consists of finding λ and $y \neq 0$ such that $y \in D_{\mathrm{loc}}$, $\alpha^p y'' \in D$, and such that (1.1) is satisfied; this means that there exists a $g \in D$ such that

$$\begin{aligned} &\alpha^p(x)y''(x) = g(x) \quad \text{a.e.,} \\ (2.3) \qquad &g''(x) = \lambda\alpha(x)y(x) \quad \text{a.e.,} \\ &y(0) = y(1) = 0, \qquad g(0) = g(1) = 0. \end{aligned}$$

If (in a formal calculation) (1.1) is multiplied by y and integrated, then after partial integrations, taking (1.1b), (1.1c) into account, we find that λ satisfies

$$\lambda = \frac{\int_0^1 \alpha^p(x)[y''(x)]^2\,dx}{\int_0^1 \alpha(x)[y(x)]^2\,dx}.$$

(If y satisfies the conditions above, this can be justified.)

In most treatments (see e.g. [9], [7], [11]) the eigenvalue problem is replaced by the problem of minimizing this Rayleigh quotient. It is therefore of interest to show that the minimum is attained for some y. The exact problem is to find the minimum of

$$(2.4) \qquad R(y) = \frac{\int_0^1 \alpha^p(x)[y''(x)]^2\,dx}{\int_0^1 \alpha(x)[y(x)]^2\,dx}$$

as y varies over $D_\alpha \backslash \{0\}$. One can show by direct methods that this problem has a solution y. Then, as a consequence of the necessary conditions for minimum, one can show that $\alpha^p y'' \in D$ and that (1.1) is satisfied (note in particular that (1.1c) results from the necessary conditions). We shall however proceed differently and transform (1.1) to an eigenvalue problem for an integral operator. The following lemma will then be useful.

LEMMA 2.1. *The differential equation*

$$\begin{cases} y'' = f & \text{in } (0,1), \\ y(0) = y(1) = 0 \end{cases}$$

has the unique solution

$$(2.5) \qquad y(x) = \int_0^1 h(x,t)f(t)\,dt, \qquad x \in [0,1],$$

where

$$(2.6) \qquad h(x,t) = \begin{cases} -(1-x)t & \text{for } 0 \leqq t \leqq x, \\ -x(1-t) & \text{for } x \leqq t \leqq 1, \end{cases}$$

if

(a) $f \in L_1(0,1)$, *in which case* $y \in D$ *and* $\alpha^{-p/2}y \in L_2(0,1)$, *or*

(b) $f = \alpha^{-p/2}v$, *where* $v \in L_2(0,1)$, *in which case* $y \in D_\alpha$.

The proof consists of an easy verification, so we omit the details. Let us only remark that we use that

$$(2.7) \qquad |h(x,t)| \leqq \min(x(1-x), t(1-t)),$$

and that the function $t \mapsto t(1-t)\alpha^{-p/2}(t)$ belongs to $L_2(0,1)$.

THEOREM 2.1. *Let* α *satisfy* (2.1) *and* (2.2). *Then the eigenvalue problem* (1.1) (*or* (2.3)) *has infinitely many eigenvalues,* $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots$. *The Rayleigh quotient* (2.4) *attains its minimum* λ_1 *if and only if* y *is an eigenfunction corresponding to* λ_1. *The least eigenvalue* λ_1 *is simple and the corresponding eigenfunction can be chosen to be positive in* $(0,1)$.

Proof. By a double application of Lemma 2.1 we see that a function $y \neq 0$ such that $y \in D_{\text{loc}}$ and $\alpha^p y'' \in D$ satisfies (1.1) if and only if

$$y(x) = \lambda \int_0^1 h(x,t)\alpha^{-p}(t)\left[\int_0^1 h(t,\tau)\alpha(\tau)y(\tau)\,d\tau\right]dt. \tag{2.8}$$

We have then used that $t \mapsto \alpha^{-p/2}(t)\int_0^1 h(t,\tau)\alpha(\tau)y(\tau)\,d\tau$ belongs to $L_2(0,1)$. This fact also implies that $y \in D_\alpha$.

Let us introduce the integral operator K_α defined by

$$(K_\alpha v)(x) = \int_0^1 h(x,t)\alpha^{1/2}(x)\alpha^{-p/2}(t)v(t)\,dt. \tag{2.9}$$

For the kernel

$$k_\alpha(x,t) = h(x,t)\alpha^{1/2}(x)\alpha^{-p/2}(t)$$

we have

$$\int_0^1\int_0^1 k_\alpha^2(x,t)\,dx\,dt \leqq \int_0^1 \alpha(x)\,dx \cdot \int_0^1 t^2(1-t)^2\alpha^{-p}(t)\,dt < \infty,$$

and therefore K_α is a Hilbert–Schmidt operator from $L_2(0,1)$ to $L_2(0,1)$. This means in particular that K_α is compact. Its adjoint K_α^* is defined by

$$(K_\alpha^* v)(x) = \int_0^1 h(x,t)\alpha^{1/2}(t)\alpha^{-p/2}(x)v(t)\,dt, \qquad v \in L_2(0,1).$$

We see then that (2.8) can be written

$$z = \lambda(K_\alpha K_\alpha^*)z \tag{2.10}$$

where

$$z(x) = \alpha^{1/2}(x)y(x).$$

Thus the eigenvalue problem (1.1) is equivalent to the eigenvalue problem (2.10) for the compact selfadjoint operator $K_\alpha K_\alpha^*$. From the general theory for such operators we know that there are infinitely many eigenvalues $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots$, each with finite multiplicity.

By Lemma 2.1(b) there is a one-to-one correspondence between $L_2(0,1)$ and D_α given by

$$v = \alpha^{p/2}y'', \quad y = \alpha^{-1/2}K_\alpha v, \quad v \in L_2(0,1), \quad y \in D_\alpha. \tag{2.11}$$

Since the eigenvalues of $K_\alpha K_\alpha^*$ are also the eigenvalues of $K_\alpha^* K_\alpha$, we see from (2.4) and (2.11) that

$$\lambda_1 = \min_{\substack{v \in L_2(0,1)\\ v \neq 0}} \frac{\|v\|^2}{\|K_\alpha v\|^2} = \min_{\substack{y \in D_\alpha\\ y \neq 0}} R(y) \tag{2.12}$$

where $\|\cdot\|$ is the norm in $L_2(0,1)$. Thus the Rayleigh quotient $R(y)$ attains its minimum for some $y_0 \in D_\alpha \backslash \{0\}$ and it is also easy to see that this happens if and only if y_0 is an eigenfunction corresponding to λ_1.

It follows from general properties for compact operators with positive kernels that λ_1 is simple and that the corresponding eigenfunction y can be chosen such that $y(x) > 0$ a.e.; see e.g. [4, pp. 287-288]. But then (2.3) shows that $y(x) > 0$ for all $x \in (0,1)$. □

3. A nonexistence result. If condition (2.2) is not satisfied, then the analysis in § 2 breaks down. Let us in this section assume that $\alpha(x) \geqq 0$, $\alpha \in L_1(0,1)$, $\alpha^{-p} \in L_1(a,b)$ for each $[a,b] \subset (0,1)$, and $\int_0^1 x^2(1-x)^2\alpha^{-p}(x)\,dx = \infty$. The problem of minimizing $R(y)$ over $D_\alpha \backslash \{0\}$ still makes sense, but we shall show that it has no solution in this case.

Let, for $0 < \delta < \frac{1}{2}$, χ_δ be the characteristic function of the interval $[\delta, 1-\delta]$. Define

$$v_\delta(x) = -x(1-x)\alpha^{-p/2}(x)\chi_\delta(x),$$

$$y_\delta(x) = \int_0^1 h(x,t)\alpha^{-p/2}(t)v_\delta(t)\,dt,$$

$$A_\delta = \int_\delta^{1-\delta} x^2(1-x)^2\alpha^{-p}(x)\,dx = \int_0^1 v_\delta^2(x)\,dx.$$

It follows from Lemma 2.1(b) that $y_\delta'' = \alpha^{-p/2}v_\delta$ and $y_\delta \in D_\alpha \backslash \{0\}$. Let $x_0 \in (0,1)$ be such that $y_\delta(x_0) = \max_{x\in[0,1]} y_\delta(x)$. By using (2.6) and the fact that $y_\delta'(x_0) = 0$, we easily find that $y_\delta(x_0) \geqq \frac{1}{2}A_\delta$. Since $y_\delta'' \leqq 0$, y_δ is concave and

$$y_\delta(x) \geqq y_\delta(x_0)\frac{x}{x_0} \quad \text{for } 0 \leqq x \leqq x_0,$$

$$y_\delta(x) \geqq y_\delta(x_0)\frac{1-x}{1-x_0} \quad \text{for } x_0 \leqq x \leqq 1.$$

From this we find that $\int_0^1 \alpha(x)y_\delta^2(x)\,dx \geqq ky_\delta^2(x_0)$ for a certain constant $k > 0$ (independent of x_0). Therefore

$$R(y_\delta) = \frac{\int_0^1 \alpha^p(x)y_\delta''^2(x)\,dx}{\int_0^1 \alpha(x)y_\delta^2(x)\,dx} = \frac{A_\delta}{\int_0^1 \alpha(x)y_\delta^2(x)\,dx} \leqq \frac{4}{kA_\delta}.$$

But $A_\delta \to \infty$ as $\delta \to 0^+$. Thus

$$\inf_{y \in D_\alpha \backslash \{0\}} R(y) = 0.$$

But $R(y)$ can obviously not attain the value 0.

4. Maximizing the least eigenvalue. Now we want to let α vary in the class of functions satisfying (2.1), (2.2) and (1.2), and we want to maximize the least eigenvalue λ_1 of (1.1). We consider however also a slightly larger class.

Let, for $p \geqq 1$, $\mathscr{A}_p$ be the class of all measurable functions $\alpha : [0,1] \to \mathbb{R}$ such that

$$\alpha(x) \geqq 0 \text{ a.e.}, \qquad 0 < \int_0^1 \alpha(x)\,dx \leqq 1, \tag{4.1}$$

$$A_p(\alpha) = \int_0^1 x^2(1-x)^2\alpha^{-p}(x)\,dx < \infty. \tag{4.2}$$

Let $\mathscr{A}_p'$ be the class of those α in $\mathscr{A}_p$ for which $\int_0^1 \alpha(x)\,dx = 1$.

For each $\alpha \in \mathscr{A}_p$ there are eigenvalues of the eigenvalue problem (1.1) (see § 2), and in particular there is a smallest one given by (2.12); let us denote it by $\lambda(\alpha)$.

The main result of this section is the following theorem. We prove it now for $p > 1$, whereas the result for $p = 1$ will be obtained in § 5.

THEOREM 4.1. *There is an* $\alpha_0 \in \mathscr{A}_p'$ *such that*

$$\lambda(\alpha_0) = \sup_{\alpha \in \mathscr{A}_p} \lambda(\alpha).$$

In the proof the following lemma will be used.

LEMMA 4.1. *Let $(x, u)\mapsto h(x, u)$ be a function defined for $x\in[0, 1]$ and $u\in I$, where $I\subseteq\mathbb{R}$ is a closed interval, with values in $[0,\infty]$. Assume that $h(\cdot, u)$ is measurable for each $u\in I$, and that $h(x, \cdot)$ is continuous and convex for each $x\in[0, 1]$. If $u_n:[0, 1]\to I$, $n=1, 2, \cdots$, is a sequence of functions in $L_r(0, 1)$ such that $u_n\to u$ weakly in $L_r(0, 1)$ as $n\to\infty$ for some r, $1\leqq r<\infty$, then $u(x)\in I$ a.e., and*

$$\int_0^1 h(x, u(x))\,dx\leqq\liminf_{n\to\infty}\int_0^1 h(x, u_n(x))\,dx. \tag{4.3}$$

A similar result can be found in [5, p. 7], so we only remark that the proof of Lemma 4.1 is based on Mazur's Theorem and Fatou's Lemma.

In the proof of Theorem 4.1 we need an estimate of $\lambda(\alpha)$ in terms of $A_p(\alpha)$. This estimate is first proved in a separate lemma.

LEMMA 4.2. *There is a constant $C>0$ such that*

$$\frac{1}{A_p(\alpha)}\leqq\lambda(\alpha)\leqq\frac{C}{[A_p(\alpha)]^{1-1/p}}$$

for all $\alpha\in\mathcal{A}_p$.

Proof. If $v_0\in L_2(0, 1)\backslash\{0\}$ achieves minimum in (2.12), we get from the Cauchy-Schwarz inequality, using (2.12), (2.9), (4.2) and (4.1), the following:

$$\lambda(\alpha)=\frac{\|v_0\|^2}{\|K_\alpha v_0\|^2}\geqq\frac{\|v_0\|^2}{\int_0^1\alpha(x)A_p(\alpha)\|v_0\|^2\,dx}\geqq\frac{1}{A_p(\alpha)}.$$

To get an estimate of $\lambda(\alpha)$ from above, choose

$$v(x)=-x(1-x)\alpha^{-p/2}(x)$$

and define

$$y(x)=\int_0^1 h(x, t)\alpha^{-p/2}(t)v(t)\,dt.$$

Since $v\in L_2(0, 1)$, Lemma 2.1(b) shows that $y\in D_\alpha$ and $y''=\alpha^{-p/2}v\leqq 0$. By the same arguments as in § 3 we obtain

$$y(x_0)=\max_{x\in[0,1]}y(x)\geqq\tfrac{1}{2}A_p(\alpha), \tag{4.4}$$

and

$$\begin{aligned}\|K_\alpha v\|^2&=\int_0^1\alpha(x)y^2(x)\,dx\\ &\geqq y^2(x_0)\left[x_0^{-2}\int_0^{x_0}x^2\alpha(x)\,dx+(1-x_0)^{-2}\int_{x_0}^1(1-x)^2\alpha(x)\,dx\right].\end{aligned} \tag{4.5}$$

Now, by Hölder's inequality

$$\begin{aligned}\int_0^{x_0}x^{1+1/p}(1-x)^{1/p}\,dx&=\int_0^{x_0}[x(1-x)\alpha^{-p/2}(x)]^{1/p}x\alpha^{1/2}(x)\,dx\\ &\leqq x_0^{(p-1)/(2p)}\left[\int_0^{x_0}x^2(1-x)^2\alpha^{-p}(x)\,dx\right]^{1/(2p)}\\ &\quad\cdot\left[\int_0^{x_0}x^2\alpha(x)\,dx\right]^{1/2},\end{aligned}$$

so that

$$x_0^{-2}\int_0^{x_0}x^2\alpha(x)\,dx\geqq x_0^{-(3-1/p)}A_p^{-1/p}(\alpha)\left[\int_0^{x_0}x^{1+1/p}(1-x)^{1/p}\,dx\right]^2,$$

and in the same way

$$(1-x_0)^{-2}\int_{x_0}^{1}(1-x)^2\alpha(x)\,dx \geqq (1-x_0)^{-(3-1/p)}A_p^{-1/p}(\alpha)\left[\int_{x_0}^{1}x^{1/p}(1-x)^{1+1/p}\,dx\right]^2.$$

From (4.4) and (4.5) we then obtain

$$\|K_\alpha v\|^2 \geqq kA_p^{2-1/p}(\alpha)$$

for some constant $k>0$ (which is independent of x_0). For our choice of v, $\|v\|^2=A_p(\alpha)$, so that

$$\lambda(\alpha)\leqq\frac{\|v\|^2}{\|K_\alpha v\|^2}\leqq\frac{1}{kA_p^{1-1/p}(\alpha)}. \qquad \square$$

Proof of Theorem 4.1 *for* $p>1$. Let $p>1$ and

$$\lambda_0=\sup_{\alpha\in\mathscr{A}_p}\lambda(\alpha).$$

From Lemma 4.2 we deduce that $A_p^{-1/p}(\alpha)\leqq C$ and $\lambda(\alpha)\leqq C^p$ for all $\alpha\in\mathscr{A}_p$, and therefore $0<\lambda_0<\infty$. Let $\alpha_n\in\mathscr{A}_p$ be a maximizing sequence, i.e., $\lambda(\alpha_n)\to\lambda_0$ as $n\to\infty$. We may assume that $\lambda(\alpha_n)\geqq\lambda_0/2$ for all n. Then Lemma 4.2 gives

$$A_p(\alpha_n)\leqq\left[\frac{C}{\lambda(\alpha_n)}\right]^{p/(p-1)}\leqq\left(\frac{2C}{\lambda_0}\right)^{p/(p-1)}=C_1\quad\text{for all } n. \tag{4.6}$$

Let $\beta_n=\alpha_n^{1/2}$ and $g_n(x)=x(1-x)\beta_n^{-p}(x)$. Then (4.1), (4.2) and (4.6) show that β_n and g_n are uniformly bounded in $L_2(0,1)$. Therefore we can find a subsequence (we may assume that it is the original sequence) such that $\beta_n\to\bar\beta$ and $g_n\to g$ weakly in $L_2(0,1)$ as $n\to\infty$ for some $\bar\beta\geqq0$ and $g\geqq0$ in $L_2(0,1)$. Define

$$\beta(x)=\left[\frac{x(1-x)}{g(x)}\right]^{1/p}.$$

By Lemma 4.1

$$\int_0^1\beta(x)\phi(x)\,dx\leqq\liminf_{n\to\infty}\int_0^1[x(1-x)]^{1/p}g_n^{-1/p}(x)\phi(x)\,dx=\int_0^1\bar\beta(x)\phi(x)\,dx$$

for all $\phi\in L_2(0,1)$ such that $\phi\geqq0$. Thus

$$0\leqq\beta(x)\leqq\bar\beta(x)\text{ a.e.} \tag{4.7}$$

Lemma 4.1 also gives

$$\int_0^1\bar\beta^2(x)\,dx\leqq\liminf_{n\to\infty}\int_0^1\beta_n^2(x)\,dx\leqq1. \tag{4.8}$$

If we put $\alpha_0=\beta^2$, (4.7), (4.8) and the fact that $A_p(\alpha_0)=\int_0^1g^2(x)\,dx<\infty$ show that $\alpha_0\in\mathscr{A}_p$.

Let $v_0\in L_2(0,1)\backslash\{0\}$ be such that

$$\lambda(\alpha_0)=\frac{\|v_0\|^2}{\|K_{\alpha_0}v_0\|^2}. \tag{4.9}$$

We want to show that $\lambda(\alpha_n)\to\lambda(\alpha_0)$ as $n\to\infty$. We have that

$$\lambda(\alpha_n)\leqq\frac{\|v_0\|^2}{\|K_{\alpha_n}v_0\|^2}, \tag{4.10}$$

and

$$(4.11)\qquad \begin{aligned}(K_{\alpha_n}v_0)(x)&=\alpha_n^{1/2}(x)\int_0^1 h(x,t)\alpha_n^{-p/2}(t)v_0(t)\,dt\\&=\beta_n(x)\int_0^1\frac{h(x,t)}{t(1-t)}g_n(t)v_0(t)\,dt=\beta_n(x)y_n(x)\end{aligned}$$

where $y_n(x)$ is defined as the integral. If we also define

$$y(x)=\int_0^1\frac{h(x,t)}{t(1-t)}g(t)v_0(t)\,dt=\int_0^1 h(x,t)\alpha_0^{-p/2}(t)v_0(t)\,dt,$$

then, since $g_n\to g$ weakly, $y_n(x)\to y(x)$ as $n\to\infty$ for each $x\in[0,1]$.

Let ε, $0<\varepsilon<1$, be arbitrary, and choose $\delta>0$ such that

$$(4.12)\qquad \int_\delta^{1-\delta}\alpha_0(x)y^2(x)\,dx>(1-\varepsilon)\int_0^1\alpha_0(x)y^2(x)\,dx=(1-\varepsilon)\|K_{\alpha_0}v_0\|^2.$$

Now we have

$$|y_n(x)|\leqq\int_0^1 g_n(t)|v_0(t)|\,dt\leqq C_1^{1/2}\|v_0\|\quad\text{for all } x\in[0,1],$$

and (using (2.6))

$$|y_n'(x)|\leqq\frac{1}{\delta}\int_0^1 t(1-t)\alpha_n^{-p/2}(t)|v_0(t)|\,dt\leqq\frac{1}{\delta}C_1^{1/2}\|v_0\|\quad\text{for all } x\in[\delta,1-\delta],$$

so that the sequence $\{y_n\}_1^\infty$ is uniformly bounded and equicontinuous on $[\delta,1-\delta]$. The Ascoli-Arzelà Theorem and the fact that $y_n(x)\to y(x)$ pointwise show that $y_n\to y$ uniformly on $[\delta,1-\delta]$ as $n\to\infty$. Using (4.11), Lemma 4.1, (4.7) and (4.12), we then get

$$\begin{aligned}\liminf_{n\to\infty}\|K_{\alpha_n}v_0\|^2&\geqq\liminf_{n\to\infty}\int_\delta^{1-\delta}\beta_n^2(x)y_n^2(x)\,dx\\&=\liminf_{n\to\infty}\int_\delta^{1-\delta}\beta_n^2(x)y^2(x)\,dx\\&\geqq\int_\delta^{1-\delta}\bar\beta^2(x)y^2(x)\,dx\geqq\int_\delta^{1-\delta}\beta^2(x)y^2(x)\,dx\\&=\int_\delta^{1-\delta}\alpha_0(x)y^2(x)\,dx>(1-\varepsilon)\|K_{\alpha_0}v_0\|^2.\end{aligned}$$

From (4.10) and (4.9) we then get

$$\lambda_0=\lim_{n\to\infty}\lambda(\alpha_n)\leqq\limsup_{n\to\infty}\frac{\|v_0\|^2}{\|K_{\alpha_n}v_0\|^2}\leqq\frac{1}{1-\varepsilon}\frac{\|v_0\|^2}{\|K_{\alpha_0}v_0\|^2}=\frac{1}{1-\varepsilon}\lambda(\alpha_0).$$

But $\varepsilon>0$ is arbitrary, so therefore $\lambda(\alpha_0)=\lambda_0=\sup_{\alpha\in\mathscr{A}_p}\lambda(\alpha)$.

From the fact that $\lambda(\mu\alpha_0)=\mu^{p-1}\lambda(\alpha_0)$ for $\mu>1$, we find that $\int_0^1\alpha_0(x)\,dx=1$, so that $\alpha_0\in\mathscr{A}_p'$.

Let us also remark that in fact $\bar\beta=\beta$ and $\beta_n\to\beta$ strongly in $L_2(0,1)$. We have

$$\begin{aligned}1=\int_0^1\alpha_0(x)\,dx&=\int_0^1\beta^2(x)\,dx\leqq\int_0^1\bar\beta^2(x)\,dx\\&\leqq\liminf_{n\to\infty}\int_0^1\beta_n^2(x)\,dx\leqq\limsup_{n\to\infty}\int_0^1\beta_n^2(x)\,dx\leqq1,\end{aligned}$$

so that $\bar\beta(x)=\beta(x)$ a.e. and $\|\beta_n\|\to\|\beta\|$ as $n\to\infty$. This together with the weak convergence of β_n towards β implies strong convergence. □

5. Necessary conditions. Assume that $\alpha_0 \in \mathscr{A}'_p$, $p \geqq 1$ and $v_0 \in L_2(0, 1) \backslash \{0\}$ are such that

$$\lambda_0 = \lambda(\alpha_0) = \max_{\alpha \in \mathscr{A}_p} \lambda(\alpha) = \frac{\|v_0\|^2}{\|K_{\alpha_0} v_0\|^2} = \min_{\substack{v \in L_2(0,1) \\ v \neq 0}} \frac{\|v\|^2}{\|K_{\alpha_0} v\|^2}.$$

Let $\delta > 0$ be arbitrary. Put

$$I_\delta = \{x \in [0, 1]\colon \alpha_0(x) \geqq \delta\}$$

and let $\phi : [0, 1] \to \mathbb{R}$ be an arbitrary measurable function such that $|\phi(x)| \leqq 1$ for all x, $\phi(x) = 0$ for $x \in [0, 1] \backslash I_\delta$, $\int_0^1 \phi(x)\, dx = 0$, and define

$$\alpha_\varepsilon(x) = \alpha_0(x) + \varepsilon \phi(x) \quad \text{for } |\varepsilon| \leqq \frac{\delta}{2}.$$

It follows that $\alpha_\varepsilon \in \mathscr{A}'_p$. We consider the perturbed operator $H_\varepsilon = K^*_{\alpha_\varepsilon} K_{\alpha_\varepsilon}$ and find that its kernel $h_\varepsilon(x, t)$ can be expanded in a convergent power series

$$h_\varepsilon(x, t) = h_0(x, t) + \varepsilon h^{(1)}(x, t) + \cdots + \varepsilon^n h^{(n)}(x, t) + \cdots \text{ for } |\varepsilon| \leqq \frac{\delta}{2}$$

where

$$|h^{(n)}(x, t)| \leqq x(1-x)\alpha_0^{-p/2}(x) \cdot t(1-t)\alpha_0^{-p/2}(t) \cdot K^n$$

for some constant $K > 0$. For $h^{(1)}$ we get

$$(5.1)\qquad \begin{aligned} h^{(1)}(x, t) &= \alpha_0^{-p/2}(x)\alpha_0^{-p/2}(t) \\ &\cdot \left\{ \int_0^1 h(x, s)h(t, s)\phi(s)\, ds - \frac{p}{2}\left[\frac{\phi(x)}{\alpha_0(x)} + \frac{\phi(t)}{\alpha_0(t)}\right] \int_0^1 \alpha_0(s) h(x, s) h(t, s)\, ds \right\}. \end{aligned}$$

Let $h^{(n)}(x, t)$ be the kernel of a selfadjoint Hilbert-Schmidt operator $H^{(n)}$ from $L_2(0, 1)$ to $L_2(0, 1)$. We can now apply a theorem of Rellich (see [13, § 2] or [4, pp. 171-172]) which says that the eigenvalue $\mu_\varepsilon = \lambda^{-1}(\alpha_\varepsilon)$ of H_ε and a suitable corresponding eigenfunction v_ε depend analytically on ε (note that $\mu_0 = \lambda_0^{-1}$ is a simple eigenvalue according to Theorem 2.1). Write

$$H_\varepsilon = H_0 + \varepsilon H^{(1)} + \cdots, \quad v_\varepsilon = v_0 + \varepsilon v^{(1)} + \cdots \quad \text{and} \quad \mu_\varepsilon = \mu_0 + \varepsilon \mu^{(1)} + \cdots$$

where we assume that $\|v_0\| = 1$. By identifying the coefficients of ε in the equation $H_\varepsilon v_\varepsilon = \mu_\varepsilon v_\varepsilon$ we get, after scalar multiplication by v_0

$$\mu^{(1)} = (H^{(1)} v_0, v_0).$$

From the extremal property of λ_0 we get $0 = (d/d\varepsilon)\mu_\varepsilon|_{\varepsilon=0} = \mu^{(1)}$, so that $(H^{(1)} v_0, v_0) = 0$. From (5.1) we obtain after some computations

$$(5.2)\qquad 0 = (H^{(1)} v_0, v_0) = \int_0^1 \frac{\phi(x)}{\alpha_0(x)} [(K_{\alpha_0} v_0)^2(x) - p\mu_0 v_0^2(x)]\, dx.$$

Now $y_0 = \alpha_0^{-1/2} K_{\alpha_0} v_0$ is an eigenfunction of (1.1) with $\alpha = \alpha_0$ corresponding to the eigenvalue $\lambda_0 = \mu_0^{-1}$. Since (5.2) holds for all ϕ of the form described above, we get

$$(5.3)\qquad y_0^2(x) - p\mu_0 \alpha_0^{-1}(x) v_0^2(x) = -a$$

a.e. on I_δ, where a is a constant. Since $\alpha_0(x) > 0$ a.e. on $[0, 1]$, and $\delta > 0$ is arbitrary, (5.3) holds a.e. on $[0, 1]$. If (5.3) is multiplied by α_0 and integrated, we get

$$a = (p-1)\|K_{\alpha_0} v_0\|^2 \geqq 0.$$

Since $y_0'' = \alpha_0^{-p/2} v_0$, we find that y_0 must satisfy

$$p\alpha_0^{p-1}(x) y_0''^2(x) - \lambda_0 y_0^2(x) = k \quad \text{a.e.,} \tag{5.4}$$

where

$$k = (p-1)\lambda_0 \int_0^1 \alpha_0(x) y_0^2(x)\, dx. \tag{5.5}$$

Equations (5.4) and (5.5) constitute the necessary conditions for optimality for our problem. They can be utilized to devise a numerical algorithm as in [7] and [9]. From them we can also obtain more information on the behavior of α_0 and y_0. We consider now the case $p > 1$, where we know that an optimal α_0 exists. From (2.3) we get

$$\alpha_0^p(x) y_0''(x) = g(x) \quad \text{a.e.,} \tag{5.6}$$

$$g''(x) = \lambda_0 \alpha_0(x) y_0(x) \quad \text{a.e.,} \tag{5.7}$$

$$g(0) = g(1) = 0, \tag{5.8}$$

for some $g \in D$. From Theorem 2.1 we know that we may assume that $y_0(x) > 0$ in $(0, 1)$. Then (5.7) and (5.8) imply that $g(x) < 0$ in $(0, 1)$. From (5.4)-(5.6) we get

$$\alpha_0(x) = \left[\frac{pg^2(x)}{k + \lambda_0 y_0^2(x)}\right]^{1/(p+1)} \quad \text{where } k > 0. \tag{5.9}$$

We can redefine α_0 on a set of measure zero so that (5.9) holds for all x. Then we see that α_0 is continuous on $[0, 1]$ and positive in $(0, 1)$. Combining (5.7) and (5.9) we find that g satisfies the equation

$$g''(x) = \psi(x)[g^2(x)]^{1/(p+1)} \tag{5.10}$$

for some function ψ continuous on $[0, 1]$. We want to prove that $g'(0) \neq 0$. Assume $g'(0) = 0$. Then (5.10) implies that $|g(x)| \leqq C \cdot x^{2(p+1)/(p-1)}$ for some constant $C > 0$. From (5.6) and (5.9) we then get $y_0''(x) \leqq -C_1 x^{-2}$, and hence $y_0(x) \leqq C_1 \ln x + C_2$ in a neighborhood of 0, for certain constants $C_1 > 0$ and C_2, and that is impossible. Thus $g'(0) \neq 0$. Then $g(x) = O(x)$ as $x \to 0$, and from (5.9) and (5.6) we get

$$\alpha_0(x) = O(x^{2/(p+1)}) \quad \text{as } x \to 0,$$

and $y_0''(x) = O(x^{-(p-1)/(p+1)})$, so that $y_0'' \in L_1(0, \frac{1}{2})$ and

$$y_0(x) = y_0'(0)x + O(x^{(p+3)/(p+1)}) \quad \text{as } x \to 0.$$

The behavior at $x = 1$ is similar.

We can now settle the question of existence in the case $p = 1$, which was left open in the proof of Theorem 4.1. If there is a solution for $p = 1$, then λ_0, α_0 and y_0 must satisfy (5.6)-(5.8) and

$$y_0''^2(x) = \lambda_0 y_0^2(x).$$

As we noted above, (5.6) implies that y_0 and y_0'' are of opposite sign, so that

$$y_0''(x) = -\sqrt{\lambda_0}\, y_0(x), \qquad y_0(0) = y_0(1) = 0, \tag{5.11}$$

which has the solution $\lambda_0 = \pi^4$, $y_0(x) = c_1 \sin \pi x$, $c_1 \neq 0$. For g we obtain the same equation, so that $g(x) = c_2 \sin \pi x$, $c_2 \neq 0$, and $\alpha_0(x) = g(x)/y_0''(x) = -c_2/(\pi^2 c_1) = c_3$. The condition $\int_0^1 \alpha_0(x)\, dx = 1$ gives $c_3 = 1$.

By an argument partly inspired by [12] we can show that the necessary conditions, which λ_0, α_0 and y_0 satisfy, also are sufficient, i.e., that $\alpha_0(x) \equiv 1$ really is optimal. To that end, consider the Rayleigh quotient

$$R(\alpha, y)=\frac{\int_0^1 \alpha(x)[y''(x)]^2\,dx}{\int_0^1 \alpha(x)[y(x)]^2\,dx}$$

where $\alpha \in \mathscr{A}_1'$, $y \in D_\alpha \backslash\{0\}$. Let $y_\alpha \in D_\alpha \backslash\{0\}$ be such that

$$\lambda(\alpha)=R(\alpha, y_\alpha)=\min_{y \in D_\alpha \backslash\{0\}} R(\alpha, y).$$

Since y_0'' is bounded, $y_0 \in D_\alpha$ for all α (see Definition 2.1), so that

$$R(\alpha, y_\alpha) \leqq R(\alpha, y_0)=\lambda_0;$$

in the last step we used (5.11). Since $\lambda_0=\lambda(\alpha_0)$ by construction, we see that $\lambda(\alpha) \leqq \lambda(\alpha_0)$ for all $\alpha \in \mathscr{A}_1'$, which means that α_0 is optimal.

REFERENCES

[1] D. C. BARNES, *Extremal problems for eigenvalues with applications to buckling, vibration and sloshing*, this Journal, 16 (1985), pp. 341-357.

[2] E. R. BARNES, *Some max-min problems arising in optimal design studies*, in Control Theory of Systems Governed by Partial Differential Equations, A. K. Aziz, J. W. Wingate and M. J. Balas, eds., Academic Press, New York, 1977, pp. 177-208.

[3] M. P. BENDSØE, *On obtaining a solution to optimization problems for solid, elastic plates by restriction of the design space*, J. Structural Mech., 11 (1983-84), pp. 501-521.

[4] J. A. COCHRAN, *The Analysis of Linear Integral Equations*, McGraw-Hill, New York, 1972.

[5] B. DACOROGNA, *Weak continuity and weak lower semicontinuity of nonlinear functionals*, Lecture Notes in Mathematics, 922, Springer-Verlag, Berlin-Heidelberg-New York, 1982.

[6] E. J. HAUG AND B. ROUSSELET, *Design sensitivity analysis in structural mechanics.* II. *Eigenvalue variations*, J. Structural Mech., 8 (1980), pp. 161-181.

[7] B. L. KARIHALOO AND F. J. NIORDSON, *Optimum design of vibrating cantilevers*, J. Optim. Theory Appl., 11 (1973), pp. 638-654.

[8] V. G. LITVINOV, *Optimal control of the natural frequency of a plate of variable thickness*, USSR Comput. Math. and Math. Phys., 19 (1980), pp. 70-86.

[9] F. J. NIORDSON, *On the optimal design of a vibrating beam*, Quart. Appl. Math., 23 (1965), pp. 47-53.

[10] N. OLHOFF, *Optimization of vibrating beams with respect to higher order natural frequencies*, J. Structural Mech., 4 (1976), pp. 87-122.

[11] ———, *Optimal design with respect to structural eigenvalues*, in Theoretical and Applied Mechanics, (Proc. 15th Internat. Congr., Univ. Toronto, Toronto, Ontario, Canada, 1980), North-Holland, Amsterdam, 1980, pp. 133-149.

[12] W. PRAGER AND J. E. TAYLOR, *Problems of optimal structural design*, J. Appl. Mech., 35 (1968), pp. 102-106.

[13] F. RELLICH, *Störungstheorie der Spektralzerlegung*, Math. Ann., 113 (1937), pp. 600-619.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

012

A NOTE ON THE IDENTIFIABILITY OF DISTRIBUTED PARAMETERS IN ELLIPTIC EQUATIONS*

CARMEN CHICONE† AND JÜRGEN GERLACH†

Abstract. For u and f given smooth functions on a bounded domain Ω we consider solutions of the PDE $-\text{div}\,(a\nabla u)=f$ for the parameter a. This problem arises in the identification of the flow of groundwater. We say a is identifiable if, for given u and f, a is unique. Our main result shows that a is identifiable on the points in Ω which are the closure of the interior of the set of points which stay in Ω for all positive time (or negative time) under the flow of the gradient field ∇u. We also show a is identifiable on Ω if the set of critical points of u has nonempty interior and the co-normal derivative of u is specified on $\partial\Omega$.

Key words. parameter identification, uniqueness, inverse problems

AMS(MOS) subject classifications. 35R30, 76S05

1. Introduction. In this paper we investigate the following situation: A bounded region $\Omega\subset R^n$ with smooth boundary $\partial\Omega$ and a function $u\in C^2(\bar\Omega)$ are given. We consider solutions $a(x)$ of the equation

$$\text{div}\,(a\nabla u)=0, \tag{1}$$

where $a\in C^1(\bar\Omega)$. Obviously, $a(x)\equiv 0$ solves (1). The main question which we are going to address in this note is the following: What are the conditions on the function $u(x)$ so that $a(x)\equiv 0$ is the only solution of (1)?

Background. This problem arises in the identification of the flow of groundwater. If we denote the pressure head of an aquifer by $u(x)$, its transmissivity by $a(x)$, and if we denote external sources such as wells by $f(x)$, we obtain the equation

$$-\text{div}\,(a\nabla u)=f \tag{2}$$

in the case of a steady flow. The actual flow rates q are then calculated from $q=a\nabla u$. Suppose that f and u are known everywhere, and suppose that (2) is satisfied for two transmissivity functions $a_1(x)$ and $a_2(x)$; then their difference $a(x):=a_1(x)-a_2(x)$ satisfies (1). If $a(x)\equiv 0$ is the only possible solution of (1), then the data u and f can be explained by exactly one transmissivity function. In this case we call the parameter $a(x)$ identifiable from the data u. In § 4 we will consider test conditions under which $a(x)$ is identifiable.

A related question can be asked from the point of view of dynamical systems: Suppose we are given a gradient flow ∇u; when is it possible to multiply the gradient field by a nonzero function $a(x)$ so that the resulting vector field $v=a\nabla u$ is divergence free?

Uniqueness of $a(x)$ plays an important role in numerical schemes for parameter identifications, e.g., this assumption is needed for the methods suggested by Kravaris and Seinfeld [7], and Gerlach and Guenther [4]. Kitamura and Nakagiri [6] study the identifiability for a one-dimensional time dependent problem, and Kunisch and White [9] investigate the identifiability of a discretized version of (2) for several finite element approximations for one space dimension. In higher dimensions Falk [3] considers the case where the flow satisfies $\nabla u\cdot\nu\geqq\sigma>0$ for some fixed vector $\nu\in R^n$, and some constant σ, and uses inflow data for stability estimates. Similarly, Richter [11] considers

* Received by the editors May 7, 1986; accepted for publication (in revised form) October 29, 1986.

† Department of Mathematics, University of Missouri, Columbia, Missouri 65211.

the special cases where $|\nabla u|>0$, $\Delta u>0$, and $\max\{|\nabla u|, \Delta u\}>0$, and uses inflow data on the boundary for estimates. Alessandrini considers the uniqueness problem in two spatial variables with $u\equiv 0$ on $\partial\Omega$ for $f=0$ [2], and $f=\delta(x-x_0)$ [1]. In this paper we obtain uniqueness of the parameter on subregions of Ω which are solely determined by the observation u. In particular, uniqueness is not linked to flow data on the boundary.

2. Notation and preliminaries. We first note two special cases of the identification problem:

(1) If u is constant in a subregion of Ω, the gradient will vanish identically there, and $a(x)$ can be chosen arbitrarily.

(2) If u is harmonic on Ω, (1) will be satisfied for any constant function $a(x)$.

In addition we observe that (1) can be viewed as a first-order partial differential equation for $a(x)$, and we can rewrite it in the form

$$\nabla a\cdot\nabla u+a\Delta u=0. \tag{3}$$

The gradient field of u defines the characteristic curves for this equation, and $a(x)$ could be found by integration along the characteristics, provided Cauchy data for a are known along a noncharacteristic manifold. The linear character of (3) implies that $a(x)$ does not change sign along characteristic curves, and in particular, if $a(P)=0$ at some point P, it vanishes identically along the characteristic through that point. We also recall that the function u itself is strictly increasing along flow lines of the gradient field.

In what follows Ω is an open, bounded and connected subset of R^n with smooth boundary $\partial\Omega$. The function $u(x)$ belongs to $C^2(\bar\Omega)$, i.e., there exists an open set D such that $\bar\Omega\subset D$ and $u\in C^2(D)$. The set of all critical points of u is denoted by C.

Next we consider the gradient field of u. Its trajectories are denoted by $\phi_t(P)$, i.e., if $x(t)$ is a solution of

$$\frac{dx}{dt}=\nabla u,\qquad x(0)=P, \tag{4}$$

we write $\phi_t(P):=x(t)$. It is understood that we take the maximal t-intervals so that $\phi_t(P)\in D$.

Now we define subsets of $\bar\Omega$ by

$$\Omega_+:=\{P\in\bar\Omega\colon \phi_t(P)\in D-\bar\Omega \text{ for some } t>0\},$$

$$\Omega_-:=\{P\in\bar\Omega\colon \phi_t(P)\in D-\bar\Omega \text{ for some } t<0\},$$

$$\Omega_0:=\Omega_+\cap\Omega_-.$$

Any one of these sets may be empty, or may be equal to $\bar\Omega$, and none of them contains critical points of u. The complements of Ω_+ and Ω_- are denoted in the following way:

$$\Omega_f:=\bar\Omega-\Omega_+\quad\text{and}\quad\Omega_b:=\bar\Omega-\Omega_-.$$

The set Ω_f contains all points P of $\bar\Omega$ for which the trajectories $\phi_t(P)$ remain in $\bar\Omega$ for all $t>0$. In particular, the singular points of u are contained in the intersection of Ω_f and Ω_b.

The structure of the sets just defined is important to our considerations. We need the following easy proposition, which we state without proof.

PROPOSITION 1. *Ω_+, Ω_-, and Ω_0 are relatively open in $\bar\Omega$. Thus, also Ω_f, Ω_b, and Ω_1 are closed subsets of R^n.*

DEFINITION. The parameter $a(x) \in C^1(\bar{\Omega})$ is called identifiable at a point $P \in \bar{\Omega}$ from the data $u(x)$, if for any solution $a(x)$ of the equation $\operatorname{div}(a\nabla u) = 0$ on $\bar{\Omega}$ necessarily $a(P) = 0$. We call the parameter identifiable on $S \subset \bar{\Omega}$, if it is identifiable for all $P \in S$.

PROPOSITION 2. *$a(x)$ is not identifiable in Ω_0.*

Proof. Let $P \in \Omega_0$. By Proposition 1 and the fact that P is a nonsingular point of u, there is an open $(n-1)$-dimensional disk $S \subset \Omega_0$ of radius $r > 0$ contained in the level set M of u through P.

Let $f\colon S \to R$ be a smooth bump function with $0 \leqq f(s) \leqq 1$, and $f \equiv 0$ on $S - S_0$, where S_0 is the disk in M of radius $r/2$. We define a function $a(x)$ in $\bar{\Omega}$ as follows: If for some t we have $Q = \phi_t(P_0)$, where $P_0 \in S$, we define $a(Q)$ as $a(t)$, where $a(\tau)$ is the solution of (3) with initial value $a(0) = f(P_0)$. Else we set $a(Q) = 0$. It follows that in $\bar{\Omega}$, a satisfies $\operatorname{div}(a\nabla u) = 0$, and $a(P) = 1$. Thus, a is not identifiable at P, and P was arbitrary in Ω_0. □

3. The main result. In the previous part we have seen that $a(x)$ is not identifiable on Ω_0, and in the introduction it was pointed out that $a(x)$ is not identifiable in regions where $u(x)$ is identically constant. Our goal now is to show that $a(x)$ is identifiable in all other parts of $\bar{\Omega}$. To this end we set

$$\Omega_1 := \bar{\Omega} - \Omega_0 \quad \text{and} \quad \Omega_2 := \Omega_1 - C.$$

LEMMA 1. *Let $V_0 \subset \Omega$ be any volume, and denote by $V(t)$ the volume which is obtained from V_0 by following the gradient flow for time t, i.e.*

$$V(t) := \{P\colon \exists Q \in V_0 \text{ with } P = \phi_t(Q)\}.$$

Then

$$\int_{V_0} a(x)\,dx = \int_{V(t)} a(x)\,dx$$

for all t for which $V(t) \subset \Omega$.

Proof. We apply the Reynolds transport theorem (cf. [10, p. 443]) and the divergence theorem to obtain the following:

$$\begin{aligned}\frac{d}{dt}\left(\int_{V(t)} a(x)\,dx\right) &= \int_{V(t)} \partial_t a(x)\,dx + \int_{\partial V(t)} a\nabla u \cdot n\,dx \\ &= \int_{V(t)} \operatorname{div}(a\nabla u)\,dx = 0.\end{aligned}$$

Therefore the integral of $a(x)$ over any volume which follows the flow remains constant. □

LEMMA 2. *Suppose there exists an open set $U \subset \Omega_2$, such that either $U \subset \Omega_b$ or $U \subset \Omega_f$. Then $a \equiv 0$ on U.*

Proof. We assume $P \in U \subset \Omega_f$, and $a(P) > 0$, the other cases being similar.

Since a is continuous, there is an open set $W \subset \Omega_2$ on which a is everywhere positive. Moreover, since P is not a critical point of u, the level set of u through P is a smooth section for the flow ϕ_t. Let S denote an open disk in this level set containing P and contained in W. Also define $R_s := [s, \infty)$ and $I_s := [0, s)$. The map $\phi\colon R_s \times S \to \Omega$ given by $\phi(t, Q) = \phi_t(Q)$ is a diffeomorphism onto its image which we denote by V_s. Moreover, by the remarks following (3), a is positive on V_s. If we set $W_s := \phi(I_s \times S)$, we find that $V_s \cup W_s = V_0$.

Now since V_s is an open subset of the bounded region Ω, $a(x)$ is integrable over V_s and, in fact, its integral over V_s is a finite positive number. But clearly

$$\int_{V_s} a(x)\,dx + \int_{W_s} a(x)\,dx = \int_{V_0} a(x)\,dx,$$

so

$$\int_{V_s} a(x)\,dx < \int_{V_0} a(x)\,dx,$$

which contradicts Lemma 1. (See Fig. 1).

THEOREM 1. *$a(x)$ is identifiable on the closure of* int (Ω_2).

Proof. Suppose int (Ω_2) is not empty, and let P belong to this set. If there exists a neighborhood U of P which belongs entirely to either Ω_f or Ω_b, then $a(P)=0$ by virtue of Lemma 2. Otherwise, every open neighborhood U of P contains points of Ω_f and Ω_b.

Let us consider the remaining case, and assume WLOG that $P \in \Omega_f \cap \Omega$. Let U be an open neighborhood of P and set $V := U \cap \Omega_+$ (recall that Ω_+ is the complement of Ω_f, and relatively open by Proposition 1). Then V is not empty, open, and $P \in \partial V$. Therefore $a(x)$ vanishes on V due to Lemma 2, $a(x)$ being continuous, and $P \in \partial V$ implies $a(P)=0$. □

To illustrate Proposition 2 and Theorem 1, consider the following simple example. Let Ω denote the interior of the unit disk in R^2 and let $u = x^3/3 - y^2/2$. The phase portrait of ∇u, shown in Fig. 2, is easily described since the origin is the only stationary point and it is a saddle node.

Now, in our notation, the portion of the region $\bar{\Omega}$ in the closed left half-plane together with the positive x-axis is Ω_2, and the portion of the region $\bar{\Omega}$ in the open first and fourth quadrants is Ω_0. Thus, by our theorems, the parameter a in div $(a\nabla u)=0$ is identifiable in the closed left half-plane, and not identifiable on the open first and fourth quadrants. However, in this example, we can make explicit calculations to illustrate our results. Moreover, we can see that a is not identifiable at points in Ω_2 along the positive x-axis. Hence we cannot replace the closure of int (Ω_2) by Ω_2 in Theorem 1.

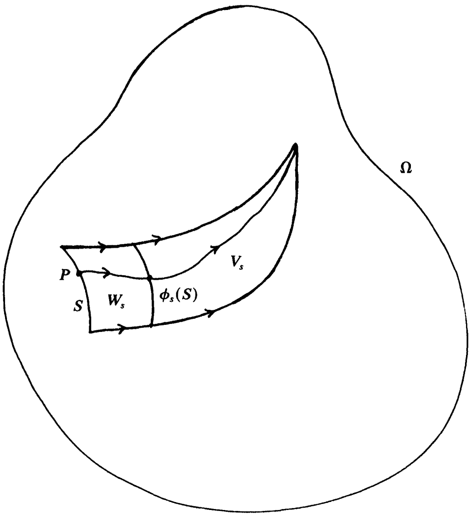


FIG. 1. *Section S and "vortex" tube given by the gradient flow of u.*

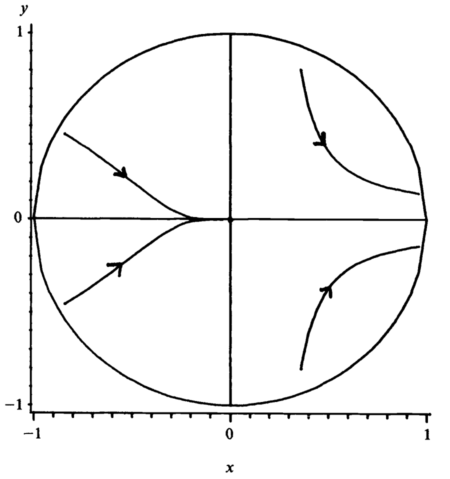


FIG. 2. *Phase portrait for the gradient of* $u = x^3/3 - y^2/2$.

For the calculations just consider the PDE for $a(x, y)$ given by $\operatorname{div}(a\nabla u) = 0$ as the first order PDE

$$x^2 a_x - y a_y + (2x-1)a = 0,$$

which has the general solution

$$a(x, y) = f_i(y/e^{1/x})/(x^2\, e^{1/x})$$

for any smooth functions f_i, where f_1 is defined for points in the left half-plane, and f_2 for points in the right half-plane subject only to the constraint that $a(x, y)$ be smooth on the y-axis.

However, for any choice of f_1 nonvanishing, $a(x, y)$ will fail to be continuous at the origin. To see this, observe that the characteristic curves of the PDE are given by $y = ce^{1/x}$. Thus, to have $a(x, y)$ defined continuously at the origin, the limit of $f_1(c)/(x^2 e^{1/x})$ must be finite as $x \to 0^-$. But the only way to insure this is to take $f_1 \equiv 0$. Therefore, any solution $a(x, y)$ must vanish in the closed left half-plane, and $a(x, y)$ is identifiable there as predicted by Theorem 1.

To see that points in the right half-plane are not identifiable we may choose $f_1 \equiv 0$, and $f_2 \equiv 1$, so

$$a(x, y) = \begin{cases} 1/x^2 e^{1/x}, & x > 0, \\ 0, & x \leqq 0. \end{cases}$$

We obtain $a(x, y)$ as a smooth parameter which is nonzero in the open right half-plane (including the positive x-axis) as predicted by Proposition 2.

4. Test conditions. In practice the function u is not arbitrary. It is obtained from the (unknown) parameter $a(x)$ and the function $f(x)$ as a solution of (2)

$$-\operatorname{div}(a\nabla u) = f$$

on Ω with appropriate boundary conditions. While for the well-posedness of this problem it is necessary that $a(x) \geqq a_0 > 0$, we may drop this last requirement in our approach. We assume that the functions a, u and f are given so that (2) is satisfied, and ask for conditions which imply that a is unique.

In this section we will use the term "identifiable" in a slightly modified way. For fixed f and u the parameter a is called identifiable at a point $P \in \bar{\Omega}$ if any parameter which satisfies (2), with possibly prescribed Cauchy data, agrees with a at the point

P. Identifiability on subsets of $\bar{\Omega}$ is defined similarly. The connection to our original definition becomes obvious when we consider the difference of two solutions for a.

PROPOSITION 3. *Suppose a is a solution of* (2), *satisfying* $a(x) \geqq a_0 > 0$, *and no Cauchy data for a are prescribed. Then if f vanishes on Ω, a is not identifiable on $\bar{\Omega}$.*

The proof of this result is trivial. However, Proposition 3 clearly shows that no information about the flow rates $q = a\nabla u$ can be obtained if the rates at which water is added to or taken from the aquifer are identically zero throughout the region Ω. If f vanishes on a subregion U of Ω, the argument in the proof cannot be repeated directly without introducing discontinuities if U is adjacent to a region on which a is identifiable.

PROPOSITION 4. *Let u be a solution of* (2) *with $u \equiv u_0$ on $\partial\Omega$, where u_0 is a constant, and suppose that* int $(C) = \emptyset$. *Then a is identifiable on $\bar{\Omega}$.*

Proof. Let $P \in \bar{\Omega}$, and assume that P is not a critical point of u. Then the trajectory $\phi_t(P)$ contains at most one point of $\partial\Omega$, since u increases along any trajectory, and u is constant on the boundary. Therefore $\Omega_0 = \emptyset$, $\bar{\Omega} = \Omega_1$ and $\Omega_2 = \bar{\Omega} - C$. Thus int $(\Omega_2) = \Omega - C$, and the closure of int (Ω_2) is all of $\bar{\Omega}$, since int $(C) = \emptyset$. The identifiability of a on $\bar{\Omega}$ now follows from Theorem 1. □

PROPOSITION 5. *Let u be a solution of* (2) *with $a\partial u/\partial n = g$ on $\partial\Omega$, and assume that* int $(C) = \emptyset$. *Then a is identifiable on $\bar{\Omega}$.*

Proof. Suppose we have two functions $a_1(x)$, and $a_2(x)$ which meet the conditions stated in the hypotheses; then their difference $a(x) := a_1(x) - a_2(x)$ satisfies div $(a\nabla u) = 0$ on Ω. Also $a\partial u/\partial n = 0$ on $\partial\Omega$.

Suppose $P \in \Omega_+ \cup \Omega_-$, and assume $P \in \Omega_+$, the case $P \in \Omega_-$ being similar. Then there is a point $Q \in \partial\Omega$ such that $Q = \phi_t(P)$ and $\phi_s(Q) \in (D - \bar{\Omega})$ for small $s > 0$.

There are two cases. First suppose $a(Q) = 0$. Since a satisfies the linear equation (3) along $\phi_t(Q)$ it follows that $a(P) = 0$. If $a(Q) \neq 0$, we find a neighborhood U of Q on $\partial\Omega$ on which a is nonzero. Hence $\partial u/\partial n \equiv 0$ on U. But in this case $\phi_t(t)$ will remain in $U \cap \partial\Omega$, which contradicts the choice of Q. Hence we have shown that $a(P) = 0$ on $\Omega_+ \cup \Omega_-$, and since $\Omega_0 \subset \Omega_+ \cup \Omega_-$, it follows that $a \equiv 0$ on Ω_0.

To complete the proof, we recall that Theorem 1 implies $a \equiv 0$ on the closure of int (Ω_2). Since $\bar{\Omega} = \Omega_0 \cup \Omega_2 \cup C$, and int $(C) = \emptyset$, the continuity of $a(x)$ implies $a(x) \equiv 0$ on $\bar{\Omega}$. □

Proposition 5 agrees with a result of Falk's [3] without using the assumption that $\nabla u \cdot \nu \geqq \sigma > 0$ on Ω for some fixed vector ν and a constant σ.

REFERENCES

[1] G. ALESSANDRINI, *On the identification of the leading coefficient of an elliptic equation*, Boll. Un. Mat. Ital. C (6), 4, (1985), pp. 87-111.

[2] ———, *An identification problem for an elliptic equation in two variables*, Ann. Mat. Pura Appl., to appear.

[3] R. S. FALK, *Error estimates for the numerical identification of a variable coefficient*, Math. Comp., 40 (1983), pp. 537-546.

[4] J. GERLACH AND R. B. GUENTHER, *Remarks on parameter identification* II, *Convergence of the method*, Numer. Math., submitted.

[5] K. H. HOFFMANN AND J. SPREKELS, *On the identification of the coefficients of elliptic problems by asymptotic regularization*, Numer. Funct. Anal. Optim., 7 (1984/5), pp. 157-177.

[6] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, SIAM J. Control Optim., 15 (1977), pp. 785-802.

[7] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217-241.

[8] K. KUNISCH, *Inherent identifiability: rate of convergence for parameter estimation problems*, preprint, Univ. of Graz, Austria, 1985.

[9] K. KUNISCH AND L. W. WHITE, *Identifiability under approximation for an elliptic boundary value problem*, preprint, University of Graz, Austria, 1985.

[10] C. C. LIN AND L. A. SEGEL, *Mathematics Applied to Deterministic Problems in the Natural Sciences*, Macmillan, New York-London, 1974.

[11] G. R. RICHTER, *An inverse problem for the steady state diffusion equation*, this Journal, 41 (1981), pp. 210-221.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

013

EXISTENCE OF GENERALIZED SOLUTIONS OF A NONLINEAR DIFFUSION CAUCHY PROBLEM*

CARMEN CORTAZAR† AND MANUEL ELGUETA†

Abstract. In this paper we prove the existence of a generalized solution of the initial value problem $\partial u/\partial t = \partial/\partial x\, \Phi(u, \partial\Sigma(u)/\partial x)$; $u(x, 0) = u_0(x)$ and the uniqueness of such a solution under certain conditions.

Key words. existence, parabolic, nonlinear, Cauchy problem

AMS(MOS) subject classifications. 35K65, 35D05

1. Introduction. In this note we prove the existence of at least one generalized solution of the initial value problem

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\Phi\left(u,\frac{\partial\Sigma(u)}{\partial x}\right), \tag{1.1}$$

$$u(x,0)=f(x). \tag{1.2}$$

This is a generalization of the equation $\partial u/\partial t = \partial^2\Sigma(u)/\partial x^2$ that includes $\partial u/\partial t = \partial/\partial x(|\partial u/\partial x|^{N-1}\partial u/\partial x)$ and $\partial u/\partial t = \partial/\partial x(|\partial u^{\lambda}/\partial x|^{N-1}\,\partial u/\partial x)$.

By a *generalized solution* to the above problem we understand a function $u(x, t)$ defined on $S = \mathbb{R}\times[0,\infty)$ so that

(i) $u(x, t)$ is continuous in $S=\mathbb{R}\times[0,\infty)$;

(ii) $\partial\Sigma(u)/\partial x$, which exists in the sense of distributions, belongs to L^{∞};

(iii) $$\iint\left[u\phi_t-\Phi\left(u,\frac{\partial\Sigma(u)}{\partial x}\right)\phi_x\right]dx\,dt + \int_{-\infty}^{+\infty} f(x)\phi(x,0)\,dx=0 \quad \forall\phi\in C_0^{\infty}(S).$$

Throughout this paper the functions $\Phi:[0,\infty)\times\mathbb{R}\to\mathbb{R}$ and $\Sigma:[0,\infty)\to[0,\infty)$ will satisfy

$$\Phi\in C^1([0,\infty)\times\mathbb{R}), \Sigma\in C^1([0,\infty)), \Sigma\in C^2([0,\infty)), \tag{1.3}$$

$$\begin{gathered} y_1<y_2\Rightarrow\Phi(z,y_1)<\Phi(z,y_2) \quad \forall z\in[0,\infty), \\ \Phi(z,-y)=-\Phi(z,y) \quad \forall(z,y)\in[0,\infty)\times\mathbb{R}, \end{gathered} \tag{1.4}$$

(1.5) there exist constants $\alpha_1>0$, $\alpha_2>0$, $\beta>0$, $C>0$ and a function $\psi\in C^{\infty}([0,\infty))$, $\psi(s)>0$ for all $s>0$ and $\int_0^z \psi^{1/\alpha_1}(s)\Sigma'(s)s\,ds<\infty$ so that $|\Phi(z,y)|\leqq\psi(z)|y|^{\alpha_1}$ if $|y|\leqq\beta$ and $|\Phi(z,y)|\geqq C|y|^{\alpha_2}$ if $|y|\geqq\beta$,

(1.6) $\Sigma'(s)\geqq 0$ and $\Sigma'(s)>0$ if $s>0$.

We will produce a generalized solution to problem (1.1), (1.2) as the limit of classical solutions of strictly parabolic problems that converge in a suitable way to

* Received by the editors March 15, 1984; accepted for publication June 19, 1986. This work was supported in part by Dirección de Investigación de la Universidad Católica de Chile through DIUC Grant and by Fondo Nacional de Desarrollo Cientifico y Tecnológico.

† Universidad Católica de Chile, Facultad de Matemáticas, Casilla 114-D, Santiago, Chile.

problem (1.1), (1.2). Namely, for $n=0,1,2,\cdots$, let $\Phi_n:[0,\infty)\times\mathbb{R}\to\mathbb{R}$ and $\Sigma_n:[0,\infty)\to[0,\infty)$ be functions that satisfy hypotheses (1.3)-(1.6) with the constants α_1, α_2, β, C and the function ψ independent of n, and such that $\Sigma_n\in C^2([0,\infty))$. Assume moreover that

$$\frac{\partial\Phi_n}{\partial y}(z,y)\geqq a_n\gneqq 0 \quad \forall n, \tag{1.7}$$

$$\Sigma_n'(s)\geqq a_n\gneqq 0 \quad \forall n, \tag{1.8}$$

$$\Phi_0(z,y)\geqq\Phi_n(z,y) \quad \forall z,\quad \forall y\geqq 0,\quad n=0,1,2,\cdots, \tag{1.9}$$

$$\Sigma_0'(s)\geqq\Sigma_n'(s) \quad \forall s\geqq 0,\quad n=0,1,2,\cdots, \tag{1.10}$$

and

(1.11) $\Phi_n(z,y)\to\Phi(z,y)$ as $n\to+\infty$ uniformly on compact subsets,

(1.12) $\Sigma_n'(s)\to\Sigma'(s)$ as $n\to+\infty$ uniformly on compact subsets and $\Sigma_n(0)\to\Sigma(0)$ as $n\to+\infty$,

(1.13) Let $f_n:\mathbb{R}\to\mathbb{R}$ so that $f_n\to f$ on $L^1(\mathbb{R})$ as $n\to+\infty$ and $\|f_n(x)\|_\infty\leqq M<\infty$ with M independent of n. Let $u_n(x,t)$ be the solution in S of the strictly parabolic problem,

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\Phi_n\left(u,\frac{\partial\Sigma_n(u)}{\partial x}\right), \tag*{(1.1)$_n$}$$

$$u(x,0)=f_n(x). \tag*{(1.2)$_n$}$$

(For the existence of such a solution see [13].) We will refer to $(1.1)_n$, $(1.2)_n$ as an approximating sequence to problem (1.1), (1.2).

More precisely we will prove the following.

THEOREM 1. *Let $u_n(x,t)$ be a solution of problem* $(1.1)_n$, $(1.2)_n$ *with Φ_n, Σ_n and f_n as in the preceding paragraph and such that*

$$\left\|\frac{\partial}{\partial x}\Phi_n\left(f_n,\frac{\partial\Sigma_n(f_n)}{\partial x}\right)\right\|_1\leqq P \tag{1.14}$$

with P independent of n. Then there exists a subsequence of $\{u_n(x,t)\}_{n=1}^\infty$ which converges uniformly on compact subsets of S to a generalized solution $u(x,t)$ of problem (1.1), (1.2). *Moreover, for a fixed $t\geqq 0$ the functions $u_n(\cdot,t)$ converge to $u(\cdot,t)$ in $L^1(\mathbb{R})$.*

As a corollary we obtain the following existence result.

COROLLARY 1. *If Φ and Σ satisfy hypotheses* (1.3)-(1.7) *and f, $\partial f/\partial x$, $\partial^2 f/\partial x^2$ and $\partial^2\Sigma(f)/\partial x^2$ belong to $L^1(\mathbb{R})$ then the initial value problem* (1.1), (1.2) *has at least a generalized solution in S.*

Related existence results appear in [4], [6], [7], and [13]-[15].

The next theorem shows that no matter which sequence of approximating problems is taken, the solution we obtain is the same. In particular this proves that the sequence $\{u_n(x,t)\}_{n=1}^\infty$ of Theorem 1 itself converges to a generalized solution of problem (1.1), (1.2).

THEOREM 2. *Let $\Phi_n^{(i)}$, $\Sigma_n^{(i)}$, $f_n^{(i)} i=1,2$; $n=0,\cdots$, be sequences of functions that satisfy hypotheses* (1.3)-(1.14).

Let $u^{(i)}(x,t) i=1,2$ be defined by

$$u^{(i)}(x,t)=\lim_{n\to+\infty}u_n^{(i)}(x,t)$$

where $u_n^{(i)}(x, t)$ *is the solution of problem*

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\Phi_n^{(i)}\left(u, \frac{\partial \Sigma_n^{(i)}(u)}{\partial x}\right), \qquad u(x,0)=f_n^{(i)}(x).$$

Then $u^{(1)}(x, t)=u^{(2)}(x, t)$ *for all* $(x, t)\in S$.

Although we have not been able to prove uniqueness of the generalized solution of problem (1.1), (1.2) we have the following uniqueness theorem.

THEOREM 3. *Let* $u(x, t)$ *and* $v(x, t)$ *be two generalized solutions of problem* (1.1), (1.2). *In addition to hypotheses* (1.3)-(1.7) *assume that* $\Phi\in C^\infty([0,\infty)\times\mathbb{R})$, $\Sigma\in C^\infty[0,\infty))$ *and* $\partial\Phi/\partial y(z, y)\geqq a>0$, *for all* $(z, y)\in[0,\infty)\times\mathbb{R}$. *Then* $u\equiv v$.

For some uniqueness results see [3], [5] and [15].

We would like to note that several properties of the solutions of strictly parabolic problems are inherited by the limit solution given by Theorem 1, among them the comparison theorems as they appear in [8]. This permits us, using the same techniques as in [9] and [12], to study asymptotic behaviour of the solutions so obtained. The details will appear somewhere else.

2. Proof of Theorem 1. Let $u_n(x, t)$ be the solution of problem $(1.1)_n$, $(1.2)_n$; it is known that these are classical solutions. Since $\|u_n(\cdot,\cdot)\|_\infty\leqq\|f_n\|_\infty\leqq M$ we obtain, by (1.10), $\|\Sigma_n(u_n(\cdot,\cdot))\|_\infty\leqq\Sigma_0(M)$. We will prove now that the sequence $\Sigma_n(u_n)$ is equicontinuous in S.

By Theorem 1 of [8] and (1.14) we get

$$\left\|\frac{\partial}{\partial x}\Phi_n\left(u_n, \frac{\partial\Sigma_n(u_n)}{\partial x}\right)\right\|_1\leqq\left\|\frac{\partial}{\partial x}\Phi_n\left(f_n, \frac{\partial\Sigma_n(f_n)}{\partial x}\right)\right\|_1\leqq P \quad \forall n.$$

Therefore

$$\left\|\Phi_n\left(u_n, \frac{\partial\Sigma_n(u_n)}{\partial x}\right)\right\|_\infty\leqq P$$

and, by (1.5), recalling that the constants do not depend on n, we obtain

$$\left\|\frac{\partial\Sigma_n}{\partial x}(u_n)\right\|_\infty\leqq C \quad \forall n\geqq 0.$$

Hence, for a fixed t, $\Sigma_n(u_n(\cdot, t))$ is Lipschitz continuous in x, with Lipschitz constant C independent of t and n.

On the other hand, since the solution $u_n(x, t)$ is given by a contraction semigroup,

$$\|u_n(\cdot, t_1)-u_n(\cdot, t_2)\|_1\leqq P\quad|t_1-t_2|$$

(see [8], [13]).

Therefore, by (1.10),

$$\|\Sigma_n(u_n(\cdot, t_1))-\Sigma_n(u_n(\cdot, t_2))\|_1\leqq C\cdot P\quad|t_1-t_2|.$$

This, and the uniform Lipschitz continuity in x, imply by a standard argument that $\{\Sigma_n(u_n)\}_{n=0}^\infty$ is equicontinuous in S.

Now the Ascoli-Arzela Theorem gives a subsequence $\{\Sigma_{n_k}(u_{n_k})\}_{k=0}^\infty$ and hence a subsequence $\{u_{n_k}\}_{k=0}^\infty$ that converges uniformly on compact subsets of S to a continuous function u. An immediate consequence of this is that $\partial\Sigma_{n_k}(u_{n_k})/\partial x$ converges weakly to $\partial\Sigma(u)/\partial x$. Since $\|\partial\Sigma_{n_k}(u_{n_k})/\partial x\|_\infty$ are uniformly bounded we get $\partial\Sigma(u)/\partial x\in L^\infty$.

We will now prove that $u(x, t)$ satisfies (iii) of the definition of a generalized solution.

We define

$$H_n(z, \eta)=y \Leftrightarrow \Phi_n(z, y)=\eta \quad \text{and} \quad H(z, \eta)=y \Leftrightarrow \Phi(z, y)=\eta.$$

Now, since for a fixed t,

$$\left\| \frac{\partial}{\partial x} \Phi_{n_k}\left(u_{n_k}, \frac{\partial \Sigma_{n_k}(u_{n_k})}{\partial x}\right)(\cdot, t)\right\|_1 \leqq P$$

by the inbedding [11, Thm. 11.2, p. 31], we have that for a fixed interval $[-A, A]$ there exists a subsequence $\{n_{k'}\}_{k'=0}^{\infty}$ of $\{n_k\}$ so that

$$\Phi_{n_{k'}}\left(u_{n_{k'}}, \frac{\partial \Sigma_{n_{k'}}(u_{n_{k'}})}{\partial x}\right)(\cdot, t) \rightarrow \eta(\cdot, t)$$

as $k' \rightarrow +\infty$ in $L^1[-A, A]$ and so we can extract another subsequence $\{n_{k''}\}$ so that

$$\eta_{n_{k''}}(\cdot, t)=\Phi_{n_{k''}}\left(u_{n_{k''}}, \frac{\partial \Sigma_{n_{k''}}(u_{n_{k''}})}{\partial x}\right)(\cdot, t) \rightarrow \eta(\cdot, t)$$

as $k'' \rightarrow +\infty$ a.e. in $[-A, A]$. But

$$\frac{\partial \Sigma_{n_{k''}}(u_{n_{k''}})}{\partial x}(x, t)=H_{n_{k''}}(u_{n_{k''}}, \eta_{n_{k''}})$$

and if we let $k'' \rightarrow +\infty$, since

$$\frac{\partial \Sigma_{n_{k''}}(u_{n_{k''}})}{\partial x} \rightarrow \frac{\partial \Sigma(u)}{\partial x}$$

weakly as $k'' \rightarrow +\infty$, we get

$$\frac{\partial \Sigma(u)}{\partial x}(x, t)=H(u, \eta)(x, t) \quad \text{a.e. in } [-A, A].$$

So

$$\eta(x, t)=\Phi\left(u(x, t), \frac{\partial \Sigma(u)}{\partial u}(x, t)\right) \quad \text{a.e. in } [-A, A]$$

and

$$\Phi_{n_{k''}}\left(u_{n_{k''}}(\cdot, t), \frac{\partial \Sigma_{n_{k''}}(u_{n_{k''}})}{\partial x}(\cdot, t)\right) \rightarrow \Phi\left(u(\cdot, t), \frac{\partial \Sigma(u)}{\partial x}(\cdot, t)\right)$$

as $k'' \rightarrow +\infty$ in $L^1([-A, A])$.

Therefore the same is true for the whole subsequence $\{u_{n_k}\}_{k=1}^{\infty}$. Since

$$\left\| \Phi_{n_k}\left(u_{n_k}, \frac{\partial \Sigma_{n_k}(u_{n_k})}{\partial x}\right)(\cdot, \cdot)\right\|_{\infty} \leqq P$$

the dominated convergence theorem implies that $\Phi_{n_k}(u_{n_k}, \partial \Sigma_{n_k}(u_{n_k})/\partial x)$ converges to $\Phi(u, \partial \Sigma(u)/\partial x)$ in $L^1([-A, A] \times [0, T])$ for our rectangle $[-A, A] \times [0, T]$.

Finally, letting $k \rightarrow +\infty$ in

$$\iint_S \left[u_{n_k} \phi_t-\Phi_{n_k}\left(u_{n_k}, \frac{\partial \Sigma_{n_k}(u_{n_k})}{\partial x}\right) \phi_x\right] dx\, dt+\int_{-\infty}^{+\infty} f_{n_k}(x) \phi(x, 0)\, dx=0,$$

we obtain that $u(x, t)$ is a generalized solution of (1.1), (1.2).

We have proved that the subsequence $\{u_{n_k}(x, t)\}_{k=1}^{\infty}$ converges uniformly on compact subsets to a generalized solution of problem (1.1), (1.2). It remains to prove that $\{u_{n_k}(\cdot, t)\}_{k=1}^{\infty}$ converges in $L^1(\mathbb{R})$ to $u(\cdot, t)$. Now we need the following lemma.

LEMMA 1. *Let $u_n(x, t)$ be a solution of problem $(1.1)_n$, $(1.2)_n$. Then given $\varepsilon > 0$ and $T > 0$ there exists $A > 0$ so that*

$$\int_{|y| \geqq A} u_n(y, t)\, dy \leqq \varepsilon \quad \forall n, \quad \forall t \in [0, T].$$

Proof. The main tool in the proof is the comparison theorem [8, Thm. 3] (see also [16]).

We prove first the case $f_n = f \in C_0^{\infty}$. In this case let $g \in C_0^{\infty}$ symmetric with respect to $x = 0$, increasing for $x \leqq 0$ and $f(x) \leqq g(x)$ for all x.

Let $v_n(x, t) n = 0, 1, 2, \cdots$, be the solution of

$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial x} \Phi_n\left(v, \frac{\partial \Sigma_n(v)}{\partial x}\right), \qquad v(x, 0) = g(x).$$

According to the above-mentioned comparison theorem, we have

$$\int_{|y| \geqq A} u_n(y, t)\, dy \leqq \int_{|y| \geqq A} v_n(y, t)\, dy \leqq \int_{|y| \geqq A} v_0(y, t) \leqq \int_{|y| \geqq A} v_0(y, T)\, dy.$$

As $v_0(\cdot, T) \in L^1(\mathbb{R})$ the lemma is proved in the case $f_n = f \in C_0^{\infty}$. The general case follows now from the well-known fact that

$$\|u_n(\cdot, t) - \tilde{u}_n(\cdot, t)\|_1 \leqq \|u_n(\cdot, 0) - \tilde{u}_n(\cdot, 0)\|_1$$

where u_n and $\tilde{u}_n$ are solutions of $(1.1)_n$. This ends the proof of Lemma 1.

Lemma 1 and the fact that $u_{n_k} \to u$ uniformly on compact subsets of S imply that $u_{n_k}(\cdot, t) \to u(\cdot, t)$ in $L^1(\mathbb{R})$. This proves Theorem 1.

Remark. We would like to observe that an alternative proof to Theorem 1 could be obtained by using nonlinear semigroup theory, in particular P. Benilan's result that states that if A_n, $n = 1, 2, \cdots, \infty$ is an m-accretive operator in a Banach space and $A_n \to A_\infty$ as $n \to +\infty$ in a suitable way; then for the corresponding semigroups $S_n(t)$ one has $S_n(t) \to S_\infty(t)$ as $n \to +\infty$. For a precise statement of this result see [1], [2] or [10].

Proof of Corollary 1. From the hypothesis it is easy to construct sequences Φ_n, Σ_n that satisfy the requirements of Theorem 1 with $f_n = f$, for all n.

3. Proof of Theorem 2. First we need the following lemma.

LEMMA 2. *Let $u_n(x, t)$ be a solution of problem $(1.1)_n$, $(1.2)_n$ and assume that $0 < a \leqq f_n(x)$, for all $x \in [-A, A]$ for all n. Then given $T > 0$ there exists $b > 0$ independent of n so that $u_n(x, t) > b$ for all $(x, t) \in [-A, A] \times [0, T]$.*

Proof. Let g be C_0^{∞}, symmetric with respect to $-A$ increasing for $x \leqq -A$, $g(-A) \neq 0$, $g(x) \leqq f_n(x)$ for all $x \in \mathbb{R}$.

Let $v_n(x, t)$ be a solution of $(1.1)_n$ with initial condition g. The comparison theorem ([8], [16]) implies

$$\int_{-\infty}^{x} v_n(y, t) \leqq \int_{-\infty}^{x} v_0(y, T) \qquad \forall x \leqq -A, \quad \forall t \in [0, T].$$

Therefore, since $\int_{-\infty}^{-A} v_n = \int_{-\infty}^{-A} v_0 = \int_{-\infty}^{-A} g$,

$$\int_{x}^{-A} v_n(y, t)\, dy \geqq \int_{x}^{0} v_0(y, T)\, dy \quad \forall x \leqq -A.$$

Hence

$$v_n(-A, t) \geqq v_0(-A, T).$$

Pointwise comparison (see [8], [13]) gives

$$u_n(-A, t) \geqq v_0(-A, T) > 0.$$

Similarly,

$$u_n(A, t) \geqq v_0^+(A, T) > 0$$

where $v_0^+(A, T)$ is a solution to $(1.1)_0$ with initial condition $g^+ \in C_0^\infty$, symmetric with respect to A, increasing for $x \leqq A$, $g^+(A) \neq 0$, $g^+(x) \leqq f_n(x)$ for all $x \in \mathbb{R}$.

Since, by hypothesis, $u_n(x, 0) = f_n(x) \geqq a > 0$ for $x \in [-A, A]$, the minimum principle for parabolic equations implies that

$$u_n(x, t) > \min(a, v_0(-A, T), v_0^+(A, T)) = b > 0$$

for all $(x, t) \in [-A, A] \times [0, T]$ and the lemma is proved.

Proof of Theorem 2. Let $\varepsilon > 0$ be given. There exists $\tilde{f}$: $\mathbb{R} \to \mathbb{R}$ and $N > 0$ so that $\tilde{f}$, $\partial \tilde{f}/\partial x$, $\partial^2 \tilde{f}^2/\partial x^2 \in L^1(\mathbb{R})$ and $\|f_n - \tilde{f}\| \leqq \varepsilon$ for all $n \geqq N$.

Let $\tilde{u}_n^{(i)}(x, t)$ be the solution of $(1.1)_n$ with initial condition $\tilde{u}_n^{(i)}(x, 0) = \tilde{f}(x)$.

It is well known that

$$\|\tilde{u}_n^{(i)}(\cdot, t) - u_n^{(i)}(\cdot, t)\|_1 \leqq \|\tilde{f} - f_n\|_1. \tag{3.1}$$

Since

$$\begin{aligned} e^{-t} \int_{-\infty}^{x} [u_n^{(1)}(y, t) - u_n^{(2)}(y, t)]\, dy = {} & e^{-t} \int_{-\infty}^{x} [u_n^{(1)}(y, t) - \tilde{u}_n^{(1)}(y, t)]\, dy \\ & + e^{-t} \int_{-\infty}^{x} [\tilde{u}_n^{(1)}(y, t) - \tilde{u}_n^{(2)}(y, t)]\, dy \\ & + e^{-t} \int_{-\infty}^{x} [\tilde{u}_n^{(2)}(y, t) - u_n^{(2)}(y, t)]\, dy, \end{aligned}$$

and ε is arbitrary, by (3.1) in order to show that $u^{(1)} \equiv u^{(2)}$ it suffices to show that for

$$J_n(x, t) = e^{-t} \int_{-\infty}^{x} [\tilde{u}_n^{(1)}(y, t) - \tilde{u}_n^{(2)}(y, t)]\, dy$$

we have

$$\max_{(x,t) \in S} |J_n(x, t)| \to 0 \quad \text{as } n \to +\infty. \tag{3.2}$$

In order to prove (3.2) assume that

$$\overline{\lim_{n \to +\infty}} \max_{(x,\, t) \in S} J_n(x, t) = d > 0;$$

then there exists $n > 0$ and $T > 0$ such that

$$\max_{(x,t) \in S} J_n(x, t) = \max_{(x,\, t) \in \mathbb{R} \times [0,\, T]} J_n(x, t) \geqq \frac{d}{2} > 0. \tag{3.3}$$

By Lemma 1 we can find $A>0$ so that

$$\max_{(x,t)\in S} J_n(x,t)=J_n(x_n,t_n)$$

and $(x_n, t_n)\in[-A, A]\times[0, T]$.

Since the maximum of $J_n(x, t)$ is attained at (x_n, t_n), the relations $(\partial/\partial x)J_n(x_n, t_n)=0$, $(\partial^2/\partial x^2)J_n(x_n, t_n)\leqq 0$ and $(\partial J_n/\partial x)(x_n, t_n)=0$ imply

$$\tilde{u}_n^{(1)}(x_n,t_n)=\tilde{u}_n^{(2)}(x_n,t_n), \tag{3.4}$$

$$\frac{\partial\tilde{u}_n^{(1)}}{\partial x}(x_n,t_n)\leqq\frac{\partial\tilde{u}_n^{(2)}}{\partial x}(x_n,t_n), \tag{3.5}$$

$$\begin{aligned} J_n(x_n,t_n)=\Phi_n^{(1)}&\left(\tilde{u}_n^{(1)}(x_n,t_n),\frac{\partial\Sigma_n^{(1)}(\tilde{u}_n^{(1)})}{\partial x}(x_n,t_n)\right)\\ &-\Phi_n^{(2)}\left(\tilde{u}_n^{(2)}(x_n,t_n),\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}(x_n,t_n)\right). \end{aligned} \tag{3.6}$$

So from (3.3) and (3.6) we obtain

$$0<\frac{d}{2}\leqq I_n^{(1)}+I_n^{(2)}+I_n^{(3)}$$

where

$$I_n^{(1)}=\Phi_n^{(1)}\left(\tilde{u}_n^{(1)},\frac{\partial\Sigma_n^{(1)}(\tilde{u}_n^{(1)})}{\partial x}\right)-\Phi\left(\tilde{u}_n^{(1)},\frac{\partial\Sigma^{(1)}(\tilde{u}_n^{(2)})}{\partial x}\right),$$

$$I_n^{(2)}=\Phi\left(\tilde{u}_n^{(2)},\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right)-\Phi_n^{(2)}\left(\tilde{u}_n^{(2)},\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right),$$

$$I_n^{(3)}=\Phi\left(\tilde{u}_n^{(1)},\frac{\partial\Sigma_n^{(1)}(u_n^{(1)})}{\partial x}\right)-\Phi\left(\tilde{u}_n^{(2)},\frac{\partial\Sigma_n^{(2)}(u_n^{(2)})}{\partial x}\right),$$

evaluated at the point (x_n, t_n).

From the fact that $\Phi_n^{(i)}\to\Phi$ uniformly on compact subsets as $n\to+\infty$ we have that $I_n^{(1)}$, $I_n^{(2)}\to 0$ as $n\to+\infty$. Therefore in order to obtain the contradiction we are looking for it suffices to prove that $\overline{\lim}\, I_n^{(3)}\leqq 0$ as $n\to+\infty$.

Using (3.4) and (3.5) and the monotonicity of the function Φ on the second variable we get

$$\begin{aligned} I_n^{(3)}&\leqq\left|\Phi\left(\tilde{u}_n^{(2)},\frac{\partial\Sigma_n^{(1)}(\tilde{u}_n^{(2)})}{\partial x}\right)-\Phi\left(\tilde{u}_n^{(2)},\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right)\right|\\ &\leqq K\left|\frac{\partial\Sigma_n^{(1)}(\tilde{u}_n^{(2)})}{\partial x}-\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right| \end{aligned}$$

where K is a Lipschitz constant for the C^1 function $\Phi(\cdot,\cdot)$ in a suitable compact set and does not depend on n.

Or, setting $\sigma_n^{(i)}(s)=(\Sigma_n^{(i)})'(s)$, we get

$$I_n^{(3)}\leqq K\left\|\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right\|_\infty\cdot\left|\frac{\sigma_n^{(1)}(\tilde{u}_n^{(2)})-\sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\sigma_n^{(2)}(\tilde{u}_n^{(2)})}\right|.$$

But we know that

$$\left\|\frac{\partial\Sigma_n^{(2)}(\tilde{u}_n^{(2)})}{\partial x}\right\|_\infty\leqq K,$$

independently of n and, by Lemma 5 and the hypothesis $\sigma_n^{(2)} \to \sigma$ uniformly on compact subsets as $n \to +\infty$, we have $|\sigma_n^{(2)}(u_n(x_n, t_n))| \geqq c > 0$ for all n large enough where c is independent of n.

Therefore

$$I_n^{(3)} \leqq C|\sigma_n^{(1)}(\tilde{u}_n^{(2)}(x_n, t_n)) - \sigma_n^{(2)}(\tilde{u}_n^{(2)}(x_n, t_n))|$$

for a suitable constant C and letting $n \to +\infty$ we get $\overline{\lim}\, I_n^{(3)} \leqq 0$.

Analogously we can prove that

$$\underline{\lim}_{n \to +\infty} \max_{(x,t) \in S} J_n(x, t) \geqq 0$$

and consequently (3.2) is proved.

4. Proof of Theorem 3. The proof of Theorem 3 is based on the following lemma whose proof is exactly the same as the proof of Theorem 7 in [8], and we will not give it here.

LEMMA 3. *Let the hypothesis of Theorem* 3 *hold. If $u(x, t)$ is a generalized solution of problem* (1.1), (1.2), *then $u(x, t)$ is a classical solution of* (1.1) *in a neighborhood of any point (x_0, y_0) so that $u(x_0, t_0) > 0$.*

Now we can prove Theorem 3. We note that we can assume that $u(x, t)$ is the solution given by Theorem 1 to problem (1.1), (1.2). From this it follows that

$$u(x, t) = \lim_{n \to +\infty} u_n(x, t) \quad \text{in } L^1$$

where $u_n(x, t)$ is a solution of $(1.1)_n$ with initial condition

$$u_n(x, 0) = f_n(x)$$

where $f_n(x) > 0$ for all $x \in \mathbb{R}$ and $f_n \to f$ in L^1.

Now we essentially repeat the argument of the proof of Theorem 2.

Set

$$J_n(x, t) = e^{-t} \int_{-\infty}^{x} (u_n(y, t) - v(y, t))\, dy.$$

Assume that

$$\max_{(x,t) \in S} \int_{-\infty}^{x} (u(y, t) - v(y, t))\, dy = d > 0.$$

Then for n large enough we have

$$\max_{(x,t) \in S} J_n(x, t) = J_n(x_n, t_n) \geqq \frac{d}{2} > 0.$$

At the point (x_n, t_n) where this maximum is attained we have

$$v(x_n, t_n) = u_n(x_n, t_n) > 0.$$

So by Lemma 3, $v(x, t)$ is a classical solution in a neighborhood of (x_n, t_n); this lets us take higher order derivatives at this point to obtain, as in the proof of Theorem 2,

$$\frac{\partial u_n}{\partial x}(x_n, t_n) \leqq \frac{\partial v}{\partial x}(x_n, t_n)$$

and

$$0<\frac{d}{2}<J_n(x_n,t_n)=\Phi\left(u_n(x_n,t_n),\frac{\partial\Sigma(u_n)}{\partial x}\right)-\Phi\left(v(x_n,t_n),\frac{\partial\Sigma(v)}{\partial x}\right)\leqq 0,$$

which is a contradiction. Consequently

$$\int_{-\infty}^{x} u(y,t)\,dy \leqq \int_{-\infty}^{x} u(y,t)\,dy \quad \forall x,\quad \forall t.$$

Analogously we obtain the reverse inequality to get $\int_{-\infty}^{x} u(y,t)\,dy=\int_{-\infty}^{x} v(y,t)\,dy$ and hence $u\equiv v$.

REFERENCES

[1] Ph. Benilan, *Equations d'évolution dans un espace de Banach quelconque et applications*, Thesis, Univ. Orsay, 1972.

[2] Ph. Benilan and M. G. Crandall, *The continuous dependence on Φ of solutions of $u_t-\Delta\Phi(u)=0$*, Indiana Univ. Math. J., 30 (1981), pp. 162–177.

[3] J. E. Bouillet and C. Atkinson, *Qualitative properties of a generalized diffusion equation* II, preprint.

[4] H. Brezis, *On some degenerate non-linear parabolic equations*, Proc. Amer. Math. Soc., 18 (1970).

[5] H. Brezis and M. G. Crandall, *Uniqueness of solutions of the initial-value problem for $u_t-\Delta(u)=0$*, J. Math. Pures Appl., 58 (1979), pp. 153–163.

[6] J. R. Cannon and E. DiBenedetto, *On the existence of solution of boundary value problems in fast chemical reactions.* Boll. Un. Mat. Ital. B (5), 15 (1978), pp. 835–843.

[7] ———, *On the existence of weak-solutions to an n-dimensional Stefan problem with nonlinear boundary conditions*, this Journal, 11 (1980), pp. 632–645.

[8] C. Cortazar, *The application of dissipative operators to nonlinear diffusion equations*, J. Differential Equations, 47 (1983), pp. 1–23.

[9] C. Cortazar and M. Elgueta, *The asymptotic behaviour of the solution of a nonlinear diffusion equation*, this Journal, 16 (1985), pp. 251–258.

[10] L. G. Evans, *Applications of nonlinear semigroup theory to certain partial differential equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978.

[11] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[12] S. Kamin (Kamenomostskaya), *Similar solutions and the asymptotics of filtration equations*, Arch. Rational Mech. Anal., 60 (1976), pp. 171–183.

[13] O. A. Ladyzenskaja, V. A. Solonnikov and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Amer. Math. Soc. Transl., 23, Providence, RI, 1968.

[14] J. L. Lions, *Quelques méthodes de résolution des problemes aux limites non linéaires*, Dunod, Paris, 1960.

[15] O. Oleinick, *On some degenerate quasilinear parabolic equations*, Seminari 1962–1963, Edizioni Cremonese, Rome, 1965.

[16] J. L. Vazquez, *Symétrisation pour $u_t=\Delta\Phi(u)$ et applications*, C.R. Acad. Sci. Paris Sér. I. Math., 295 (1982), pp. 71–74.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

014

THREE-DIMENSIONAL TEMPERATURE RESPONSE TO IMPULSIVE INPUT OUTSIDE A SPHERICAL RESERVOIR*

JAMES A. COCHRAN†

Abstract. The classical expression for the Green's function associated with heat conduction external to a spherical reservoir involves an infinite series in which the radial/time contributions are given by Laplace-type integrals containing cross products of spherical Bessel functions. This expression is most useful, numerically, for large distances and/or long times. In this paper, using a combination of transform and integral equations techniques, alternative expressions are derived for the radial/time components predominantly in terms of *polynomials* in the time variable. These new representations are thus valuable for short-time heat conduction calculations. Moreover, when re-expressed in a quantum-mechanical setting, they provide particularly useful closed-form expressions in terms of familiar functions for all the higher-order partial waves of the two-particle density matrix under the assumption of hard-core interaction.

Key words. Green's functions, heat conduction, quantum mechanical scattering, spherical Bessel functions

AMS(MOS) subject classifications. Primary 35C10, 35K05, 42C10, 81F05; secondary 33A40

1. Introduction.

The three-dimensional temperature distribution $u(\vec{r}, \vec{r}'; t)$ in an unbounded homogeneous medium exterior to a spherical fixed-temperature reservoir, which results from an impulsive input at time $t=0$, satisfies the combined boundary/initial value problem

$$\begin{aligned} &\nabla^2 u = k\frac{\partial u}{\partial t} \quad |\vec{r}|, |\vec{r}'| > 1, \quad t>0; \\ &u \equiv 0 \quad \text{for } |\vec{r}| \leqq 1, \\ &\lim_{t\to 0} u = \delta(\vec{r}-\vec{r}'). \end{aligned} \tag{1}$$

Here k is a positive constant, $\vec{r}'$ is the impulsive source point and the reservoir has been both localized to the unit ball about the origin and given the normalized temperature zero. If separation of variables in spherical coordinates is then applied, the desired solution should have the following form, owing to symmetry and boundedness considerations:

$$u(\vec{r}, \vec{r}'; t) = \sum_{n=0}^{\infty} \left(\frac{2n+1}{4\pi}\right) u_n(r, r'; t) P_n(\cos\theta). \tag{2}$$

In this expression θ is the angle between the position vectors $\vec{r}$ and $\vec{r}'$, $r \equiv |\vec{r}|$, $r' \equiv |\vec{r}'|$, P_n is the Legendre polynomial of nonnegative order n and $u_n(r, r'; t)$ is the bounded solution of

$$\begin{aligned} &\frac{\partial}{\partial r}\left(r^2\frac{\partial u_n}{\partial r}\right) - n(n+1)u_n = kr^2\frac{\partial u_n}{\partial t}, \qquad r, r' > 1, \quad t>0, \\ &u_n(1, r'; t) = 0, \\ &\lim_{t\to 0} u_n(r, r'; t) = \frac{\delta(r-r')}{rr'}. \end{aligned} \tag{3}$$

* Received by the editors June 30, 1986; accepted for publication September 17, 1986. This research was performed while the author was visiting as a Gordon Fellow at Deakin University, Victoria, Australia. This work was supported in part by the Division of Mathematics and Computing at Deakin.

† Department of Mathematics, Washington State University, Pullman, Washington 99164-2930.

The classical representation [2, p. 382] for the solution of (3) can be expressed as the Laplace-type integral

$$-\frac{1}{2\pi}\int_0^\infty \frac{D_n(vr)D_n(vr')}{h_n^{(1)}(v)h_n^{(2)}(v)} e^{-v^2 t/k} v^2\, dv \qquad (n=0,1,2,\cdots)$$

where $D_n(z)\equiv h_n^{(1)}(z)h_n^{(2)}(v)-h_n^{(2)}(z)h_n^{(1)}(v)$ and the $h_n^{(1,2)}$ are the spherical Hankel functions of the first and second kinds, respectively, of nonnegative integer order n. When this integral representation is employed in (2), an expression for the Dirichlet heat conduction Green's function is obtained which is most useful, computationally, when r (or r') and t are large.

The problem (1) arises in other contexts as well, e.g., the quantum mechanical interaction of two particles under the assumption of hard cores. Recently, alternative representations have been derived in this setting for the lower-order solutions to (3) ([8], $n=1$, 2; see also [9], $n=0$) which show them to be polynomials in t modified by familiar functions such as exponentials, error functions and, of course, spherical Hankel functions. In this paper, we continue in this same spirit and, using a combination of transform and integral equations techniques, derive an analogous representation for the general solution of (3) in the case of arbitrary n. This expression proves to be of a form particularly advantageous in short-time heat conduction calculations.

The next section of this paper contains this new general solution of (3) along with remarks concerning the nature of the solution and its behavior in various limiting regimes. A word or two about computation of the solution components is included also. The major details of the solution derivation are presented in § 3.

2. General form of the solution $u_n(r, r'; t)$. As we shall show in the next section, the solution of (3) may be expressed as the three-term summation

$$u_n(r, r'; t)\equiv I_n(r, r'; t)+J_n(r, r'; t)+K_n(r, r'; t) \tag{4}$$

with each term in this representation having particular characteristics.

Singular term. The initial term in (4) is given by the formula

$$\begin{aligned} I_n(r, r'; t) &\equiv e^{-k(r^2+r'^2)/4t}(-i)^n\left(\frac{\pi}{4}\right)\left(\frac{k}{\pi t}\right)^{3/2} h_n^{(2)}\left(\frac{ikrr'}{2t}\right) \\ &= \frac{e^{-k(r-r')^2/4t}}{rr'}\sqrt{\frac{k}{4\pi t}}\sum_{m=0}^{n}(-1)^m\left(n+\tfrac{1}{2}, m\right)\left(\frac{t}{krr'}\right)^m \end{aligned} \tag{5}$$

where $h_n^{(2)}$ is again the spherical Hankel function of the second kind of nonnegative integer order n, $i\equiv\sqrt{-1}$ and $(n+\frac{1}{2}, m)\equiv(n+m)!/\{m!(n-m)!\}$. This component of u_n is a "singular" solution of the equation in (3) and gives rise to the prescribed delta function behavior as $t\to 0$.

Balancing term. The second term in the representation (4) is also a polynomial in t modified by $\sqrt{t}$ and multiplied by an exponential. One expression for J_n takes the form

$$\begin{aligned} J_n(r, r'; t) \equiv &-\frac{e^{-k(r+r'-2)^2/4t}}{rr'}\sqrt{\frac{k}{4\pi t}} \\ &\cdot\left\{1+\sum_{m=0}^{n}\sum_{l=0}^{n}\sum_{p=m+l}^{m+l+n}(-1)^{m+l}\frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, p-m-l)}{r^m(r')^l 2^p}\right. \\ &\left.\cdot\sum_{j=n}^{p-2}A_{p-j}(\alpha; t)S_{j+1}(n)\right\}. \end{aligned} \tag{6}$$

Here, for $m \geqq 0$,

$$A_{m+2}(\alpha;\,t) \equiv \sum_{j=0}^{[m/2]} \sum_{l=0}^{j} \frac{(\alpha)^{m+2l-2j}(4t/k)^{j+1-l}}{(2j+2)!\,(m-2j)!} \left\{ \frac{(j+1)!}{l!} - \frac{(m-2j)\Gamma(j+3/2)}{2\Gamma(l+3/2)} \right\}, \tag{7}$$

with $\alpha \equiv r + r' - 2$, while the S_j are constants, depending upon n, which are generated recursively from

$$\begin{aligned} S_{n+1}(n) &= \frac{(-2)^n}{(n+\frac{1}{2},\, n)}, \\ S_{n+p}(n) &= -\sum_{l=n+1}^{n+p-1} \frac{(n+\frac{1}{2},\, l-p)}{(n+\frac{1}{2},\, n)} (-2)^{n+p-l} S_l(n) \end{aligned} \tag{8}$$

($p \geqq 2$). In (7) and elsewhere, $[x]$ designates the greatest integer $\leqq x$. Moreover, throughout this paper we adopt the usual convention of ignoring summations in which the upper index is strictly less than the lower index.

Alternatively, J_n can be expressed in a manner which more clearly exhibits its dependence upon t, namely

$$J_n(r, r';\, t) \equiv \frac{e^{-k(r+r'-2)^2/4t}}{rr'} \sqrt{\frac{k}{4\pi t}} \sum_{m=0}^{n} f_m(n;\, r, r') \left(\frac{t}{k}\right)^m \tag{9}$$

where

$$f_0(n;\, r, r') = -1$$

and

$$f_n(n; r, r') = (-1)^{n+1} \frac{(n+\frac{1}{2},\, n)}{(rr')^n}.$$

In this representation the intermediate f_m $(1 \leqq m \leqq n-1)$ are given recursively by

$$\begin{aligned} &\sum_{m=[(p-n+1)/2]}^{n} f_m(n;\, r, r') \sum_{l=a}^{b} (n+\tfrac{1}{2}, p-l-m)(m+\tfrac{1}{2}, l)(r+r'-2)^{m-l} \\ &\qquad = -\sum_{m=c}^{n} \sum_{l=d}^{e} (-1)^{l+m+p} \frac{(n+\frac{1}{2},\, m)(n+\frac{1}{2},\, l)(n+\frac{1}{2}, p-l-m)}{r^m (r')^l} \end{aligned} \tag{10}$$

with p an integral parameter and a, b, c, d, e integral limits which satisfy

$$\left.\begin{aligned} a &= p-m-n \\ b &= m \\ c &= p-2n \\ d &= p-m-n \\ e &= n \end{aligned}\right\} \text{for } 2n+1 \leqq p \leqq 3n,$$

$$\left.\begin{aligned} a&=0\\ b&=p-m\\ c&=0\\ d&=0\\ e&=p-m \end{aligned}\right\}\quad \text{for } n+1\leqq p\leqq 2n.$$

The term $J_n(r, r'; t)$, like $I_n(r, r'; t)$, is symmetric in r, r' and "balances out" the contribution of I_n when $r=1$, i.e.,

$$J_n(1, r'; t) = -I_n(1, r'; t).$$

Last term. The remaining term in (4) is only present if $n \geqq 1$ in which case the appropriate formula is

$$\begin{aligned} K_n(r, r'; t) &\equiv \tfrac{1}{2}(-i)^n\, e^{-k(r+r'-2)^2/4t}\\ &\cdot \sum_{j=1}^{n} \lambda_j^3 C_j h_n^{(1)}(i\lambda_j) h_n^{(2)}(i\lambda_j r) h_n^{(2)}(i\lambda_j r')\, e^{-\lambda_j(r+r'-1)}\\ &\cdot w\left\{i\left(\lambda_j\sqrt{\frac{t}{k}}+\frac{r+r'-2}{2\sqrt{t/k}}\right)\right\}. \end{aligned} \tag{11}$$

Here $h_n^{(1)}$ and $h_n^{(2)}$ are the spherical Hankel functions of the first and second kinds, respectively, of order n, the λ_j $(j=1, 2, \cdots, n)$ are the n roots of

$$h_n^{(2)}(i\lambda)=0,$$

the C_j $(j=1, 2, \cdots, n)$ are the (unique) solutions of the simultaneous equations

$$\sum_{j=1}^{n} C_j/\lambda_j^m = -\delta_{1m} \qquad (m=1, 2, \cdots, n), \tag{12}$$

formed with these roots and w is the complementary error function as given by

$$w(iz) \equiv e^{z^2}\operatorname{erfc}(z) \equiv \frac{2}{\sqrt{\pi}} e^{z^2}\int_z^\infty e^{-x^2}\,dx \tag{13}$$

(see [1, p. 297]).

In view of the nature of the λ_j, $K_n(1, r'; t)$ vanishes. The three-term summation (4) therefore has the appropriate behavior both for $r=1$ and as $r\to r'$. Hence the presence of K_n may be viewed as ensuring that u_n satisfies the differential equation given in (3).

Component computation. Numerical evaluation of $u_n(r, r'; t)$ for fixed n, k and varying r, r', and t is straightforward for small to moderate values of n. With appropriate organization, the bulk of the computation of I_n and J_n can be performed with integer arithmetic. Evaluation of K_n, however, is more demanding.

The spherical Hankel functions are given by

$$h_n^{(1,2)}(iz) = \frac{i^{\mp(n+1)}}{iz}\, e^{\mp z} \sum_{m=0}^{n} \frac{(n+\frac{1}{2}, m)}{(\pm 2z)^m}, \tag{14}$$

and the zeros λ_j $(j=1, 2, \cdots, n)$ of $h_n^{(2)}(i\lambda)$ therefore lie in the right half of the complex λ-plane (see [1, p. 373]). (The λ_j occur in complex conjugate pairs, except of course, when n is odd, then one of the λ_j is real.) If λ_j is real, the corresponding C_j is real as well. Otherwise the associated C_j also occur in complex conjugate pairs. This behavior, coupled with the nature of the complementary error function, implies the expected reality of K_n.

3. Derivation of the solution $u_n\ (r, r'; t)$. The boundary/initial value problem (3) is ideally suited to application of the Laplace transform. If we define

$$U_n(r, r'; s) \equiv \mathscr{L}(u_n) \equiv \int_0^\infty u_n(r, r'; t)\, e^{-st}\, dt,$$

then U_n is the bounded (in r) solution of the ordinary boundary-value problem

$$(15)\qquad (r^2 U_n')' - (r^2 ks + n(n+1))U_n = -k\delta(r-r'), \qquad U_n(1, r'; s) = 0.$$

The solutions underlying the homogeneous equation in (15) are the spherical Hankel functions of order n and argument $i\sqrt{ks}\,r$, and in view of (14), only $h_n^{(1)}(iz)$ is bounded for large z. The desired solution of (15), therefore, is essentially nothing more than the ordinary Green's function associated with this problem (see [4, p. 239], for example). We thus deduce for $r \leqq r'$

$$(16)\quad U_n(r, r'; s) \equiv \frac{k}{2}\sqrt{ks}\{h_n^{(1)}(i\sqrt{ks}\,r)h_n^{(2)}(i\sqrt{ks}) - h_n^{(2)}(i\sqrt{ks}\,r)h_n^{(1)}(i\sqrt{ks})\}\frac{h_n^{(1)}(i\sqrt{ks}\,r')}{h_n^{(1)}(i\sqrt{ks})},$$

with a symmetric expression valid in the regime $r \geqq r'$. To determine u_n we shall invert U_n. Indeed, the classical representation for u_n given in the Introduction is equivalent to the *formal* inverse Laplace transform of U_n.

The easier inversion. The "singular" component I_n in the representation (4) arises as a result of the following.

THEOREM 1.

$$\mathscr{L}^{-1}\left\{-\frac{k}{2}\sqrt{ks}\,h_n^{(2)}(i\sqrt{ks}\,r)h_n^{(1)}(i\sqrt{ks}\,r')\right\} = I_n(r, r'; t).$$

Proof. We establish the result in the forward direction assuming $r \leqq r'$. From (5)

$$\begin{aligned}\mathscr{L}(I_n(r, r'; t)) &= \mathscr{L}\left\{e^{-k(r^2+r'^2)/4t}(-i)^n\left(\frac{\pi}{4}\right)\left(\frac{k}{\pi t}\right)^{3/2} h_n^{(2)}\left(\frac{ikrr'}{2t}\right)\right\}\\ &= \sqrt{\frac{k}{4\pi}}\sum_{m=0}^{n}(-1)^m\left(n+\frac{1}{2}, m\right)\left(\frac{1}{krr'}\right)^{m+1} k\int_0^\infty t^{m-1/2}e^{-st}\, e^{-k(r-r')^2/4t}\, dt\\ &= -k\sqrt{ks}\sum_{m=0}^{n}(-1)^m\left(n+\frac{1}{2}, m\right)\left(\frac{r'-r}{2rr'\sqrt{ks}}\right)^{m+1} i^m h_m^{(1)}(i(r'-r)\sqrt{ks})\end{aligned}$$

by virtue of [7, 3.471(9), p. 340], and the relationship of the modified Bessel function $K_{m+1/2}(z)$ to $h_m^{(1)}(iz)$. Hence, using (14) and expanding the resulting powers of $r'-r$,

$$\begin{aligned}\mathscr{L}(I_n) &= k\, e^{(r-r')\sqrt{ks}}\sum_{m=0}^{n}\sum_{l=0}^{m}(-1)^m\frac{(n+\frac{1}{2}, m)(m+\frac{1}{2}, l)(r'-r)^{m-l}}{(rr')^{m+1}(2\sqrt{ks})^{m+l+1}}\\ &= k\, e^{(r-r')\sqrt{ks}}\sum_{m=0}^{n}\sum_{l=0}^{m}(-1)^m\frac{(n+\frac{1}{2}, m)(m+\frac{1}{2}, l)}{(2\sqrt{ks})^{m+l+1}}\\ &\quad\cdot\sum_{j=0}^{m-l}\binom{m-l}{j}(-1)^j r^{j-m-1}(r')^{-l-j-1}.\end{aligned}$$

Changing the index j into $m-j$, and then interchanging the last summation with the first two, results in

$$\mathscr{L}(I_n) = ke^{(r-r')\sqrt{ks}} \sum_{j=0}^{n} \sum_{m=j}^{n} \sum_{l=0}^{j} (-1)^j \frac{(n+\frac{1}{2}, m)(m+\frac{1}{2}, l)}{(2\sqrt{ks})^{m+l+1}}$$

$$\cdot \binom{m-l}{m-j} r^{-j-1}(r')^{j-l-m-1},$$

and replacing l with $l-m+j$ this can be rewritten as

$$\mathscr{L}(I_n) = k\, e^{(r-r')\sqrt{ks}} \sum_{j=0}^{n} \sum_{m=j}^{n} \sum_{l=m-j}^{m} (-1)^j \frac{(n+\frac{1}{2}, m)(m+\frac{1}{2}, l-m+j)}{(2\sqrt{ks})^{j+l+1}}$$

$$\cdot \binom{2m-l-j}{m-j} r^{-j-1}(r')^{-l-1}$$

$$= k\, e^{(r-r')\sqrt{ks}} \sum_{j=0}^{n} \sum_{m=0}^{n} \sum_{l=0}^{n} \frac{(-1)^j r^{-j-1}(r')^{-l-1}}{(2\sqrt{ks})^{j+l+1}}$$

$$\cdot \frac{(n+m)!\,(l+j)!}{m!\,(n-m)!\,(l+j-m)!\,(m-j)!\,(m-l)!}.$$

In the last expression we have written out explicitly the various factorials and taken advantage of the fact that the summand vanishes for $0 \leqq m < j$, $0 \leqq l < m-j$ and $m < l \leqq n$. The summation over m only involves factorials and fortuitously can be carried out using a variant of Saalschutz's formula (see [5, p. 66]) to give, using (14),

$$\mathscr{L}(I_n) = k\, e^{(r-r')\sqrt{ks}} \sum_{j=0}^{n} \sum_{l=0}^{n} \frac{(-1)^j r^{-j-1}(r')^{-l-1}}{(2\sqrt{ks})^{j+l+1}}$$

$$\cdot \frac{(n+j)!\,(n+l)!}{j!\,(n-j)!\,l!\,(n-l)!}$$

$$= -\frac{k\sqrt{ks}}{2} \left\{ \frac{e^{\sqrt{ks}\,r}}{i\sqrt{ks}\,r} \sum_{j=0}^{n} \frac{(n+\frac{1}{2}, j)}{(-2\sqrt{ks}\,r)^j} \right\}$$

$$\cdot \left\{ \frac{e^{-\sqrt{ks}\,r'}}{i\sqrt{ks}\,r'} \sum_{l=0}^{n} \frac{(n+\frac{1}{2}, l)}{(2\sqrt{ks}\,r')^l} \right\}$$

$$= -\frac{k\sqrt{ks}}{2} h_n^{(2)}(i\sqrt{ks}\,r)\, h_n^{(1)}(i\sqrt{ks}\,r'). \qquad \square$$

The remaining components. Inversion of the rest of $U_n(r, r'; s)$ as given by (16) is detailed in the lemmas and theorems which follow. The general method is believed to be of wide applicability. Our approach recognizes the transform at hand as a quotient and begins with an inversion of the numerator. The inverse transform of the quotient itself is then shown to satisfy an inhomogeneous integral equation of Volterra type. Solution of this integral equation yields *three* terms, two of which can be recognized as the J_n and K_n of (4). The third term is a multi-summation which, surprisingly, sums to zero for arbitrary n.

LEMMA 1.

$$\frac{k}{2}\sqrt{ks}\, h_n^{(1)}(i\sqrt{ks}\,r)\, h_n^{(1)}(i\sqrt{ks}\,r')\, h_n^{(2)}(i\sqrt{ks}) / h_n^{(1)}(i\sqrt{ks}) = \frac{P}{Q}$$

where

$$P \equiv -k\,e^{-\sqrt{ks}\,(r+r'-2)} \sum_{m=0}^{n} \sum_{l=0}^{n} \sum_{j=0}^{n} (-1)^j \frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, j)}{r^{m+1}(r')^{l+1}} \cdot \left(\frac{1}{2\sqrt{ks}}\right)^{j+l+m+1}$$

and

$$Q \equiv \sum_{m=0}^{n} (n+\tfrac{1}{2}, m)\left(\frac{1}{2\sqrt{ks}}\right)^{m}.$$

Proof. We prove by substitution using (14). □

Lemma 2.

$$\mathscr{L}^{-1}\left(\frac{k\,e^{-\alpha\sqrt{ks}}}{\sqrt{ks}}\right) \equiv I^{(0)}(\alpha;\, t) \equiv \frac{e^{-k\alpha^2/4t}}{\sqrt{\pi t/k}}$$

and for integer $n \geqq 1$

$$\mathscr{L}^{-1}\left(\frac{k\,e^{-\alpha\sqrt{ks}}}{(\sqrt{ks})^{n+1}}\right) \equiv (-1)^n I^{(-n)}(\alpha;\, t)$$

$$\equiv \frac{1}{(n-1)!} \int_{\alpha}^{\infty} \frac{(\beta-\alpha)^{n-1}\, e^{-k\beta^2/4t}}{\sqrt{\pi t/k}}\, d\beta.$$

Proof. Apply iterated integration to [6, 5.6(6), p. 246]. □

Lemma 3. *For integer* $n \geqq 2$, $I^{(-n)}(\alpha;\, t)$ *can be expressed as a sum of polynomials in* α *and* t *multiplied by* $I^{(0)}(\alpha;\, t)$ *and* $I^{(-1)}(\alpha;\, t)$, *respectively. The precise relationship is*

$$I^{(-n)}(\alpha;\, t) = A_n(\alpha;\, t) I^{(0)}(\alpha;\, t) + B_n(\alpha;\, t) I^{(-1)}(\alpha;\, t)$$

where $A_n(\alpha;\, t)$ *is as defined by* (7) *and*

$$B_{n+1}(\alpha;\, t) \equiv \sum_{j=0}^{[n/2]} \frac{\alpha^{n-2j}(t/k)^j}{(n-2j)!\,j!}. \tag{17}$$

Proof. One way to establish this result is to show that A_n, B_n are the solutions of the appropriate initial-value problems. Unfortunately, while it is trivial to verify $dB_n/d\alpha = B_{n-1}$ and the necessary initial values, the demonstration of $dA_n/d\alpha = (k\alpha/2t)A_n + A_{n-1} - B_n$ requires substantial series manipulations and sophisticated knowledge regarding the sums of multi-binomial expressions. It is probably far easier, therefore, to rederive the representation from the definition of $I^{(-n)}$ given in the preceding lemma, by expanding the $(\beta-\alpha)^{n-1}$ factor and integrating by parts. Note that odd powers of β contribute to A_n and even powers to B_n. We leave the details to the interested reader. □

Lemma 4.

$$\mathscr{L}^{-1}(P) = -\sum_{m=0}^{n} \sum_{l=0}^{n} \sum_{j=0}^{n} (-1)^{m+l} \frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, j)}{r^{m+1}(r')^{l+1} 2^{j+l+m+1}} \cdot I^{(-j-l-m)}(\alpha;\, t)$$

where $\alpha \equiv r + r' - 2$.

Proof. The proof is a consequence of Lemmas 1 and 2. □

Lemma 5. *For* $n \geqq 1$, $\phi(\alpha;\, t) \equiv \mathscr{L}^{-1}(P/Q)$ *satisfies the Volterra integral equation*

$$\phi(\alpha;\, t) = \mathscr{L}^{-1}(P) - \int_{\alpha}^{\infty} \sum_{m=1}^{n} \frac{(n+\frac{1}{2}, m)}{2^m} \frac{(\beta-\alpha)^{m-1}}{(m-1)!} \phi(\beta;\, t)\, d\beta. \tag{18}$$

If $n = 0$, $\phi(\alpha;\, t) = \mathscr{L}^{-1}(P)$ *since* $Q \equiv 1$ *in this case.*

Proof. The integral equation (18) is equivalent to the initial-value problem

$$\sum_{j=0}^{n} (-1)^j \frac{(n+\frac{1}{2}, j)}{2^j} \phi^{(n-j)}(\alpha; t) = \frac{d^n}{d\alpha^n}\{\mathscr{L}^{-1}(P)\},$$

$$\lim_{\alpha\to\infty} \phi^{(j)}(\alpha; t) = 0, \qquad j = 0, 1, 2, \cdots, n-1.$$

Since differentiation with respect to the parameter α can be interchanged with transform inversion, the left-hand side of this differential equation becomes for $\phi(\alpha; t) = \mathscr{L}^{-1}(P/Q)$, from Lemmas 1, 2 and 4.

$$\mathscr{L}^{-1}\left\{\frac{1}{Q}\sum_{j=0}^{n} (-1)^j \frac{(n+\frac{1}{2}, j)}{2^j} \frac{d^{n-j}P}{d\alpha^{n-j}}\right\} = \mathscr{L}^{-1}\left\{\frac{P}{Q}\sum_{j=0}^{n} \frac{(n+\frac{1}{2}, j)}{2^j}(-1)^n(\sqrt{ks})^{n-j}\right\}$$

$$= \mathscr{L}^{-1}\{(-1)^n(\sqrt{ks})^n P\}$$

$$= \frac{d^n}{d\alpha^n}\{\mathscr{L}^{-1}(P)\}.$$

The "initial" conditions are likewise a ready consequence of the form of P, Q as given in Lemma 1. □

LEMMA 6. *The resolvent kernel associated with the Volterra integral equation* (18) *when* $n \geqq 1$ *has the representation*

$$R(\alpha, \beta) = \sum_{j=1}^{n} C_j e^{\lambda_j(\alpha-\beta)}$$

where the λ_j $(j = 1, 2, \cdots, n)$ *are the n roots of*

$$h_n^{(2)}(i\lambda) = 0$$

and the C_j $(j = 1, 2, \cdots, n)$ *are the* (*unique*) *solutions of the simultaneous equations*

$$\sum_{j=1}^{n} C_j/\lambda_j^m = -\delta_{1m} \qquad (m = 1, 2, \cdots, n). \tag{12'}$$

Proof. If $L(\alpha, \beta)$ designates the kernel of (18), we merely need to verify the validity of one of the classic Fredholm identities, say,

$$R(\alpha, \beta) = L(\alpha, \beta) + \int_\alpha^\beta L(\alpha, \gamma)R(\gamma, \beta)\, d\gamma$$

(see [3, p. 58], for example). Substituting, we obtain

$$\sum_{j=1}^{n} C_j e^{\lambda_j(\alpha-\beta)}$$

$$= -\sum_{m=1}^{n} \frac{(n+\frac{1}{2}, m)}{2^m} \frac{(\beta-\alpha)^{m-1}}{(m-1)!} - \int_\alpha^\beta \sum_{m=1}^{n} \frac{(n+\frac{1}{2}, m)(\gamma-\alpha)^{m-1}}{2^m(m-1)!} \sum_{j=1}^{n} C_j e^{\lambda_j(\gamma-\beta)}\, d\gamma$$

$$= -\sum_{m=1}^{n} \frac{(n+\frac{1}{2}, m)}{2^m} \frac{(\beta-\alpha)^{m-1}}{(m-1)!}$$

$$+ \sum_{m=1}^{n} \frac{(n+\frac{1}{2}, m)}{2^m} \sum_{j=1}^{n} C_j\left\{\sum_{l=1}^{m} \frac{(\alpha-\beta)^{m-l}}{(m-l)!\,\lambda_j^l} - \frac{e^{\lambda_j(\alpha-\beta)}}{\lambda_j^m}(-1)^m\right\}$$

or, rearranging we find

$$\sum_{j=1}^{n} C_j\, e^{\lambda_j(\alpha-\beta)} \sum_{m=0}^{n} \frac{(n+\frac{1}{2}, m)}{(-2\lambda_j)^m}$$

$$= \sum_{m=1}^{n} \frac{(\beta-\alpha)^{m-1}}{(m-1)!}\left\{\frac{-(n+\frac{1}{2}, m)}{2^m} - \sum_{j=1}^{n} C_j \sum_{p=m}^{n} \frac{(n+\frac{1}{2}, p)(-1)^{(p+m)}}{2^p \lambda_j^{p-m+1}}\right\}.$$

Observe, however, that both sides of this equation vanish owing to (12′), the definition of λ_j and (14). □

LEMMA 7. *If $n \geqq 1$, let $S_l(n) \equiv \sum_{j=1}^{n} C_j/\lambda_j^l$ where the λ_j and the C_j are as defined in Lemma 6. Then for $l > n$ we have*

$$S_{n+1}(n) = \frac{(-2)^n}{(n+\frac{1}{2}, n)}$$

and

$$S_{n+p}(n) = -\sum_{l=n+1}^{n+p-1} \frac{(n+\frac{1}{2}, l-p)}{(n+\frac{1}{2}, n)}(-2)^{n+p-l} S_l(n) \qquad (p \geqq 2). \tag{8′}$$

Proof. By virtue of the definition of λ_j, using (14) we obtain

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \frac{C_j}{(-2\lambda_j)^p}\left[\sum_{m=0}^{n} \frac{(n+\frac{1}{2}, n)}{(-2\lambda_j)^m}\right] \\ &= \sum_{l=p}^{n+p} \left(n+\frac{1}{2}, l-p\right)(-2)^{-l} S_l(n). \end{aligned} \tag{19}$$

Setting $p=1$ we easily determine S_{n+1} since $S_l = -\delta_{1l}$ $(l = 1, 2, \cdots, n)$ by (12′). The recursive expression given above for S_{n+p} with $p \geqq 2$ is likewise a ready consequence of this identity. (Note that if $p > n+1$, the lower summation index can be decreased to $n+1$ since $(n+\frac{1}{2}, l-p) = 0$ for $n+1 \leqq l < p$.) □

LEMMA 8.

$$\int_{\alpha}^{\beta} e^{\lambda(\alpha-\beta)} I^{(-n)}(\beta;\, t)\, d\beta = \sum_{m=1}^{n} I^{(-m)}(\alpha;\, t)\lambda^{m-1-n} + \lambda^{-n}\, e^{(\alpha\lambda+\lambda^2 t/k)} I^{(-1)}(\alpha + 2\lambda t/k;\, t).$$

Proof. We prove by substitution, using the definition of $I^{(-n)}$ from Lemma 2, and integration by parts. The summation term does not appear if $n=0$. □

We come now to the principal results of this subsection.

THEOREM 2. *$\phi(\alpha;\, t) \equiv \mathscr{L}^{-1}(P/Q)$ is given uniquely by*

$$\phi(\alpha;\, t) = J_n(r, r';\, t) + K_n(r, r';\, t) + L_n(r, r';\, t)$$

where $\alpha \equiv r + r' - 2$, J_n and K_n have the representations (6)–(10) *and* (11)–(13), *respectively, and*

$$\begin{aligned} L_n(r, r';\, t) \equiv I^{(-1)}(\alpha;\, t) &\sum_{m=0}^{n} \sum_{l=0}^{n} \sum_{p=m+l}^{m+l+n} (-1)^{m+l+1} \frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, p-m-l)}{r^{m+1}(r')^{l+1} 2^{p+1}} \\ &\cdot \sum_{j=n}^{p-1} B_{p-j}(\alpha;\, t) S_{j+1}(n). \end{aligned} \tag{20}$$

In this last expression the $B_n(\alpha;\, t)$ are given by (17) *for $n \geqq 2$ $(B_1 = 1)$ and the S_j are defined recursively by* (8′).

Proof. Applying fundamental integral equations theory (see [3, pp. 36, 37, 58], for example), Lemmas 5 and 6 imply that $\phi(\alpha; t)$ is given uniquely by

$$\begin{aligned}\phi(\alpha; t) &= \mathscr{L}^{-1}(P(\alpha)) + \int_\alpha^\infty R(\alpha, \beta)\mathscr{L}^{-1}(P(\beta))\, d\beta \\ &= \mathscr{L}^{-1}(P(\alpha)) + \int_\alpha^\infty \sum_{j=1}^n C_j e^{\lambda_j(\alpha-\beta)}\mathscr{L}^{-1}(P(\beta))\, d\beta \\ &= \mathscr{L}^{-1}(P(\alpha)) + \sum_{j=1}^n C_j \int_\alpha^\infty e^{\lambda_j(\alpha-\beta)}\mathscr{L}^{-1}(P(\beta))\, d\beta.\end{aligned}$$

(The interchange of integration and summation here is justified by the exponential decay of the integrand for large β.) Substituting for $\mathscr{L}^{-1}(P)$ from Lemma 4, using the results of Lemmas 8 and 3, and introducing the definition of $S_l(n)$ from Lemma 7, we easily derive the above expression (20) for L_n (if $n \geqq 1$) and the representation (6)-(8) for J_n. Neither L_n nor K_n are present unless $n \geqq 1$, in which case the relation (11) for K_n is a consequence of back-substitution using (14) and replacement of $I^{(-1)}(2z\sqrt{t/k}; t)$ by $-e^{-z^2}w(iz)$ using (13).

The alternative representation (9), (10) for J_n is of a different character. It follows from matching powers of $\sqrt{ks}$ in the Laplace transforms of $\phi(\alpha; t)$ and of $J_n(r, r'; t) + K_n(r, r'; t)$, anticipating the result of Theorem 3. Here it is helpful to recall

$$\mathscr{L}(t^{m-1/2}\, e^{-k\alpha^2/4t}) = -\sqrt{\frac{\pi}{k}}\frac{(k\alpha)^{m+1}}{(-2i\sqrt{ks})^m} h_m^{(1)}(i\alpha\sqrt{ks})$$

(see the proof of Theorem 1) and, in particular,

$$\mathscr{L}(t^{-1/2}\, e^{-k\alpha^2/4t}) = \sqrt{\pi/s}\, e^{-\alpha\sqrt{ks}},$$
$$\mathscr{L}(t^{-3/2}\, e^{-k\alpha^2/4t}) = 2\sqrt{\pi/(k\alpha^2)}\, e^{-\alpha\sqrt{ks}},$$

which together imply

$$\mathscr{L}\left(e^{-k\alpha^2/4t} w\left\{i\left(\lambda\sqrt{\frac{t}{k}} + \frac{\alpha}{2\sqrt{t/k}}\right)\right\}\right) = \frac{\sqrt{k/s}}{\lambda + \sqrt{ks}}\, e^{-\alpha\sqrt{ks}}.$$

The tedious details, which involve several summation interchanges and repeated recognition that $(l+\frac{1}{2}, m)$ vanishes unless $0 \leqq m \leqq l$, are left to the interested reader. □

THEOREM 3. *For* $n \geqq 1$, $L_n(r, r'; t) \equiv 0$.

Proof. The result may be established by showing from (20) that for arbitrary integer q $(q = 0, 1, 2, \cdots)$

$$0 = k^q \frac{d^q\{L_n(r, r'; t)/I^{(-1)}(\alpha; t)\}}{dt^q}\bigg|_{t=0}$$

$$(21) \qquad = \sum_{m=0}^{n}\sum_{l=0}^{n}\sum_{p=m+l}^{m+l+n}\sum_{j=n}^{p-1}(-1)^{m+l+1}\frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, p-m-l)}{r^{m+1}(r')^{l+1}2^{p+1}} \cdot S_{j+1}(n)\frac{\alpha^{p-j-1-2q}}{(p-j-1-2q)!}$$

where $\alpha = r + r' - 2$. The demonstration involves expanding $\alpha^{p-j-1-2q}$ in powers of r and r', collecting all the like powers together, and then verifying that the resulting coefficient vanishes in each instance. Since a considerable number of special cases needs to be examined, the remainder of the proof is relegated to the Appendix. □

Appendix. In order to complete the proof of Theorem 3 we need to show that

$$M_n \equiv \sum_{m=0}^{n} \sum_{l=0}^{n} \sum_{p=m+l}^{m+l+n} \sum_{j=n}^{p-1} \sum_{s=0}^{p-j-1-2q} \sum_{t=0}^{p-j-1-2q-s} I \tag{21'}$$

vanishes for $n \geqq 1$ and arbitrary nonnegative integer q, where

$$I \equiv (-1)^{m+l+1} \frac{(n+\frac{1}{2}, m)(n+\frac{1}{2}, l)(n+\frac{1}{2}, p-m-l)}{r^{m+1}(r')^{l+1} 2^{p+1}}$$

$$\cdot\, S_{j+1} \binom{p-j-1-2q}{s} \binom{p-j-1-2q-s}{t} \frac{r^s (r')^t (-2)^{p-j-1-2q-s-t}}{(p-j-1-2q)!}.$$

Toward this end, we replace the indices s and t with $\alpha \equiv m+1-s$ and $\beta \equiv l+1-t$ and then interchange summations until the α and β summations appear first. The demonstration now devolves to several cases.

CASE 1: $2q+1 \geqq n$.

$a < 2+2q-n$: No terms with such indices appear in M_n.

$2+2q-n \leqq \alpha \leqq n+1$: No terms with such indices appear in M_n unless $3+2q-\alpha \leqq \beta \leqq n+1$ also, in which case the resulting coefficient is

$$C(n, q; \alpha, \beta) = \sum_{m=\alpha-1}^{n} \sum_{l=\beta-1}^{n} \sum_{p=m+l+n-a}^{m+l+n} \sum_{j=n}^{p-m-l+a} I$$

where $a \equiv \alpha+\beta-3-2q$, or, rearranging and shifting indices,

$$C = \frac{2^{a-1}}{r^\alpha (r')^\beta} D(n; \alpha) D(n; \beta) E(n, q; \alpha, \beta)$$

with

$$D(n; \alpha) \equiv \sum_{m=\alpha-1}^{n} \frac{(n+\frac{1}{2}, m)}{(m+1-\alpha)!} \frac{(-1)^m}{2^m}$$

and

$$E(n, q; \alpha, \beta) \equiv \sum_{j=n}^{n+\alpha} \frac{S_{j+1}}{2^j} \left[\sum_{p=0}^{n-j+a} (-1)^{p+1} \frac{(n+\frac{1}{2}, p+j-a)}{p!} \right]$$

$$= \sum_{j=n}^{n+a} \frac{S_{j+1}}{2^j} [(-1)^{n+j} (n+\tfrac{1}{2}, j-a)].$$

If $a > 0$, this last summation is zero by virtue of the identity (19).

For the case $a = 0$ we reason as follows. $D(n; \alpha)$ can be alternatively expressed in terms of the derivative of a product as

$$D(n; \alpha) = \frac{(-2)^{1-\alpha}}{n!} (x^n (1+x)^n)^{(\alpha+n-1)} \Big|_{x=-1/2} \tag{A.1}$$

and this is readily seen to vanish whenever $\alpha+n$ is even, by symmetry. But if $a \equiv \alpha+\beta-3-2q = 0$, α and β must be of opposite parity, one of $\alpha+n$ or $\beta+n$ must be even and the product $D(n; \alpha)D(n; \beta)$ must therefore be zero.

$n+1 < \alpha$: No terms with such indices appear in M_n.

CASE 2: $2q+1 < n$.

$\alpha < 2+2q-n$: No terms with such indices appear in M_n.

$\alpha=2+2q-n$: We find $l=n$, $p=2n+m$, $j=n$, $\beta=n+1$, and hence the relevant coefficient is

$$C(n, q; 2+2q-n, n+1)=\frac{(-1)^{n+1}(n+\frac{1}{2}, n)^2}{r^{2+2q-n}(r')^{n+1}2^{2n+1}}S_{n+1}D(n; 2+2q-n)$$

which vanishes since $\alpha+n=2+2q$ is even (see (A.1) above).

$2+2q-n<\alpha\leqq 2q+2$: Precisely the same situation occurs as in Case 1.

$2q+2<\alpha\leqq n+1$: No terms with such indices appear in M_n unless $3+2q-\alpha\leqq\beta\leqq n+1$. For $\beta\leqq 3+2q-\alpha+n$, the analysis again duplicates that of Case 1. Since $(n+\frac{1}{2}, p-m-l)=0$ for $p<m+l$, rearrangements are possible when $3+2q-\alpha+n<\beta\leqq n+1$ which also make this situation equivalent to Case 1.

$n+1<\alpha$: No terms with such indices appear in M_n.

Acknowledgment. Special thanks are due to Dr. Anthony Klemm for proposing the investigation of which this research is a part.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965; NBS Applied Math Series 55, U.S. Dept. of Commerce, Washington, DC, 1964.

[2] H. S. CARSLAW AND J. C. JAEGER, *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford, 1959.

[3] J. A. COCHRAN, *The Analysis of Linear Integral Equations*, McGraw-Hill, New York, 1972.

[4] ———, *Applied Mathematics: Principles, Techniques, and Applications*, Wadsworth International, Belmont, CA, 1982.

[5] A. ERDÉLYI, et al., *Higher Transcendental Functions, Vol. I*, McGraw-Hill, New York, 1953.

[6] ———, *Tables of Integral Transforms, Vol. I*, McGraw-Hill, New York, 1954.

[7] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series and Products*, Academic Press, New York, 1965.

[8] A. D. KLEMM, *Exact solutions for a quantum hard-core system*, Phys. Lett. A, 110 (1985), pp. 246-248.

[9] S. Y. LARSEN, *Quantum-mechanical pair-correlation function of hard spheres*, J. Chem. Phys., 48 (1968), pp. 1701-1708.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

015

EXISTENCE OF A SOLUTION TO A CERTAIN PLANE PREMIXED FLAME PROBLEM WITH TWO-STEP KINETICS*

C. M. BRAUNER† AND CL. SCHMIDT-LAINÉ†

Abstract. We consider a nonlinear differential system modelling a two-step reaction in a plane premixed flame. The unknowns are two functions u and v (temperature and mass fraction) and a parameter δ associated with the burning rate. We introduce a normalized problem which is first studied on a bounded interval. Upper and lower solutions induce a priori estimates which enable us to pass to the limit of a doubly infinite interval. We obtain the existence of a solution and we provide an explicit value for δ which is related to the L^2-norm of $w=u-v$.

Key words. nonlinear system, unbounded domain, upper/lower solutions, combustion, premixed flame, 2-step kinetics

AMS(MOS) subject classifications. 34A34, 34A45, 34A40, 80A25, 80A30

1. Physical framework. In a recent paper [6], G. Joulin, A. Liñan, G. S. S. Ludford, N. Peters and one of the authors introduced a two-step irreversible reaction for a steady plane flame, with chain-branching/chain-breaking kinetics:

$$A+X\rightarrow 2X, \qquad 2X+M\rightarrow 2P+M. \tag{1.1}$$

Radical X is obtained in the production step, which has a very large activation energy θ and provides product P in the recombination step for which the activation energy is taken to be zero; A is the reactant and M a third body.

This two-step scheme is presented as an alternative to classical single-step kinetics [3] and allows the description of a wider range of phenomena.

The equations are derived in the stretched flame zone, described by the one-dimensional space variable η, $-\infty<\eta<+\infty$, by assuming a fast recombination, i.e., that both production and recombination of radicals take place in the same thin zone. The system reads (see [6, p. 423])

$$u''=q_1 r\delta(u-v)v\,e^{-u}+\delta(u-v)^2, \qquad v''=r\delta(u-v)v\,e^{-u}, \tag{1.2}$$

and the associated boundary conditions are

$$\begin{aligned} &u=-\eta+o(1), \quad v=-\eta+o(1) \quad \text{as } \eta\rightarrow-\infty,\\ &u=o(1), \quad v=o(1) \quad \text{as } \eta\rightarrow+\infty. \end{aligned} \tag{1.3}$$

The unknowns are positive functions u and v, representing temperature and mass fraction of the reactant and a positive constant δ, representing the burning rate. The parameters q_1 and q_2 are the proportions of the total heat released in the first and second steps of the reaction, so that $q_1+q_2=1$. Physical considerations require the recombination step to be exothermic, so that $q_2>0$. Finally r is a positive parameter, corresponding to the ratio of the two reaction rates. The boundary conditions (1.3) are obtained by matching with expansions on either side of the flame sheet [6, p. 422].

* Received by the editors September 23, 1985; accepted for publication (in revised form) June 17, 1986.

† Ecole Centrale de Lyon, Département de Mathématiques, Informatique et Systèmes, BP 163, 69131 Ecully Cedex, France.

The mathematical problem is the following: q_1, q_2 and $r>0$ being given, find two functions $u>0$, $v>0$ and the constant $\delta>0$ satisfying (1.2), (1.3). We refer again to [6] for a numerical treatment of this problem, leading to curves in the (r, δ)-plane.

It is particularly convenient for our study to deal with an equivalent formulation of the system involving the radical mass fraction $w=u-v>0$. It consists of the equations

$$v''=r\delta vw\,e^{-v}\,e^{-w}, \qquad w''=-q_2r\delta vw\,e^{-v}\,e^{-w}+\delta w^2, \tag{1.4}$$

together with the boundary conditions

$$\begin{aligned} &v=-\eta+o(1), \quad w=o(1) \quad \text{as } \eta\to-\infty,\\ &v=o(1), \quad w=o(1) \quad \text{as } \eta\to+\infty. \end{aligned} \tag{1.5}$$

We remark that q_1 is no longer involved. In the sequel the only hypothesis is $q_2>0$, which has a physical basis.

2. Main results. Let us introduce the following problem in the x-variable, which is obtained from (1.4) by taking δ equal to 1:

$$\frac{d^2v}{dx^2}=rvw\,e^{-v}\,e^{-w}, \qquad \frac{d^2w}{dx^2}=-q_2rvw\,e^{-v}\,e^{-w}+w^2. \tag{2.1}$$

We consider this nonlinear system subject to the following boundary conditions:

$$v'\to0 \quad \text{as } x\to+\infty, \qquad w\to0 \quad \text{as } x\to\pm\infty, \tag{2.2}$$

together with the normalization condition

$$v(0)=1. \tag{2.3}$$

In this paper, we will prove the following theorem.

THEOREM 2.1. *For any $r>0$ fixed, there exists a solution $(v, w)\in(C^\infty(\mathbb{R}))^2$ to problem* (2.1)-(2.3) *such that*

$$0<w<M_0=q_2\frac{r}{e}, \tag{2.4}$$

$$\begin{aligned} &v>0, \quad v'<0, \quad v\to+\infty \quad \textit{as } x\to-\infty,\\ &v\to0 \quad \textit{as } x\to+\infty. \end{aligned} \tag{2.5}$$

Moreover, $w\in H^2(\mathbb{R})$ and, as $x\to-\infty$, $v(x)=-l(x-x_0)+$ E.S.T., where $x_0\in\mathbb{R}$ and

$$l=\frac{1}{q_2}\int_{-\infty}^{+\infty} w^2(x)\,dx=\frac{1}{q_2}\,\|w\|^2_{L^2(\mathbb{R})}. \tag{2.6}$$

To return to the initial problem (1.4), (1.5), we make the transformation

$$\eta=l(x-x_0). \tag{2.7}$$

Then (2.1) becomes

$$\begin{aligned} &\frac{d^2v}{d\eta^2}=\frac{1}{l^2}\,rvw\,e^{-v}\,e^{-w},\\ &\frac{d^2w}{d\eta^2}=-\frac{1}{l^2}\,q_2rvw\,e^{-v}\,e^{-w}+\frac{1}{l^2}\,w^2 \end{aligned} \tag{2.8}$$

and, as $\eta \to -\infty$, $v(\eta) = -\eta + \text{E.S.T.}$ Thus it is clear that (1.4), (1.5) is solved by the pair $(v(\eta), w(\eta))$ with

$$\delta = \frac{1}{l^2}. \tag{2.9}$$

Therefore, we have determined

$$\delta = \frac{q_2}{\int_{-\infty}^{+\infty} w^2(\eta)\, d\eta}. \tag{2.10}$$

In the sequel, we give a detailed demonstration of Theorem 2.1. A sketch of the proof is the following: First, we exhibit positive upper and lower solutions of the problem in a formal way. Next, we consider the system (2.1) on a bounded interval $[-a, b]$ with suitable boundary conditions and we prove the existence of a solution (v, w) in a closed convex set K, by a fixed point argument (the convex K involves the upper and lower solutions). We derive a priori estimates in the $\mathscr{C}^1$-norm as well as in the H^1-norm. Finally, we let a and b tend to $+\infty$ and prove the existence of a limit which solves (2.1)–(2.3). Properties (2.5), (2.6) appear as a "spinoff" of the existence proof.

Problem (2.1) is "nearly homogeneous" at $\eta = \pm\infty$, i.e., the exponential terms are effectively constant there. In the reference [8] the stability of the homogeneous problem

$$v'' = r\delta v w, \qquad w'' = -q_2 r \delta v w + \delta w^2$$

is investigated near the equilibrium point $0 \in \mathbb{R}^4$ corresponding to $\eta \to +\infty$. The authors exhibit a critical value $r = 1$, corresponding to a bifurcation point. This value of r has also been pointed out in [6] by means of formal asymptotic expansions.

As an extension to the problem, it is possible to consider, instead of the exponential e^{-v}, some function $\theta(v)$ with algebraic decay as $v \to +\infty$. Then the asymptotic expansion of $v(\eta)$ as $\eta \to -\infty$ will be governed by the behavior of θ. See e.g. [1].

Finally, let us mention that a summary of this work has been presented as a Note aux Comptes Rendus [2]. An alternative proof by a topological shooting method has been announced by S. P. Hastings, C. Lu and Y. H. Wan [5]. The idea of considering the problem in a bounded domain and taking an infinite domain limit is already used in [4] in the case of one-step kinetics (see also [10]).

3. Upper and lower solutions in a formal setting. In this section, we perform a formal analysis of the system (2.1)–(2.3), which is a nonlinear eigenvalue problem: it admits the trivial solution $w = 0$, $v = 1$. In order to prove the existence of a nontrivial solution, a positive lower solution $\underline{w}$ is needed. The main trick is a logarithmic transformation, but it turns out that a large part of the analysis lies on elementary properties of the mapping $t \to t e^{-t}$.

Let us state first the main comparison principles used, which go back to Nagumo. Consider a nonlinear BVP on some interval (α, β), namely: $(*)$ $-z'' + F(x, z) = 0$, $z(\alpha) = z(\beta) = 0$, where F is continuous (say). An upper solution $\bar{z}$ to $(*)$ is a C^2 function satisfying: $-\bar{z}'' + F(x, \bar{z}) \geqq 0$, $\bar{z}(\alpha) \geqq 0$, $\bar{z}(\beta) \geqq 0$. A lower solution $\underline{z}$ satisfies the opposite inequalities. If $\underline{z} \leqq \bar{z}$, then the problem $(*)$ has a solution z such that $\underline{z} \leqq z \leqq \bar{z}$ on $[\alpha, \beta]$. For more general statements, see e.g. [9].

Part 1. Assume that (2.1)–(2.3) has a smooth positive solution (v, w). We remark that $v e^{-v} e^{-w}$ is bounded by $1/e$ so on $(0, +\infty)$ a lower solution to (2.1a) is

$$\underline{v}'' = \frac{r}{e} \underline{v}, \quad \underline{v}(0) = 1, \quad \underline{v}' \to 0 \quad \text{as } x \to +\infty. \tag{3.1}$$

Since $v''>0$, $v'<0$, and $v<1$ for $x>0$, $v>1$ for $x<0$. So we define

$$\underline{v}(x)=e^{-cx},\quad c=\sqrt{r/e}\quad \text{for } x>0,$$
$$\underline{v}(x)=1\quad \text{for } x\leqq 0.^1 \tag{3.2}$$

Integrating (2.1a) backwards from $+\infty$ yields

$$v'^2(x)=2r\int_{+\infty}^{x} w\,e^{-w}v\,e^{-v}v'\,ds\leqq\frac{2r}{e}\int_{x}^{+\infty}-vv'\,e^{-v}\,ds$$
$$=\frac{2r}{e}[(v+1)\,e^{-v}]_0^\infty<\frac{2r}{e},$$

so

$$0<-v'(x)<M_1=\sqrt{2r/e}. \tag{3.3}$$

We define

$$\bar{v}(x)=-M_1x+1\quad \text{for } x<0,$$
$$\bar{v}(x)=1\quad \text{for } x\geqq 0.^1 \tag{3.4}$$

Part 2. We look for an upper solution $\bar{w}$ of the form

$$\bar{w}=M_0=Cst>0. \tag{3.5}$$

Replacing w by M_0 and using (2.1b), we obtain

$$M_0^2-q_2rv\,e^{-v}M_0\,e^{-M_0}>M_0^2-\frac{q_2r}{e}M_0,$$

so we may choose

$$M_0=\frac{q_2r}{e}. \tag{3.6}$$

Remark 3.1. The choice of $\bar{w}$ may be improved by taking the root of

$$M_0\,e^{-y}=y. \tag{3.7}$$

Part 3. Finally, we look for a lower solution of (2.1b), i.e., a C^2 function satisfying

$$\underline{w}<\bar{w},$$
$$-\underline{w}''+w^2-q_2rv\,e^{-v}\underline{w}\,e^{-\underline{w}}\leqq 0, \tag{3.8}$$
$$\underline{w}\to 0\quad \text{as } x\to\pm\infty.$$

Minorizing $v\,e^{-v}$ by $\underline{v}\,e^{-\bar{v}}$, we obtain

$$v\,e^{-v}\geqq\frac{1}{e}e^{-cx}\quad \text{if } x>0,\qquad v\,e^{-v}\geqq\frac{1}{e}e^{M_1x}\quad \text{if } x<0. \tag{3.9}$$

When we set $\rho=\max(c,M_1)=M_1$, (3.9) yields

$$v\,e^{-v}\geqq\frac{1}{e}e^{-\rho|x|},\qquad -\infty<x<+\infty. \tag{3.10}$$

So we may replace (3.8) by

$$\underline{w}<\bar{w},$$
$$-\underline{w}''+\underline{w}^2-M_0\,e^{-M_0}\,e^{-\rho|x|}\underline{w}\leqq 0, \tag{3.11}$$
$$\underline{w}\to 0\quad \text{as } x\to\pm\infty.$$

To avoid the trivial lower solution $\underline{w}=0$, we introduce the logarithmic transformation

$$f=-\ln \underline{w}. \tag{3.12}$$

[1] Note that $\underline{v}$ and $\bar{v}$ are genuine lower and upper solutions to (2.1) only on $(0,+\infty)$, i.e., they verify: $\underline{v}''\geqq r\underline{v}w\,e^{-v}\,e^{-w}$ and $\bar{v}''\leqq r\bar{v}w\,e^{-\bar{v}}\,e^{-w}$. On $(-\infty,0)$ it holds that $\underline{v}\leqq v\leqq\bar{v}$. Therefore we do not need $\underline{v}$ and $\bar{v}$ to be smooth at the origin.

Therefore the equivalent is to find $f>0$ such that

$$e^{-f}<\min(1, M_0),$$

$$f''-f'^2+e^{-f}\leqq M_0\, e^{-M_0}\, e^{-\rho|x|}, \tag{3.13}$$

$$f\to+\infty \quad \text{as } x\to\pm\infty.$$

LEMMA 3.1. *A convenient choice of f is*

$$f=\frac{1}{2\rho}M_0\,e^{-M_0}\left(\frac{1}{\rho}e^{-\rho|x|}+|x|+\gamma\right),\qquad \gamma=Cst. \tag{3.14}$$

Proof. Remark that f is even and C^2. Since

$$f''-f'^2-M_0\,e^{-M_0}\,e^{-\rho|x|}=-f'^2-\tfrac{1}{2}M_0\,e^{-M_0}\,e^{-\rho|x|}<0$$

and does not vanish at $\pm\infty$, it is obviously possible to choose the constant γ such that the maximum of e^{-f} is small enough to satisfy simultaneously (3.13a and b). □

4. The problem on a bounded interval. Let a and b be two positive real numbers. We consider the system (2.1) on the interval $(-a, b)$, viz.,

$$-v''+rv\,e^{-v}w\,e^{-w}=0,\qquad -w''+w^2=q_2rv\,e^{-v}w\,e^{-w}, \tag{4.1}$$

with the following conditions:

$$\begin{aligned} &v(0)=1, && v'(b)=0,\\ &w(-a)=\underline{w}(-a), && w(b)=\underline{w}(b). \end{aligned} \tag{4.2}$$

We formulate it as a fixed point problem:

$$(v, w)=T[(v, w)] \tag{4.3}$$

in the following way: consider the space $E=\{\mathscr{C}^1([-a, b])\}^2$ and the closed convex set $K\subset E$:

$$K=\{(v, w)\in E,\ \underline{w}\leqq w\leqq \bar{w},\ \underline{v}\leqq v\leqq \bar{v},\ 0\leqq -v'\leqq M_1,\ |w'|\leqq m\} \tag{4.4}$$

where the functions $\bar{v}$, $\bar{w}$ and $\underline{v}$, $\underline{w}$, as well as the constant M_1, have been defined in the preceding section; they do not depend on a and b. Here only the constant m depend on a and b.

For $(v, w)\in K$, we define the operator T by

$$(V, W)=T[(v, w)]. \tag{4.5}$$

First V is solution of the nonlinear problem

$$\begin{aligned} &-V''+rw\,e^{-w}V\,e^{-V}=0,\\ &V(0)=1,\qquad V'(b)=0. \end{aligned} \tag{4.6}$$

LEMMA 4.1. *V is the unique solution of the* BVP (4.6) *on* $[0, b]$. *Moreover,* $\underline{v}\leqq V\leqq 1$.

Proof. $\underline{v}$ defined by (3.1) is clearly a lower solution to (4.6), while 1 is an upper solution. So (4.6) admits a solution such that $\underline{v}\leqq V\leqq 1$. The uniqueness of V is guaranteed by the monotonicity of the mapping $t\to t\,e^{-t}$ for $0\leqq t\leqq 1$. □

LEMMA 4.2. *On* $[-a, 0]$, *V is the unique solution of the nonlinear* IVP (4.6a) *with*

$$V(0)=1,\qquad V'(0^-)=V'(0^+). \tag{4.7}$$

Proof. The IVP (4.6a), (4.7) has a local solution V in some interval $\xi \leqq x \leqq 0$ which can be extended to $[-a, 0]$, for $1 \leqq V \leqq e^{-cx}$. □

Next W is given by the nonlinear BVP

$$\begin{aligned} &-W''+W^2=q_2rv\,e^{-v}w\,e^{-w},\\ &W(-a)=\underline{w}(-a), \qquad W(b)=\underline{w}(b). \end{aligned} \tag{4.8}$$

LEMMA 4.3. *W is the unique solution of* (4.8) *on* $[-a, b]$. *Moreover* $\underline{w} \leqq W \leqq \bar{w} = M_0$.

Proof. (i) Let us check that M_0 is an upper solution

$$M_0^2-q_2rv\,e^{-v}w\,e^{-w} \geqq M_0^2-\frac{q_2r}{e}w \geqq 0 \quad \text{for } w \leqq M_0.$$

(ii) Now we verify that $\underline{w}$ is a lower solution:

$$-\underline{w}''+\underline{w}^2-q_2rv\,e^{-v}w\,e^{-w} \leqq -\underline{w}''+\underline{w}^2-\frac{q_2r}{e}\,e^{-\rho|x|}\underline{w}\,e^{-M_0} \leqq 0$$

after (3.10), (3.11) and $w \geqq \underline{w}$.

(iii) So (4.8) admits a solution such that $\underline{w} \leqq W \leqq \bar{w}$. The uniqueness follows from the monotonicity of the mapping $t \to t^2$ for $t \geqq 0$. □

LEMMA 4.4. *T is a continuous compact mapping from K into itself.*

Proof. (i) T maps K into itself. We already know $V \geqq \underline{v}$, $\underline{w} \leqq W \leqq \bar{w}$. Since $V''>0$, $V' \leqq 0$. From (4.6) we find for $-a \leqq x < b$:

$$\begin{aligned} V'^2(x) &= 2r\int_b^x w\,e^{-w}\,V e^{-V}\,V'\,ds \\ &\leqq \frac{2r}{e}\int_x^b -VV'\,e^{-V}\,ds \\ &\leqq \frac{2r}{e} = M_1^2, \end{aligned}$$

and clearly $V(x) \leqq \bar{v}(x) = -M_1x+1$ on $[-a, 0]$.

Finally, let us consider W'. Let x_0 be a point where $W'(x_0)=0$. Writing $W'(x) = \int_{x_0}^x W''(s)\,ds$, we find $m=2(a+b)M_0^2$.

(ii) T is a continuous and compact mapping. Whenever $(v, w) \in K$, V'' and W'' are clearly bounded on $[-a, b]$, which induces the boundedness of Range (T) in $(C^2[-a, b])^2$. Therefore Range (T) is relatively compact in E.

To show that T is continuous, take a sequence $(v_n, w_n) \in K$, converging to (v, w) in E; let $(V_n, W_n) = T[v_n, w_n]$. Since Range (T) is relatively compact in K, there exists a subsequence (V_{n_p}, W_{n_p}) converging to $(V, W) \in K$, in E. We shall show that $(V, W) = T[(V, W)]$: let ϕ be a smooth function with compact support in $(-a, b)$; then (4.6)-(4.8) yield

$$\begin{aligned} &\int_{-a}^b V'_{n_p}\phi'\,dx + r\int_{-a}^b w_{n_p}\,e^{-V_{n_p}}\,e^{-w_{n_p}}v_{n_p}\phi\,dx = 0,\\ &\int_{-a}^b W'_{n_p}\phi'\,dx + \int_{-a}^b W_{n_p}^2\phi\,dx = q_2r\int_{-a}^b v_{n_p}w_{n_p}\,e^{-v_{n_p}}\,e^{-w_{n_p}}\phi\,dx, \end{aligned} \tag{4.9}$$

and the convergence of the above integrals is straightforward. Finally $T[(v_n, w_n)] \to T[(v, w)]$ by contradiction. □

THEOREM 4.1. *Problem* (4.1), (4.2) *has a solution* (v, w) *belonging to* $K \cap (C^\infty[-a, b])^2$.

Proof. From Lemma 4.4, T is a continuous and compact mapping from K into K and Schauder's Fixed-Point Theorem implies the existence of a fixed point of T in K. The smoothness of v and w is straightforward. □

We may establish further estimates independently of a and b.

THEOREM 4.2. *There exist positive constants* M_2, M_3 *and* M_4, *independent of a and b, such that*

$$(4.10)\qquad \begin{aligned} &|w'| \leqq M_2,\\ &\int_{-a}^{b} w^2\,dx \leqq M_3, \qquad \int_{-a}^{b} w'^2\,dx \leqq M_4. \end{aligned}$$

Proof. (i) Let us estimate w'; let x_0 be a point where $w'(x_0)=0$; an integration of (4.1b) leads to

$$\begin{aligned} \frac{w'^2(x)}{2} &= q_2 r[v\,e^{-v}\,e^{-w}(w+1)]_{x_0}^{x} - q_2 r\int_{x_0}^{x} e^{-w}(w+1)v'\,e^{-v}\,ds \\ &\quad + q_2 r\int_{x_0}^{x} e^{-w}(w+1)v'v\,e^{-v}\,ds + \left[\frac{w^3}{3}\right]_{x_0}^{x} \\ &\leqq 2q_2 r \sup\{v\,e^{-v}\,e^{-w}(w+1)\} + q_2 r\sup\{e^{-w}(w+1)\}\int_{-a}^{b} -v'\,e^{-v}\,ds \\ &\quad + q_2 r\sup\{e^{-w}(w+1)\}\int_{-a}^{b} -vv'\,e^{-v}\,ds + \frac{2}{3}M_0^3. \end{aligned}$$

Hence

$$(4.11)\qquad M_2 = 2\left[q_2 r\left(1+\frac{1}{e}\right) + \frac{M_0^3}{3}\right]^{1/2}.$$

(ii) We are now going to estimate $\int_{-a}^{b} w^2\,dx$. First, we combine (4.1) to get

$$(4.12)\qquad w'' + q_2 v'' = w^2.$$

By integrating (4.12) and making use of the previous estimates, we obtain

$$(4.13)\qquad M_3 = 2M_2 + 2q_2 M_1.$$

(iii) Finally, we multiply (4.1b) by w and integrate on $(-a, b)$ to obtain

$$\begin{aligned} \int_{-a}^{b} w'^2\,dx &= q_2 r\int_{-a}^{b} v\,e^{-v}\,e^{-w}w^2\,dx - \int_{-a}^{b} w^3\,dx + [ww']_{-a}^{b} \\ &\leqq q_2 r\frac{1}{e}\int_{-a}^{b} w^2\,dx + M_0\int_{-a}^{b} w^2\,dx + 2M_0M_2. \end{aligned}$$

Therefore

$$(4.14)\qquad M_4 = 2M_0(M_2 + M_3).$$

5. Passage to the limit. In this section, we will let a and b tend to infinity. Inasmuch as our proof may have some numerical meaning and since the behavior of v near ±∞ is substantially different, it is interesting to let a and b tend independently to infinity, although it is obviously possible to take $a = b$.

Let a_n and b_n be two sequences tending to infinity, as $n \to +\infty$. We consider a sequence of solution to (4.1), (4.2) on the interval $[-a_n, b_n]$, denoted by (v_n, w_n), satisfying

$$v_n'' = r v_n w_n e^{-v_n} e^{-w_n}, \qquad w_n'' = -q_2 r v_n w_n e^{-v_n} e^{-w_n} + w_n^2, \tag{5.1}$$

with

$$\begin{aligned} &v_n(0) = 1, \qquad v_n'(b_n) = 0, \\ &w_n(-a_n) = \underline{w}(a_n), \qquad w_n(b_n) = \underline{w}(b_n). \end{aligned} \tag{5.2}$$

Next we extend (v_n, w_n) to $\mathbb{R}$ by

$$\begin{aligned} \tilde{v}_n(x) &= v_n(x) \quad \text{if } x \in [-a_n, b_n] \\ &= v_n(-a_n) \quad \text{if } x < -a_n \\ &= v_n(b_n) \quad \text{if } x > b_n, \\ \tilde{w}_n(x) &= w_n(x) \quad \text{if } x \in [-a_n, b_n] \\ &= \underline{w}(x) \quad \text{if not.} \end{aligned} \tag{5.3}$$

Note that $\underline{w}$ given by (3.12), (3.14) is obviously in $H^1(\mathbb{R})$. With Theorems 4.1 and 4.2, we have the following estimates on sequences $\tilde{v}_n$ and $\tilde{w}_n$:

$$\begin{aligned} &\underline{w} \leqq \tilde{w}_n \leqq \bar{w} = M_0, \qquad |\tilde{w}_n'| \leqq M_2 + \|\underline{w}'\|_{L^\infty(\mathbb{R})}, \\ &\|\tilde{w}_n\|^2_{H^1(\mathbb{R})} \leqq M_3 + M_4 + \|\underline{w}\|^2_{H^1(\mathbb{R})}, \\ &0 \leqq -\tilde{v}_n' \leqq M_1, \\ &\underline{v}(x) \leqq \tilde{v}_n(x) \leqq \bar{v}(X), \qquad -\infty < -X \leqq x < +\infty, \\ &\int_{-X}^{+\infty} |v_n'(x)|\, dx \leqq 1 + \bar{v}(X). \end{aligned} \tag{5.4}$$

As a consequence of these estimates, we have the following.

LEMMA 5.1. *There exists a subsequence (v_{n_k}, w_{n_k}) and a pair (v, w) such that, as $n_k \to +\infty$,*

$$\tilde{v}_{n_k} \to v, \ \tilde{w}_{n_k} \to w \quad \textit{uniformly on compact subsets of } \mathbb{R}. \tag{5.5}$$

Moreover, $\tilde{w}_{n_k} \to w$ in $H^1(\mathbb{R})$ weakly, $\tilde{v}_{n_k}' \to v'$ in $L^p(-X, +\infty)$ weakly for all $X > 0$, $p > 1$.

Proof. Let $X_0 > 0$ be fixed. The sequences $\tilde{v}_n$ and $\tilde{w}_n$ are bounded in $W^{1,\infty}([-X_0, X_0])$ by (5.4), so one can extract two subsequences $\tilde{v}_n^{X_0}$ and $\tilde{w}_n^{X_0}$ converging uniformly on $[-X_0, X_0]$. By means of a diagonal process one obtains two subsequences which converge uniformly on $[-X, X]$, for all $X > 0$.

On the other hand, (5.4) yields that $\tilde{w}_n$ is bounded in $H^1(\mathbb{R})$ and the sequence $\tilde{v}_n'$ is bounded in $L^p(-X, +\infty)$, $\forall X > 0, p \geqq 1$, hence the lemma. □

Now, let ϕ be a smooth test function, supp $\phi \subset (-X, X)$, $X > 0$ fixed, and let n_k be large enough so that $a_{n_k} > X$, $b_{n_k} > X$. It follows that

$$\int_{-X}^{X} v_{n_k}' \phi' \, dx = -r \int_{-X}^{X} v_{n_k} w_{n_k} e^{-v_{n_k}} e^{-w_{n_k}} \phi \, dx$$

and

$$\int_{-X}^{X} w_{n_k}' \phi' \, dx = +q_2 r \int_{-X}^{X} v_{n_k} w_{n_k} e^{-v_{n_k}} e^{-w_{n_k}} \phi \, dx - \int_{-X}^{X} w_{n_k}^2 \phi \, dx.$$

In the limit $n_k \to +\infty$, at least in the sense of distributions on $\mathbb{R}$, we find that (v, w) is a solution of[2]

$$(5.6) \qquad v'' = rvw\, e^{-v}\, e^{-w}, \qquad w'' = -q_2 rvw\, e^{-v}\, e^{-w} + w^2.$$

By bootstrapping the solution (v, w) is in fact in $(C^\infty(\mathbb{R}))^2$. We have $v(0) = 1$ and since, $w \in H^1(\mathbb{R})$,

$$(5.7) \qquad w(x) \to 0 \quad \text{as } x \to \pm\infty.$$

Also the upper and lower solutions are preserved at the limit, in fact

$$(5.8) \qquad \underline{v} \leqq v \leqq \bar{v}, \qquad \underline{w} \leqq w \leqq \bar{w}.$$

In particular, it follows that $v > 0$, $w > 0$.

6. Asymptotic properties. We have already proved the existence of a positive solution (v, w) to (5.6), i.e., (2.1). Some of the boundary conditions have been satisfied. We now complete the proof of Theorem 2.1 by a couple of lemmas.

LEMMA 6.1. *$v' < 0$; $v(x)$ and $v'(x) \to 0$ as $x \to +\infty$.*

Proof. Through the limit process, we know $v' \leqq 0$, $v' \in L^p(-X, +\infty)$, for all $X > 0$, $1 < p \leqq +\infty$. Since $vw > 0$, we have $v'' > 0$, hence $v' < 0$ and $v'(x)$ has a limit as $x \to +\infty$, which is necessarily 0. Next, $v(x)$ has a limit v_l as $x \to +\infty$ (in particular it yields $v' \in L^1(-X, +\infty)$). Assume that $v_l > 0$. Locally the behavior of w is given by the dynamical system

$$(6.1) \qquad \dot{y}_1 = y_2, \qquad \dot{y}_2 = y_1(y_1 - q_2 r\, e^{-y_1} v_l\, e^{-v_l}),$$

whose equilibrium point $(0, 0)$ is a center. Therefore w cannot tend to 0 without oscillations, a contradiction. Hence $v_l = 0$ and the degenerate equilibrium point $(0, 0)$ has a one-dimensional stable manifold in the range $y_1 > 0$, $y_2 < 0$. For a more complete stability analysis, we refer to [7], [8]. □

LEMMA 6.2. *$w \in H^2(\mathbb{R})$.*

Proof. We already know that $w \in H^1(\mathbb{R})$. From (5.6b), $|w''| \leqq 2M_0 w$; hence $w'' \in L^2(\mathbb{R})$. □

Remark 6.1. Further results on regularity can be easily obtained in higher-order Sobolev spaces. We conjecture that $w \in W^{2,1}(\mathbb{R})$ from the formal asymptotic developments of [6].

LEMMA 6.3. *As $x \to -\infty$,*

$$(6.2) \qquad v'(x) \to -l$$

with

$$(6.3) \qquad l = \frac{1}{q^2} \int_{-\infty}^{+\infty} w^2(x)\, dx$$

and there exists $x_0 \in \mathbb{R}$, such that

$$(6.4) \qquad v(x) = -l(x - x_0) + \text{E.S.T.}$$

Proof. Since $v' < 0$ and bounded from below, we may define $l = \lim_{x \to -\infty} (-v'(x))$. Let us combine the system again as in (4.12)

$$(6.5) \qquad w'' + q_2 v'' = w^2.$$

[2] As an alternative proof, it is easy to build subsequences such that the first and second derivatives converge uniformly on compact subsets of $\mathbb{R}$.

Integration of (6.5) on $\mathbb{R}$ yields, with Lemmas 6.2 and 6.3,

$$q_2 l = \int_{\mathbb{R}} w^2(x)\, dx.$$

Therefore, $v \to +\infty$ as $x \to -\infty$ and

$$v'' = rvw\, e^{-v}\, e^{-w} \xrightarrow[x \to -\infty]{} 0 \quad \text{exponentially.} \tag{6.6}$$

Relation (6.6) may clearly be integrated twice to get (6.4). x_0 is an integration constant. So Theorem 2.1 is proved. □

Acknowledgments. We are indebted to F. Conrad, G. Bayada and M. Chambat for their comments on the preliminary draft. We thank G. S. S. Ludford for his careful reading of the manuscript.

REFERENCES

[1] C. M. Brauner, W. Eckhaus, M. Garbey and A. Van Harten, *Asymptotics of a rather unusual type in a free surface problem*, this Journal, 18 (1987), pp. 812-841.

[2] C. M. Brauner and Cl. Schmidt-Lainé, *Existence d'une solution pour le problème de la flamme plane prémélangée avec cinétique à deux pas*, C.R. Acad. Sci., Paris Sér. I, 301 (1985), pp. 667-670.

[3] J. D. Buckmaster and G. S. S. Ludford, *Theory of Laminar Flames*, Cambridge Univ. Press, Cambridge, 1982.

[4] H. Berestycki, B. Nicolaenko and B. Scheurer, *Traveling wave solutions to combustion models and their singular limits*, this Journal, 16 (1985), pp.1207-1242.

[5] S. P. Hastings, C. Lu and Y. H. Wan, *A three dimensional shooting method as applied to a problem in combustion theory*, Phys. D, to appear.

[6] G. Joulin, A. Liñan, G. S. S. Ludford, N. Peters and Cl. Schmidt-Lainé, *Flames with chain branching/chain breaking kinetics*, SIAM J. Appl. Math., 45 (1985), pp. 420-434.

[7] Cl. Schmidt-Lainé, *Sur quelques problèmes non linéaires en mécanique des fluides, chimie et combustion*, Thèse d'état, Univ. Lyon I, 1985.

[8] Cl. Schmidt-Lainé and D. Serre, *Etude de stabilité d'un système non linéaire de dimension* 4 *en combustion et généralisation à une classe de problèmes homogènes de degré* 2, Physica, 21D (1986), pp. 42-60. See also C. R. Acad. Sci, Paris Sér I, 303 (1986), pp. 551-554.

[9] K. Schmitt, *Boundary value problems for nonlinear second order differential equations*, Monatsh. Math., 72 (1968), pp. 347-354.

[10] J. F. Toland, *Positive solutions of nonlinear elliptic equations. Existence and nonexistence of solutions with radial symmetry in $L_p(\mathbb{R}^n)$*, Trans. Amer. Math. Soc., 282 (1984), pp. 335-354.

SIAM J. Math. Anal.
Vol. 18, No. 5, September 1987

016

SHAPE SENSITIVITY ANALYSIS OF UNILATERAL PROBLEMS*

JAN SOKOŁOWSKI† AND JEAN-PAUL ZOLESIO‡

Abstract. After a short introduction we turn to the sensitivity analysis of an abstract variational inequality and we apply this result to the sensitivity analysis of an obstacle problem: Ω_t is a domain built by a speed vector field V, $K(\Omega_t)$ a closed convex subset of the Sobolev space $H^1(\Omega_t)$, z_t is the element of $K(\Omega_t)$ solution of an elliptic variational inequality. We characterize the shape derivative $z' = \dot{z} - \nabla z . V(0)$ (where $\dot{z}$ is the material derivative, $\dot{z} = ((d/dt)z_t \circ T_t)_{t=0}$) as the solution of another variational inequality well posed on a new convex subset $S_V(\Omega)$ of $H^1(\Omega)$ depending on the speed vector field V only by boundary expression (see Theorem 2). Finally we characterize the material derivative $\dot{\mathbf{u}}$ of the solution of the Signorini variational inequality associated with planar linear elasticity. For this purpose the material derivative has been generalized to vectors situation by $\dot{\mathbf{u}} = d/dt\ (DT_t^{-1}.\mathbf{u}_t \circ T_t)_{t=0}$. This material derivative is the solution of a variational inequality posed on a convex subset $S(\Omega)$ of $H^1(\Omega_t)^2$.

Key words. shape optimization, variational inequality

AMS(MOS) subject classifications. 35R35, 49A29, 49A52

1. Introduction. This paper is devoted to sensitivity analysis of unilateral boundary value problems with respect to the perturbations of the domain of integration.

Let us consider an example. Given a domain $\Omega \subset R^2$, let us consider a family of continuously differentiable, one-to-one mappings

$$T_t : R^2 \to R^2.$$

Ω_t is then the transformed domain; see Fig. 1. For the details we refer the reader to § 3. Here $t \in [0, \delta)$ is a real parameter. Let us consider the following obstacle problem defined in $\Omega_t = T_t(\Omega)$ for a fixed parameter $t \in [0, \delta)$. Find $Y_t \in H_0^1(\Omega_t)$ such that:

$$(1)\quad \begin{aligned} &-\Delta Y_t(x) - f(x) \leqq 0 \text{ in } \Omega_t, \\ &Y_t(x) - 1 \leqq 0 \quad \text{in } \Omega_t, \\ &(Y_t(x) - 1)(\Delta Y_t(x) + f(x)) = 0 \quad \text{in } \Omega_t \end{aligned}$$

where $f(\cdot) \in L^2(R^2)$ is a given element. It is well known [1] that there exists a unique solution to the obstacle problem.

We denote here by $E(Y_t)$ the coincidence set, by γ_t the free boundary, i.e. (see Fig. 2),

$$E(Y_t) = \{x \in \Omega_t \mid Y_t(x) = 1\}.$$

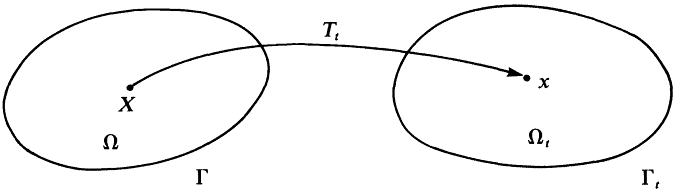


Fig. 1

* Received by the editors June 10, 1985; accepted for publication April 1, 1986.

† Systems Research Institute of Polish Academy of Sciences, 01-447 Warszawa, ul. Newelska 6, Poland.

‡ Centre National de la Recherche Scientifique, Département de physique Mathématique, Faculté des Sciences, Place E. Bataillon, 34060 Montpellier, France and E.N.S.M.P., Sophia Antipolis, Valbonne 06 565, France.

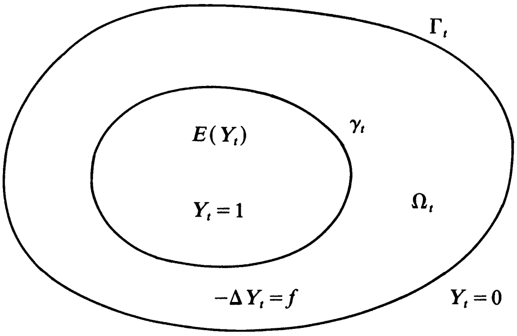


FIG. 2

Let us denote $\mu_t=-f|_{E(Y_t)}$. Since $Y_t(\cdot)$ is sufficiently smooth [1], it folllows that the element $Y_t(\cdot)\in H_0^1(\Omega_t)$ satisfies the equation

$$-\Delta Y_t(x)=f(x)+\mu_t(x) \quad \text{in } \Omega_t.$$

The obstacle problem (1) can be transported to the domain $\Omega_0\equiv\Omega$ using transformation T_t, i.e. the function $Y^t(X)\overset{\text{def}}{=}(Y_t\circ T_t)(X)$ is given by the following obstacle problem defined in the fixed domain Ω. Find $Y^t\in H_0^1(\Omega)$ such that:

$$\begin{aligned}
&-\operatorname{div}(A(t,X)\operatorname{grad} Y^t(X))-f^t(X)\leqq 0 \quad \text{in } \Omega,\\
(2)\qquad &Y^t(X)-1\leqq 0 \quad \text{in } \Omega,\\
&(Y^t(X)-1)(-\operatorname{div}(A(t,X)\operatorname{grad} Y^t(X))-f^t(X))=0 \quad \text{in } \Omega.
\end{aligned}$$

Here we denote

$$\begin{aligned}
f^t(X)&=\det(DT_t(X))(f\circ T_t)(X),\\
A(t,X)&=\det(DT_t(X))(DT_t(X))^{-1}\cdot{}^*(DT_t(X))^{-1},
\end{aligned}$$

where $DT_t(X)$ is the Jacobian matrix of transformation $T_t: R^2\to R^2$ evaluated at $X\in\Omega$.

We prove that there exists the limit

$$\overset{\circ}{Y}=\lim_{t\downarrow 0}(Y^t-Y^0)/t$$

in $H_0^1(\Omega)$. The element $\overset{\circ}{Y}\in H_0^1(\Omega)$ is given by the following variational problem (see Fig. 3). Find $\overset{\circ}{Y}\in H^1(\Omega)$ that minimizes:

$$\begin{aligned}
&I(\eta)=\frac{1}{2}\int_\Omega(\nabla\eta)^2\,dx+\int_\Omega\langle A'\cdot\nabla Y^0,\nabla\eta\rangle_{R^2}\,dx-\int_\Omega f'\eta\,dx\\
(3)\qquad &\text{subject to constraints}\\
&\eta(X)=0 \quad \text{for q.e. } X\in\operatorname{supp}\mu_0,\\
&\eta(X)\leqq 0 \quad \text{for q.e. } X\in E(Y_0)\backslash\operatorname{supp}\mu_0.
\end{aligned}$$

Here

$$\begin{aligned}
A'(X)&=\lim_{t\downarrow 0}(A(t,X)-A(0,X))/t,\\
f'(X)&=\lim_{t\downarrow 0}(f^t(X)-f^0(X))/t.
\end{aligned}$$

The element $\overset{\circ}{Y}$ is the so-called material derivative of the solution to the obstacle problem (1) in the direction of a transformations $\{T_t\}$. Using this result we prove that

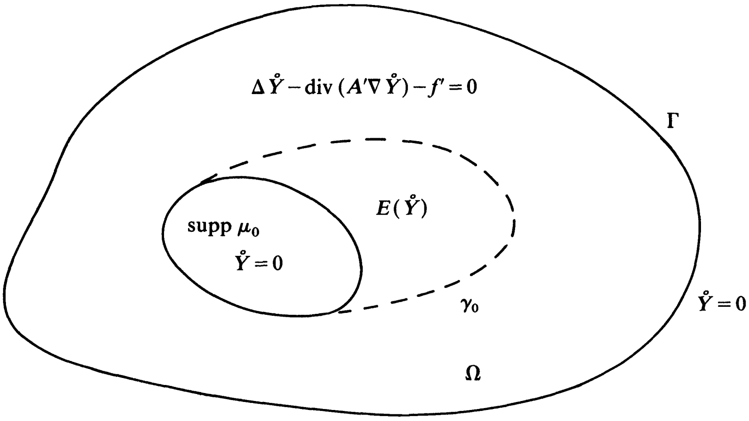


FIG. 3

for an extension $\tilde{Y}_t$ to R^2 of elements $Y_t \in H_0^1(\Omega_t)$

$$\tilde{Y}_t(X) = \begin{cases} Y_t(X), & X \in \Omega_t, \\ 0, & X \in R^2 \backslash \Omega_t, \end{cases}$$

we have the following expansion with respect to the parameter t, for $t > 0$, t small enough,

$$\tilde{Y}_t|_\Omega = Y_0 + tY' + O(t) \quad \text{in } H^1(\Omega)$$

where $\| O(t) \|_{H^1(\Omega)}/t \to 0$ with $t \downarrow 0$ and $Y' \in H^1(\Omega)$ is given by the following variational problem:

find $Y' \in H^1(\Omega)$ that minimizes

$$J(\eta) = \frac{1}{2} \int_\Omega (\nabla\eta)^2 \, dx$$

subject to constraints

$$\eta = -v \frac{\partial Y_0}{\partial n} \quad \text{on } \Gamma,$$

$$\eta = 0 \quad \text{q.e. on supp } \mu_0,$$

$$\eta \leqq 0 \quad \text{q.e. on } E(Y_0) \backslash \text{supp } \mu_0.$$

Here $v(X)$, $X \in \Gamma$ is a function uniquely determined by the family $\{T_t\}$, $t \in [0, \delta)$; for the details we refer the reader to § 3.

In this paper the continuous evolution of a given domain $\Omega \subset R^n$, $n \geqq 2$ that depends on a vector field $\mathbf{V}$ is introduced [2], [15]. The so-called material derivative [16] of the solution of the boundary value problem in a direction of the field $\mathbf{V}$ is defined. Existence of the material derivative is proved for an obstacle problem and for the Signorini variational inequality of the plane elasticity. In the case of the obstacle problem the so-called domain derivative in the direction of a field $\mathbf{V}$ is characterized as a unique solution to an auxiliary variational inequality. Such characterization of the domain derivative can be obtained under the assumption that the solution of the unilateral boundary value problem is sufficiently smooth. Shape sensitivity analysis of linear boundary value problems was investigated by many authors, in particular by J. Cea [2], [3], J.-P. Zolesio [15], [16]; see also [9], [10] and references given in [3].

Related results concerning sensitivity analysis of unilateral problems were presented in [4], [8], [11]-[14]. For a detailed description of the material derivative method we refer to the paper [16].

The plan of the paper is as follows. Notation is introduced in the next section. In § 2, differentiability properties of the solution to an abstract variational inequality with respect to a parameter are investigated. Section 3 describes the results obtained for the obstacle problem. Section 4 contains the results for the Signorini variational inequality.

1.1. Notation. Let H be a Hilbert space. Denote by H' the dual space and by $\langle\cdot,\cdot\rangle$ the duality pairing between H' and H. Let $a(\cdot,\cdot): H\times H\to R$ be a bilinear form and assume that there exist constants α and M, $0<\alpha\leqq M$, such that

$$a(\phi,\phi)\geqq\alpha\|\phi\|_H^2\quad\forall\phi\in H, \tag{1.1}$$

$$|a(\phi,\zeta)|\leqq M\|\phi\|_H\|\zeta\|_H\quad\forall\phi,\zeta\in H. \tag{1.2}$$

Consider the nonlinear mapping:

$$P: H'\to H, \tag{1.3}$$

which is defined in the following way:

for each $f\in H'$ the corresponding element $P(f)\in H$ satisfies the variational inequality

$$\begin{aligned}&P(f)\in K,\\&a(P(f),\zeta-P(f))\geqq\langle f,\zeta-P(f)\rangle\quad\forall\zeta\in K,\end{aligned} \tag{1.4}$$

where K is a convex and closed subset of H.

It is well known [5], [6] that under assumptions (1.1) and (1.2) for any $f\in H'$ there exists a unique solution to (1.4).

DEFINITION 1. Mapping (1.3) is Gateaux differentiable at $f^0\in H'$ if there exists a continuous mapping

$$Q: H'\to H \tag{1.5}$$

such that

$$\forall h\in H': \lim_{\tau\downarrow 0}\|(P(f^0+\tau h)-P(f^0))/\tau-Q(h)\|_H=0. \tag{1.6}$$

Remark. Mapping $Q(\cdot)$ is positively homogeneous,

$$Q(\tau h)=\tau Q(h)\quad\forall\tau>0,\quad\forall h\in H'. \tag{1.7}$$

However, in general, $Q(-h)\neq -Q(h)$.

2. Sensitivity analysis of an abstract variational inequality. Consider the following family of variational inequalities depending on a parameter $t\in[0,\delta)$, $\delta>0$:

$$\begin{aligned}&p^t\in K,\\&a^t(p^t,\zeta-p^t)\geqq\langle f^t,\zeta-p^t\rangle\quad\forall\zeta\in K\end{aligned} \tag{2.1}$$

where for each $t\in[0,\delta]$, $a^t(\cdot,\cdot): H\times H\to R$ is a bilinear form and $f^t\in H'$ is given. For fixed $t\in[0,\delta)$ denote $p^t=P^t(f^t)$, $f^t\in H'$.

THEOREM 1. *Assume that*

(i) *Bilinear forms* $a^t(\cdot,\cdot)$ *satisfy* (1.1), (1.2) *uniformly for* $t\in[0,\delta)$ *and there exists a linear operator* $A'\in L(H;H')$ *such that*

$$\lim_{t\downarrow 0}\sup_{\substack{\|\phi\|_H\leqq 1\\ \|w\|_H\leqq 1}}|(a^t(w,\phi)-a^0(w,\phi))/t-\langle A'w,\phi\rangle|=0; \tag{2.2}$$

(ii) *There exists an element* $f' \in H'$ *such that*

$$\lim_{t\downarrow 0} \sup_{\|\phi\|_H \leqq 1} |\langle f^t - f^0, \phi\rangle / t - \langle f', \phi\rangle| = 0; \tag{2.3}$$

(iii) $P^0(f), f \in H'$ *denotes a unique solution of the variation inequality*

$$\begin{aligned} &P^0(f) \in K, \\ &a^0(P^0(f), \xi - P^0(f)) \geqq \langle f, \xi - P^0(f)\rangle \quad \forall \xi \in K. \end{aligned} \tag{2.4}$$

Assume that the mapping $P^0(\cdot): H' \to H$ *is Gateaux differentiatiable, i.e. for* $t > 0$, t *small enough,*

$$\forall h \in H'\colon P^0(f^0 + th) = P^0(f^0) + tQ(h) + o(t) \tag{2.5}$$

where $\|o(t)\|_H/t \to 0$ *with* $t \downarrow 0$.

Then the solution $p^t \in H$ *to variational inequality* (2.1) *is right-differentiable at* $t = 0$ *in the norm of space* H:

$$p^t = p^0 + tQ(f' - A'p^0) + o(t) \tag{2.6}$$

where $\|o(t)\|_H/t \to 0$ *with* $t \downarrow 0$.

Proof. By standard argument it follows that

$$a^0(p^0 - p^t, p^0 - p^t) \leqq \langle f^t - f^0, p^0 - p^t\rangle + a^0(p^t, p^t - p^0) - a^t(p^t, p^t - p^0). \tag{2.7}$$

Using (1.1) we obtain

$$\alpha \| p^0 - p^t \|_H^2 \leqq \| f^t - f^0 \|_{H'} \| p^0 - p^t \|_H + |a^0(p^t, p^t - p^0) - a^t(p^t, p^t - p^0)|. \tag{2.8}$$

By assumption (i) there exists a constant $C < \infty$ such that for $t > 0$, t small enough,

$$|a^0(p^t, p^t - p^0) - a^t(p^t, p^t - p^0)| \leqq Ct \| p^t \|_H \| p^t - p^0 \|_H. \tag{2.9}$$

Therefore, in view of (2.9), it follows from (2.8) that

$$\| p^t - p^0 \|_H \leqq C_1 t, \qquad t \in [0, \delta). \tag{2.10}$$

Simple calculations show that the element $p^t \in H$ satisfies the following variational inequality:

$$\begin{aligned} &p^t \in K, \\ &a^0(p^t, \xi - p^t) \geqq \langle f^0 + t(f' - A'p^0), \xi - p^t\rangle + \langle r(t), \xi - p^t\rangle \quad \forall \xi \in K \end{aligned} \tag{2.11}$$

where

$$\langle r(t), \xi\rangle = \sum_{i=1}^{3} \langle r_i(t), \xi\rangle \quad \forall \xi \in H,$$

$$r_1(t) = f^t - f^0 - tf', \tag{2.12}$$

$$\langle r_2(t), \xi\rangle \overset{\text{def}}{=} a^0(p^0, \xi) - a^t(p^0, \xi) + t\langle A'p^0, \xi\rangle \quad \forall \xi \in H, \tag{2.13}$$

$$\langle r_3(t), \xi\rangle \overset{\text{def}}{=} a^0(p^t - p^0, \xi) - a^t(p^t - p^0, \xi) \quad \forall \xi \in H. \tag{2.14}$$

By assumptions (i) and (ii) it follows that $\| r_i(t) \|_{H'}/t \to 0$ with $t \downarrow 0$, $i = 1, 2$. Furthermore, in view of (2.2), for $t > 0$, t small enough,

$$|\langle r_3(t), \xi\rangle| \leqq \gamma(t) \| p^t - p^0 \|_H \| \xi \|_{H'} \quad \forall \xi \in H \tag{2.15}$$

where $\gamma(t) \downarrow 0$ with $t \downarrow 0$. Therefore, taking into account (2.10), we obtain

$$\| r_3(t) \|_{H'}/t \to 0 \quad \text{with } t \downarrow 0. \tag{2.16}$$

Thus

$$\|r(t)\|_{H'}/t \to 0 \quad \text{with } t \downarrow 0. \tag{2.17}$$

Hence

$$\begin{aligned} p^t &= P^0(f^0 + t(f' - A'p^0) + r(t)) = P^0(f^0 + t(f' - A'p^0)) + o(t) \\ &= P^0(f^0) + tQ(f' - A'p^0) + o(t) \end{aligned} \tag{2.18}$$

where

$$\|o(t)\|_H/t \to 0 \quad \text{with } t \downarrow 0.$$

3. Sensitivity analysis of an obstacle problem. This section is devoted to sensitivity analysis with respect to the parameter $t \in [0, \delta)$ of an obstacle problem posed in a domain $\Omega_t \subset R^n$.

First we introduce notation. The following notation is used for the scalar product in R^n:

$$\mathbf{a} \cdot \operatorname{grad} \zeta = \langle \mathbf{a}, \operatorname{grad} \zeta \rangle_{R^n} = \sum_{i=1}^n a_i \frac{\partial \zeta}{\partial x_i}$$

where $\mathbf{a} = \operatorname{col}(a_1, \cdots, a_n)$ is a vector and $\operatorname{grad} \zeta = \operatorname{col}(\partial\zeta/\partial x_i, \cdots, \partial\zeta/\partial x_n)$ denotes the gradient of a given function $\zeta = \zeta(x)$, $x = (x_1, \cdots, x_n) \in R^n$. Given $n \times n$ matrix $A = [a_{ij}]$, we denote by $*A$ its transpose matrix, i.e., $*A = [a_{ji}]$. As in [3], [16], we introduce a family of regular domains $\{\Omega_t\} \subset R^n$, $t \in [0, \delta)$ corresponding to a given vector field:

$$\mathbf{V}(\cdot, \cdot): [0, \delta) \times R^n \to R^n. \tag{3.1}$$

Domains Ω_t are constructed as follows. Denote by $T_t = T_t(\mathbf{V})$, $t \in [0, \delta)$ the family of mappings of the form

$$T_t \colon R^n \ni X \to x(t) \in R^n \tag{3.2}$$

where the vector function $x(\cdot)$ satisfies the following ordinary differential equation:

$$\begin{aligned} &\frac{dx}{dt}(s) = \mathbf{V}(s, x(s)), \qquad s \in [0, \delta), \\ &x(0) = X, \end{aligned} \tag{3.3}$$

and denote

$$\Omega_t = T_t(\mathbf{V})(\Omega) = \{x \in R^n \mid \exists X \in \Omega \text{ such that } x = x(t),\ x(0) = X\}. \tag{3.4}$$

Vector field $\mathbf{V}(\cdot, \cdot)$ is assumed to be regular, i.e.,

$$\mathbf{V}(t, \cdot) \in C^1(R^n, R^n) \quad \forall t \in [0, \delta), \tag{3.5}$$

$$\mathbf{V}(\cdot, x) \in C([0, \delta)) \quad \forall x \in R^n. \tag{3.6}$$

Note that from (3.3) and (3.4) it follows that $\Omega_0 \equiv \Omega$.

Denote by $DT_t(X)$ the Jacobian matrix of mapping (3.2) evaluated at $X \in R^n$. Given the family of domains $\{\Omega_t\} \subset R^n$, consider the following variational inequality, parametrized by $t \in [0, \delta)$:

$$\begin{aligned} &z_t \in K_t(\Omega_t) = \{\zeta_t \in H^1(\Omega_t) \mid \zeta = \phi \text{ on } \partial\Omega_t,\ \zeta_t(x) \geqq \psi_0(x) \text{ a.e. in } \Omega_t\}, \\ &a_t(z_t, \zeta_t - z_t) \geqq \langle f_t, \zeta_t - z_t \rangle_t \quad \forall \zeta_t \in K_t(\Omega_t) \end{aligned} \tag{3.7}$$

where $\phi(\cdot)$ and $\psi_0(\cdot)$ are given elements of $C^1(R^n)$, $\langle \cdot, \cdot \rangle_t$ is the duality pairing between $(H^1(\Omega_t))'$ and $H^1(\Omega_t)$. Bilinear forms $a_t(\cdot, \cdot)$ and element $f_t \in (H^1(\Omega_t))'$ are defined respectively by

$$\begin{aligned} a_t(y_t, \zeta_t) = \int_{\Omega_t} &\{\langle A(x) \cdot \operatorname{grad} y_t(x), \operatorname{grad} \zeta_t(x)\rangle_{R^n} \\ &+\zeta_t(x)\langle \mathbf{a}(x), \operatorname{grad} y_t(x)\rangle_{R^n} + a_0(x) y_t(x) \zeta_t(x)\}\, dx, \\ &\qquad t \in [0, \delta) \quad \forall \zeta_t, y_t \in H^1(\Omega_t), \end{aligned} \tag{3.8}$$

$$\langle f_t, \zeta_t\rangle_t = \int_{\Omega_t} \{F_0(x)\zeta_t(x) + \langle \mathbf{F}(x) + \operatorname{grad} \zeta_t(x)\rangle_{R^n}\}\, dx, \qquad t \in [0, \delta) \quad \forall \zeta_t \in H^1(\Omega_t) \tag{3.9}$$

where $A(x) = [a_{ij}(x)]_{n\times n}$, $\mathbf{a}(x) = \operatorname{col}(a_1(x), \cdots, a_n(x))$, $\mathbf{F}(x) = \operatorname{col}(F_1(x), \cdots, F_n(x))$, $a_0(x)$, $F_0(x)$ are given elements and $x \in R^n$.

We shall characterize the so-called domain derivative [16] $z' \in H^1(\Omega)$ for a solution to problem (3.7). Recall that in the case of variational inequality (3.7) the domain derivative can be defined as follows:

$$z'(x) \overset{\text{def}}{=} \frac{\partial \tilde{z}}{\partial t}(0^+, x), \qquad x \in \Omega \tag{3.10}$$

where

$$\tilde{z}(t, x) = \begin{cases} z_t(x), & x \in \Omega_t, \quad t \in [0, \delta), \\ \phi(x), & x \in R^n \backslash \Omega_t, \quad t \in [0, \delta). \end{cases} \tag{3.11}$$

THEOREM 2. *Assume that*

(i) $K_t(\Omega_t)$ *is a nonempty, closed and convex subset of space* $H^1(\Omega_t)$ *for* $t \in [0, \delta)$.

(ii) $a_{ij}(\cdot), a_i(\cdot), a_0(\cdot), F_i(\cdot), F_0(\cdot) \in C^1(R^n), \quad i, j = 1, \cdots, n,$

(iii) *There exists a constant* $\alpha > 0$ *such that*

$$a_t(\zeta_t, \zeta_t) \geqq \alpha \|\zeta_t\|^2_{H^1(\Omega_t)} \quad \forall \zeta_t \in H^1(\Omega_t), \quad \forall t \in [0, \delta); \tag{3.12}$$

(iv) $-\beta(x) \overset{\text{def}}{=} \phi(x) - \psi_0(x) > 0 \quad \forall x \in R^n$,

$\beta(\cdot), \quad 1/\beta(\cdot) \in C^1(R^n), \quad \phi \in C^2(R^n)$;

then the domain derivative $z' \in H^1(\Omega)$ *for the problem* (1) *exists and is given by the following variational inequality*:

$$\begin{aligned} &z' \in S_v(\Omega), \\ &a_0(z', \zeta - z') \geqq 0 \quad \forall \zeta \in S_v(\Omega), \end{aligned} \tag{3.13}$$

where the cone $S_v(\Omega) \subset H^1(\Omega)$ *is defined by*

$$\begin{aligned} S_v(\Omega) = \Big\{ \xi \in H^1(\Omega) \,|\, \xi = -v\frac{\partial}{\partial n}(z_0 - \phi) \textit{ on } \Gamma \xi(X) \geqq 0 \quad \textit{q.e. on } Z = Z(z_0 - \psi_0), \\ \int_Z (F_0(X) - \operatorname{div} \mathbf{F}(X) - \mathbf{a}(X) \cdot \nabla \psi_0(X) - a_0(X)\psi_0(X) \\ + \operatorname{div}(A(X) \cdot \nabla \psi_0(X)))\xi(X)\, dX = 0 \Big\}. \end{aligned}$$

For the proof of Theorem 2 we need some lemmas. Lemma 1 describes the results obtained in the case of an obstacle problem defined in the fixed domain Ω. We introduce the necessary notation. Let there be given $n\times n$ matrix function $C(t,x)=[c_{ij}(t,x)]$, vector functions $\mathbf{c}(t,x)=\mathrm{col}\,(c_1(t,x),\cdots,c_n(t,x))$, $\mathbf{H}(t,x)=\mathrm{col}\,(H_1(t,x),\cdots,H_n(t,x))$ and elements $c_0(t,x)$, $H_0(t,x)$, $t\in[0,\delta)$, $x\in\bar{\Omega}$, which are supposed to be continuous on $[0,\delta)x\bar{\Omega}$ and continuously differentiable with respect to t at $t=0^+$. Define the following:

1) bilinear form $c^t(\cdot,\cdot)$ on $H_0^1(\Omega)$

$$\begin{aligned} c^t(y,\zeta)=\int_\Omega \{&\langle C(t,x)\cdot\operatorname{grad} y(x),\operatorname{grad}\zeta(x)\rangle_{R^n}\\ &+\zeta(x)\langle \mathbf{c}(t,x),\operatorname{grad} y(x)\rangle_{R^n}+c_0(t,x)y(x)\zeta(x)\}\,dx \end{aligned}\tag{3.14}$$

$$\forall y,\zeta\in H_0^1(\Omega),\quad \forall t\in[0,\delta);$$

2) elements $\dot{h}^t\in H^{-1}(\Omega)$

$$\langle h^t,\zeta\rangle=\int_\Omega\{H_0(t,x)\zeta(x)+\langle\mathbf{H}(t,x),\operatorname{grad}\zeta(x)\rangle_{R^n}\}\,dx\tag{3.15}$$

$$\forall\zeta\in H_0^1(\Omega),\quad\forall t\in[0,\delta);$$

3) linear operator $C'\in L(H_0^1(\Omega);H^{-1}(\Omega))$

$$\begin{aligned}\langle C'y,\zeta\rangle=\int_\Omega\bigg\{&\left\langle\frac{\partial C}{\partial t}(0^+,x)\cdot\operatorname{grad} y(x),\operatorname{grad}\zeta(x)\right\rangle_{R^n}\\ &+\zeta(x)\left\langle\frac{\partial \mathbf{c}}{\partial t}(0^+,x),\operatorname{grad} y(x)\right\rangle_{R^n}+\frac{\partial c_0}{\partial t}(0^+,x)y(x)\zeta(x)\bigg\}\,dx\end{aligned}\tag{3.16}$$

$$\forall y,\zeta\in H_0^1(\Omega);$$

4) element $h'\in H^{-1}(\Omega)$

$$\langle h',\zeta\rangle=\int_\Omega\left\{\frac{\partial H_0}{\partial t}(0^+,x)\zeta(x)+\left\langle\frac{\partial\mathbf{H}}{\partial t}(0^+,x),\operatorname{grad}\zeta(x)\right\rangle_{R^n}\right\}dx\qquad\forall\zeta\in H_0^1(\Omega).\tag{3.17}$$

Consider the variational inequality

$$\begin{aligned}&w^t\in K(\Omega)=\{\zeta\in H_0^1(\Omega)\,|\,\zeta(x)\leqq\gamma\text{ a.e. in }\Omega\},\\ &c^t(w^t,\zeta-w^t)\geqq\langle h^t,\zeta-w^t\rangle\quad\forall\zeta\in K(\Omega)\end{aligned}\tag{3.18}$$

where $\gamma\in R$ is given constant. In the sequel, without loss of generality we assume that $\gamma=1$.

LEMMA 1. *Assume that there exists $\alpha>0$ such that*

$$\forall t\in[0,\delta):\quad c^t(\zeta,\zeta)\geqq\alpha\|\zeta\|^2_{H_0^1(\Omega)}\quad\forall\zeta\in H_0^1(\Omega).$$

Then

$$w^t=w^0+tq+o(t)\tag{3.19}$$

where $\|o(t)\|_{H_0^1(\Omega)}/t\to0$ with $t\downarrow0$ and $q\in H_0^1(\Omega)$. Sensitivity coefficient $q\in H_0^1(\Omega)$ satisfies the variational inequality

$$\begin{aligned}&q\in S_0(\Omega),\\ &c^0(q,\zeta-q)\geqq\langle h'-C'w^0,\zeta-q\rangle\quad\forall\zeta\in S_0(\Omega)\end{aligned}\tag{3.20}$$

where

$$S_0(\Omega)=\{\zeta\in H_0^1(\Omega)\,|\,\zeta(x)\leqq0\ q.e.\ on\ \Omega_0(w^0),\ c^0(w^0,\zeta)=\langle h^0,\zeta\rangle\}\tag{3.21}$$

and

$$\Omega_0(w^0)=\{x\in\Omega \mid w^0(x)=1\}. \tag{3.22}$$

Proof. We apply Theorem 1 to variational inequality (3.18). Assumptions (i) and (ii) of Theorem 1 are obviously satisfied by operator C' and element h'.

We verify assumption (iii) of Theorem 1. Denote by $w^0=\pi(f^0)$ the unique solution to (3.18) for $t=0$.

By a result of Mignot [8] we have

$$\forall h\in H^{-1}(\Omega)\colon \pi(f^0+\tau h)=\pi(f^0)+\tau\pi'(h)+o(\tau) \tag{3.23}$$

where element $\pi'(h)\in S_0(\Omega)$ satisfies the variational inequality

$$c^0(\pi'(h),\zeta-\pi'(h))\geqq\langle h,\zeta-\pi'(h)\rangle \quad \forall\zeta\in S_0(\Omega). \tag{3.24}$$

Hence in this case $Q(h)\overset{\text{def}}{=}\pi'(h)$, $\forall h\in H^{-1}(\Omega)$.

Next Lemma 2 below applies to an obstacle problem posed in variable domain Ω_t

$$\begin{aligned}&w_t\in K(\Omega_t),\\&b_t(w_t,\zeta_t-w_t)\geqq\langle g_t,\zeta_t-w_t\rangle \quad \forall\zeta_t\in K(\Omega_t),\end{aligned} \tag{3.25}$$

where

$$K(\Omega_t)=\{\zeta_t\in H_0^1(\Omega_t)\mid \zeta_t(x)\leqq 1 \text{ a.e. in } \Omega_t\}, \tag{3.26}$$

$$\begin{aligned}b_t(y_t,\zeta_t)=\int_{\Omega_t}&\{\langle B(x)\cdot\operatorname{grad} y_t(x),\operatorname{grad}\zeta_t(x)\rangle_{R^n}\\&+\zeta_t(x)\langle \mathbf{b}(x),\operatorname{grad} y_t(x)\rangle_{R^n}+b_0(x)y_t(x)\zeta_t(x)\}\,dx\\&\qquad\forall y_t,\zeta_t\in H_0^1(\Omega_t),\quad t\in[0,\delta),\end{aligned} \tag{3.27}$$

$$\langle g_t,\zeta_t\rangle_t=\int_\Omega\{G_0(x)\zeta_t(x)+\langle \mathbf{G}(x),\operatorname{grad}\zeta_t(x)\rangle_{R^n}\}\,dx \qquad \forall\zeta_t\in H_0^1(\Omega_t). \tag{3.28}$$

Henceforth we assume that there exists a unique solution $w_t\in H_0^1(\Omega_t)$ to (3.25) for $t\in[0,\delta)$ and denote

$$\dot w(\Omega)=\lim_{t\downarrow 0}(w_t\circ T_t-w_0)/t. \tag{3.29}$$

LEMMA 2. *Assume that*

(i) $G_0(\cdot)\in C^1(R^n)$, $\quad \mathbf{G}(\cdot)\in[C^1(R^n)]^n$,

$B(\cdot)\in[C^1(R^n)]^{n^2}$, $\quad \mathbf{b}(\cdot)\in[C^1(R^n)]^n$,

$b_0(\cdot)\in C^1(R^n)$;

(ii) *There exists a constant* $\alpha>0$ *such that*

$$b_t(\zeta_t,\zeta_t)\geqq\alpha\|\zeta_t\|^2_{H_0^1(\Omega_t)} \quad \forall\zeta_t\in H_0^1(\Omega_t),\quad \forall t\in[0,\delta); \tag{3.30}$$

then mapping

$$[0,\delta)\ni t\to w_t\circ T_t\in H_0^1(\Omega) \tag{3.31}$$

is strongly differentiable at $t=0^+$ and the element $\dot{w}(\Omega)\in H_0^1(\Omega)$ satisfies the variational inequality

$$\begin{aligned}&\dot{w}(\Omega)\in S_1(\Omega),\\ &b_0(\dot{w}(\Omega),\zeta-\dot{w}(\Omega))\geqq\langle g'-B'w_0,\zeta-\dot{w}(\Omega)\rangle\quad\forall\zeta\in S_1(\Omega)\end{aligned}\tag{3.32}$$

where

$$\begin{aligned}&S_1(\Omega)=\{\zeta\in H_0^1(\Omega)\,|\,\zeta(x)\leqq 0 \text{ q.e. on } \Omega_1(w_0),\ b_0(w_0,\zeta)=\langle g_0,\zeta\rangle\},\\ &\Omega_1(w_0)=\{x\in\Omega\,|\,w_0(x)=1\}.\end{aligned}\tag{3.33}$$

Element $g'\in H^{-1}(\Omega)$ and operator $B'\in L(H_0^1(\Omega);\ H^{-1}(\Omega))$ are defined by (3.52) *and* (3.51), *respectively.*

Proof. Using mapping T_t we transport variational inequality (3.25) defined in Ω_t to the domain $\Omega=T_t^{-1}(\Omega_t)$. Then we can apply Lemma 1 to the resulting variational inequality defined in Ω. To this end we need the following notation:

$$B^t(x)=\gamma(t,x)(^*DT_t^{-1})(x)\cdot(B\circ T_t)(x)\cdot(DT_t^{-1})(x),\tag{3.34}$$

$$\mathbf{b}^t(x)=\gamma(t,x)(DT_t^{-1})(x)\cdot(\mathbf{b}\circ T_t)(x),\tag{3.35}$$

$$b_0^t(x)=\gamma(t,x)(b_0\circ T_t)(x),\tag{3.36}$$

$$\mathbf{G}^t(x)=\gamma(t,x)(^*DT_t^{-1})(x)\cdot(\mathbf{G}\circ T_t)(x),\tag{3.37}$$

$$G_0^t(x)=\gamma(t,x)(G_0\circ T_t)(x)\tag{3.38}$$

where $\gamma(t,x)=\det(DT_t(x))$,

$$\begin{aligned}b^t(y,\zeta)=\int_\Omega&\{\langle B^t(x)\operatorname{grad}y(x),\operatorname{grad}\zeta(x)\rangle_{R^n}\\ &+\zeta(x)\langle\mathbf{b}^t(x),\operatorname{grad}y(x)\rangle_{R^n}+b_0^t(x)y(x)\zeta(x)\}\,dx\qquad\forall y,\zeta\in H_0^1(\Omega),\end{aligned}\tag{3.39}$$

$$\langle g^t,\zeta\rangle=\int_\Omega\{G_0^t(x)\zeta(x)+\langle\mathbf{G}^t(x),\operatorname{grad}\zeta(x)\rangle_{R^n}\}\,dx\quad\forall\zeta\in H_0^1(\Omega).\tag{3.40}$$

It can be verified that

$$b^t(y_t\circ T_t,\zeta_t\circ T_t)=b_t(y_t,\zeta_t)\quad\forall y_t,\zeta_t\in H_0^1(\Omega_t),\tag{3.41}$$

$$\langle g^t,\zeta_t\circ T_t\rangle=\langle g_t,\zeta_t\rangle\quad\forall\zeta_t\in H_0^1(\Omega_t).\tag{3.42}$$

Furthermore

$$y_t\circ T_t\in K(\Omega)\quad\text{iff } y_t\in K(\Omega_t).\tag{3.43}$$

Hence it follows that element $w_t\circ T_t\in H_0^1(\Omega)$ satisfies the variational inequality

$$\begin{aligned}&w_t\circ T_t\in K(\Omega),\\ &b^t(w_t\circ T_t,\zeta-w_t\circ T_t)\geqq\langle g^t,\zeta-w_t\circ T_t\rangle\quad\forall\zeta\in K(\Omega).\end{aligned}\tag{3.44}$$

By assumption (iii) (see [14]), there exists a constant $\alpha=\alpha(\delta)>0$, for $\delta>0$, sufficiently small, such that

$$b^t(\zeta,\zeta)\geqq\alpha\|\zeta\|^2_{H_0^1(\Omega)}\quad\forall\zeta\in H_0^1(\Omega),\quad\forall t\in[0,\delta).\tag{3.45}$$

Hence there exists a unique solution to (3.44). In order to apply Lemma 1 to variational inequality (3.44), we calculate derivatives with respect to t at $t=0^+$ of the coefficients of bilinear form $b^t(\cdot,\cdot)$ and linear forms $g^t\in H^{-1}(\Omega)$. We obtain

$$\frac{\partial B^t}{\partial t}(x)=(\operatorname{div}\mathbf{V})(x)B(x)-((^*D\mathbf{V})(x)\cdot B(x)+B(x)\cdot(D\mathbf{V})(x))+\hat{B}(x) \tag{3.46}$$

where

$$\hat{B}=[\hat{b}_{ij}]_{n\times n},$$

$$\hat{b}_{ij}(\mathbf{x})=\sum_{k=1}^{n}\frac{\partial}{\partial x_k}b_{ij}(x)V_k(x),\qquad i,j=1,\cdots,n,$$

$$\frac{\partial \mathbf{b}^t}{\partial t}(x)=(\operatorname{div}\mathbf{V})(x)\mathbf{b}(x)-(D\mathbf{V})(x)\cdot\mathbf{b}(x)+\hat{\mathbf{b}}(x), \tag{3.47}$$

$$\hat{\mathbf{b}}=\operatorname{col}(\hat{b}_1,\cdots,\hat{b}_n),$$

$$\hat{b}_i(x)=\sum_{k=1}^{n}\frac{\partial}{\partial x_k}b_i(x)V_k(x),\qquad i=1,\cdots,n,$$

$$\frac{\partial b_0^t}{\partial t}(x)=\operatorname{div}(b_0\mathbf{V})(x), \tag{3.48}$$

$$\frac{\partial \mathbf{G}^t}{\partial t}(x)=(\operatorname{div}\mathbf{V})(x)\mathbf{G}(x)-(^*D\mathbf{V})(x)\cdot\mathbf{G}(x)+\hat{\mathbf{G}}(x), \tag{3.49}$$

$$\frac{\partial G_0^t}{\partial t}x=\operatorname{div}(G_0\mathbf{V})(x) \tag{3.50}$$

where $\mathbf{V}(x)=\mathbf{V}(0,x)$ and $\hat{\mathbf{G}}(x)$ is defined in exactly the same way as $\hat{\mathbf{b}}(x)$.

Denote by $B'\in L(H_0^1(\Omega);H^{-1}(\Omega))$, $g'\in H^{-1}(\Omega)$

$$\begin{aligned}\langle B'y,\zeta\rangle=\int_\Omega\Big\{&\Big\langle\frac{\partial B^t}{\partial t}(x)\operatorname{grad}y(x),\operatorname{grad}\zeta(x)\Big\rangle_{R^n}\\&+y(x)\Big\langle\frac{\partial \mathbf{b}^t}{\partial t}(x),\operatorname{grad}\zeta(x)\Big\rangle_{R^n}+\frac{\partial b_0^t}{\partial t}(x)y(x)\zeta(x)\Big\}\,dx\end{aligned} \tag{3.51}$$

$$\forall y,\zeta\in H_0^1(\Omega),$$

$$\langle g',\zeta\rangle=\int_\Omega\Big\{\frac{\partial G_0^t}{\partial t}(x)\zeta(x)+\Big\langle\frac{\partial \mathbf{G}^t}{\partial t}(x),\operatorname{grad}\zeta(x)\Big\rangle_{R^n}\Big\}\,dx\quad\forall\zeta\in H_0^1(\Omega). \tag{3.52}$$

Since, in the case of variational inequality (3.44), all assumptions of Lemma 1 are satisfied it follows that

$$w_t\circ T_t=w_0+t\dot{w}(\Omega)+o(t)$$

where $\|o(t)\|_{H_0^1(\Omega)}/t\to 0$ with $t\downarrow 0$ and element $\dot{w}(\Omega)\in H_0^1(\Omega)$ is a unique solution to (3.32).

We now characterize the domain derivative

$$w'=\dot{w}-\langle\operatorname{grad}w_0,\mathbf{V}\rangle_{R^n} \tag{3.53}$$

in the case of variational inequality (3.25).

We prove that w' depends actually on $v=\langle\mathbf{V}(0),\mathbf{n}\rangle_{R^n}$, v is the normal component of the vector field $\mathbf{V}(0)$ on Γ.

LEMMA 3. *If assumptions* (i) *and* (ii) *of Lemma* 2 *are satisfied, then the domain derivative* $w' \in H^1(\Omega)$ *satisfies the variational inequality*

$$\begin{aligned} &w' \in S_v^0(\Omega), \\ &b_0(w', \zeta - w') \geqq 0 \quad \forall \zeta \in S_v^0(\Omega) \end{aligned} \tag{3.54}$$

where

$$\begin{aligned} S_v^0(\Omega) = \Bigg\{ \zeta \in H^1(\Omega) \,|\, \zeta|_\Gamma = -\mathbf{v}\frac{\partial w_0}{\partial n},\ \zeta(x) \leqq 0 \text{ } q.e. \text{ } on\, Z(w_0-1), \\ \cdot \int_{Z(w_0-1)} (G_0(x) - b_0(x) - \operatorname{div} \mathbf{G}(x))\zeta(x)\,dx = 0 \Bigg\}, \end{aligned} \tag{3.55}$$

$$Z(w_0-1) = \{x \in \Omega \,|\, w_0(x) = 1\}.$$

Proof. If the assumptions of Lemma 2 are satisfied then (see [1])

$$w_0 \in H^2(\Omega) \cap C^{1,\alpha}(\bar{\Omega}) \quad \forall \alpha < 1. \tag{3.56}$$

Hence $\langle \operatorname{grad} w_0, \mathbf{V} \rangle_{R^n} \in H^1(\Omega)$ for an arbitrary regular vector field $\mathbf{V} = \mathbf{V}(0, \cdot)$.

Denote

$$S_v^0(\Omega) = \{\phi \in H^1(\Omega) \,|\, \phi = \zeta - \langle \operatorname{grad} w_0, \mathbf{V} \rangle_{R^n},\ \zeta \in S_2(\Omega)\}. \tag{3.57}$$

Simple calculations show, by taking into account (3.56), that cone (3.57) has the form (3.55).

Note that

$$S_v^0(\Omega) - S_v^0(\Omega) = S_2(\Omega) - S_2(\Omega). \tag{3.58}$$

On the other hand, it follows from (3.53) that

$$\begin{aligned} &w' \in S_v^0(\Omega), \\ &b_0(w', \zeta - w') \geqq G(w_0, \mathbf{V}, \zeta - w') \quad \forall \zeta \in S_v^0(\Omega) \end{aligned} \tag{3.59}$$

where

$$G(w_0, \mathbf{V}, \zeta) = \langle g' - B'w_0, \zeta \rangle - b_0(\langle \operatorname{grad} w_0, \mathbf{V} \rangle_{R^n}, \zeta). \tag{3.60}$$

Element $g' \in H^{-1}(\Omega)$ and operator $B' \in L(H_0^1(\Omega);\ H^{-1}(\Omega))$ are defined by (3.52) and (3.51), respectively.

Observe that for a vector field $\mathbf{V}$ such that $\langle \mathbf{V}, \mathbf{n} \rangle_{R^n} = 0$ on Γ, the mapping $[0, \delta) \ni t \to \Omega_t \subset R^n$ is constant, i.e. $\Omega_t = \Omega$, $t \in [0, \delta)$, hence $w' = 0$ and $\dot{w} = \langle \operatorname{grad} w_0, \mathbf{V} \rangle_{R^n}$, whence

$$0 \geqq G(w_0, \mathbf{V}, \zeta) \quad \forall \zeta \in \{S_2(\Omega) - S_2(\Omega)\}. \tag{3.61}$$

If we put $\pm\mathbf{V}$ in (3.61) then we obtain

$$G(w_0, \mathbf{V}, \zeta) = 0 \quad \forall \zeta \in \{S_2(\Omega) - S_2(\Omega)\} \tag{3.62}$$

for an arbitrary vector field $\mathbf{V}$ such that $\mathbf{V}|_\Gamma \cdot \mathbf{n} = 0$, where we denote $\mathbf{V}|_\Gamma \cdot \mathbf{n} = \langle \mathbf{V}|_\Gamma, \mathbf{n} \rangle_{R^n}$.

From a general result of J. P. Zolesio [15] it follows that there exists a distribution $g_n \in D_1'(\Gamma)$ such that

$$G(w_0, \mathbf{V}, \zeta) = \langle g_n(w_0, \zeta), \mathbf{V} \cdot \mathbf{n} \rangle_{D_1'(\Gamma) \times D_1(\Gamma)} \quad \forall \zeta \in \{S_2(\Omega) - \mathrm{S}_2^0(\Omega)\} \tag{3.63}$$

and domain derivative $w' \in H^1(\Omega)$ satisfies the variational inequality

$$\begin{aligned} &w' \in S_v(\Omega), \\ &b_0(w', \zeta - w') \geqq \langle g_n(w_0, \zeta - w'), \mathbf{V} \cdot \mathbf{n} \rangle_{D_1'(\Gamma) \times D_1(\Gamma)} \quad \forall \zeta \in S_v^0(\Omega). \end{aligned} \tag{3.64}$$

Taking into account (3.56) and by integrating by parts we obtain

$$G(w_0, \mathbf{V}, \zeta) = \int_\Omega K(w_0, \zeta) \cdot \mathbf{V}\, dx + \int_\Gamma g_n(w_0, \zeta)\mathbf{V} \cdot \mathbf{n}\, d\Gamma \qquad \forall \zeta \in H^2(\Omega) \cap \{S_v^0(\Omega) - S_v^0(\Omega)\}, \tag{3.65}$$

and from (3.63) it follows that

$$\mathbf{K}(w_0, \zeta) = \mathbf{0} \quad \forall \zeta \in H^2(\Omega) \cap \{S_v^0(\Omega) - S_v^0(\Omega)\}. \tag{3.66}$$

Simple calculations show (see [14]) that

$$\int_\Gamma g_n(w_0, \zeta)\mathbf{V} \cdot \mathbf{n}\, d\Gamma = 0 \quad \forall \zeta \in H^2(\Omega) \cap H_0^1(\Omega). \tag{3.67}$$

Hence

$$G(w_0, \mathbf{V}, \zeta) = 0 \quad \forall \zeta \in H^2(\Omega) \cap \{S_v^0(\Omega) - S_v^0(\Omega)\}. \tag{3.68}$$

In order to prove that (3.68) is satisfied for an arbitrary element $\zeta \in \{S_v^0(\Omega) - S_v^0(\Omega)\}$, consider vector field $\mathbf{V}$ with compact support in some open neighbourhood of the boundary $\Gamma = \partial\Omega$. Let U, $U_1 \subset \Omega$ be open sets such that $\bar{U} \subset U_1$, $\overline{Z(w_0 - 1)} \subset U$ and let $\eta(\cdot)$ be a smooth function such that

$$\begin{aligned} &0 \leqq \eta(x) \leqq 1, \quad x \in \bar{\Omega}, \\ &\eta(x) = 1, \quad x \in U, \\ &\eta(x) = 0, \quad x \in \bar{\Omega} \backslash U_1. \end{aligned} \tag{3.69}$$

Given an element $\phi \in \{S_v^0(\Omega) - S_v^0(\Omega)\}$, let $\{\phi_n\} \subset H_0^1(\Omega) \cap H^2(\Omega)$ be a sequence such that

$$\phi_n \to \phi \quad \text{strongly in } H_0^1(\Omega). \tag{3.70}$$

Note that the sequence

$$\zeta_n = (1 - \eta)\phi_n \in \{S_v^0(\Omega) - S_v^0(\Omega)\} \cap H^2(\Omega) \tag{3.71}$$

for every vector field $\mathbf{V} \equiv 0$ on U_1.

Since

$$G(w_0, \mathbf{V}, \phi_n) = \int_\Gamma g_n(w_0, (1 - \eta)\phi_n)\mathbf{V} \cdot \mathbf{n}\, d\Gamma = 0, \qquad n = 1, 2, \cdots,$$

it follows that

$$G(w_0, \mathbf{V}, \phi) = \lim_{n \to \infty} G(w_0, \mathbf{V}, \phi_n) = 0 \qquad \forall \phi \in \{S_v^0(\Omega) - S_v^0(\Omega)\}.$$

Proof of Theorem 2. In order to use Lemma 3 we make a particular choice of coefficients of bilinear form $b_t(\cdot, \cdot)$ and linear form $\langle g_t, \cdot \rangle$ in order to obtain the solution $z_t \in H^1(\Omega_t)$ to the problem (3.7) in the form

$$z_t = \beta w_t + \phi, \qquad t \in [0, \delta) \tag{3.72}$$

where $w_t \in H_0^1(\Omega_t)$ is a unique solution to the problem (3.25), $\beta(x) = \psi_0(x) - \phi(x)$, $x \in R^n$. We select the coefficients $B(\cdot)$, $\mathbf{b}(\cdot)$, $b_0(\cdot)$ of bilinear form (3.27) and the coefficients $G_0(\cdot)$, $\mathbf{G}(\cdot)$ of linear form (3.28) in the following way:

$$B(x) = \beta^2(x)A(x), \tag{3.73}$$

$$\mathbf{b}(x) = \beta(x)({}^*A(x) - A(x)) \cdot \nabla\beta(x) + \beta^2(x)\mathbf{a}(x), \tag{3.74}$$

$$b_0(x) = \beta(x)(\beta(x)a_0(x) + \mathbf{a}(x) \cdot \nabla\beta - \operatorname{div}(A(x) \cdot \nabla\beta(x))), \tag{3.75}$$

$$\begin{aligned} G_0(x) &= \beta(x)(F_0(x) - \mathbf{a}(x) \cdot \nabla\phi(x) - a_0(x)\phi(x)) + \mathbf{F}(x) \cdot \nabla\beta(x) \\ &\quad - \langle A(x) \cdot \nabla\phi(x), \nabla\beta(x)\rangle_{R^n}, \end{aligned} \tag{3.76}$$

$$\mathbf{G}(x) = \beta(x)(\mathbf{F}(x) - A(x) \cdot \nabla\phi(x)) \tag{3.77}$$

where $A(\cdot)$, $\mathbf{a}(\cdot)$, $a_0(\cdot)$ and $F_0(\cdot)$, $\mathbf{F}(\cdot)$ are the coefficients of bilinear form (3.8) and linear form (3.9), respectively.

Using (3.73)-(3.77) we obtain

$$b_t(w_t, \xi_t) = a_t(\beta w_t, \beta\xi_t) \quad \forall w_t, \xi_t \in H_0^1(\Omega_t), \tag{3.78}$$

$$\langle g_t, \xi_t\rangle_t = \langle f_t, \beta\xi_t\rangle_t - a_t(\phi, \beta\xi_t) \quad \forall \xi_t \in H_0^1(\Omega_t), \tag{3.79}$$

therefore simple calculations show that (3.72) holds.

The domain derivative $w' \in H^1(\Omega)$ for the problem (3.25) is given by (3.54), henceforth, in view of (3.72), it follows that there exists the domain derivative $z' \in H^1(\Omega)$ for the problem (3.7) of the form

$$z' = \beta w'. \tag{3.80}$$

Furthermore, $z'/\beta \in S_v^0(\Omega)$.

Let us denote by $S_v(\Omega) \subset H^1(\Omega)$ a closed and convex cone such that $z' \in S_v(\Omega)$. Furthermore,

$$\beta\xi \in S_v(\Omega) \quad \text{iff } \xi \in S_v^0(\Omega).$$

Let $\eta \in S_v(\Omega)$ be a given element; then $\xi = \eta/\beta \in S_v^0(\Omega)$ and using (3.55) we obtain

(i) $\eta = -v(\partial/\partial n)(z_0 - \phi)$ on Γ, since $\xi = \eta/\beta = -v(\partial w_0/\partial n)$ on Γ and $w_0 = (z_0 - \phi)/(\psi_0 - \phi)$;

(ii) $\eta \geqq 0$ q.e. on $Z(z_0 - \psi_0)$ since $\eta/\beta \leqq 0$ q.e. on $Z(w_0 - 1)$ and $\beta(x) < 0$ for all $x \in R^n$;

(iii) It can be shown, in view of (3.75)-(3.77), that using the integral condition on $Z(w_0 - 1)$ we obtain

$$\begin{aligned} \int_{Z(z_0 - \psi_0)} &(F_0(x) - \operatorname{div} \mathbf{F}(x) - \mathbf{a}(x) \cdot \nabla\psi_0(x) - a_0(x)\psi_0(x) \\ &+ \operatorname{div}(A(x) \cdot \nabla\psi_0(x)))\eta(x)\, dx = 0. \end{aligned}$$

4. Sensitivity analysis of the Signorini variational inequality. Let $\Omega \in R^2$ be a domain with smooth boundary $\Gamma = \partial\Omega$. Given a vector field $\mathbf{V}(t, x)$, $t \in [0, \delta)$, $x \in R^2$, let $\{\Omega_t\} \subset R^2$ be the corresponding family of domains (3.4). We assume that $\Omega_t \subset R^2$, for each $t \in [0, \delta)$ is a bounded connected region with smooth boundary $\partial\Omega_t$ and may be expressed as

$$\partial\Omega_t = \Gamma_t^0 \cup \Gamma_t^1 \cup \Gamma_t^c \tag{4.1}$$

where $\Gamma_t^0, \Gamma_t^1, \Gamma_t^c$ are smooth, disjoint and open one-dimensional manifolds.

We assume that Ω_t lies on one side of $\partial\Omega_t$ and that the one-dimensional measures of Γ_t^0, Γ_t^c are strictly positive for $t \in [0, \delta)$.

We denote $\partial\Omega=\Gamma^0\cup\Gamma^1\cup\Gamma^c$. Furthermore, $\mathbf{n}=(n_t^1, n_t^2)$ and $\mathbf{n}=(n_1, n_2)$ denote the unit outward normals to $\partial\Omega_t$ and $\partial\Omega$, respectively. A vector function $\mathbf{V}_t$ and a tensor function $\boldsymbol{\tau}_t$ defined on Ω_t have components v_t^i, τ_t^{ij}, respectively, for $i, j=1,2$. A vector function $\mathbf{v}$ and tensor functions $\boldsymbol{\tau}$, $\mathbf{c}$ defined on R^2 or on Ω have components v_i, τ_{ij}, c_{ijkl} respectively, and $i, j, k, l=1,2$.

We use the summation convention of summing on all repeated indices. The following notation is used for the scalar product of vectors:

$$\boldsymbol{\zeta}\cdot\mathbf{n}=\zeta_i n_i \tag{4.2}$$

and for the convolution of tensors:

$$\begin{gathered}\mathbf{e}\,..\,\mathbf{c}\,..\,\boldsymbol{\tau}=c_{ijkl}e_{ij}\tau_{kl},\\ \boldsymbol{\zeta}\,.\,\boldsymbol{\tau}\,.\,\mathbf{v}=\tau_{ij}\zeta_i v_j,\\ (\boldsymbol{\tau}\cdot\mathbf{v})_i=\tau_{ij}v_j.\end{gathered} \tag{4.3}$$

Denote

$$\mathbf{H}(\Omega_t)=\{\boldsymbol{\zeta}_t\in[H^1(\Omega_t)]^2\,|\,\boldsymbol{\zeta}_t(x)=\mathbf{0}, x\in\Gamma_t^0\} \tag{4.4}$$

and let $\mathbf{K}(\Omega_t)$ be a convex and closed set of the form

$$\mathbf{K}(\Omega_t)=\{\boldsymbol{\zeta}_t\in\mathbf{H}(\Omega_t)\,|\,\boldsymbol{\zeta}_t\cdot\mathbf{n}_t\leqq 0 \text{ a.e. on } \Gamma_t^c\}. \tag{4.5}$$

Given a tensor function $\mathbf{c}$ with components $c_{ijkl}\in L^\infty_{\text{loc}}(R^2)$, $i, j, k, l=1,2$, assume that the following conditions are satisfied:

$$c_{ijkl}(x)=c_{jikl}(x)=c_{klij}(x),\quad x\in R^2,\quad i, j, k, l=1,2, \tag{4.6}$$

$\exists\nu_0>0$ such that

$$c_{ijkl}(x)e_{ij}e_{kl}\geqq\nu_0 e_{ij}e_{ij},\qquad x\in R^2 \tag{4.7}$$

for every symmetric tensor $\mathbf{e}$.

Denote $\mathbf{a}_t(\cdot,\cdot)$, $t\in[0,\delta)$, the bilinear form

$$\mathbf{a}_t(\mathbf{v}_t,\zeta_t)=\int_{\Omega_t}\boldsymbol{\varepsilon}(\mathbf{v}_t)\,..\,\mathbf{c}\,..\,\boldsymbol{\varepsilon}(\boldsymbol{\zeta}_t)\,dx \tag{4.8}$$

where the symmetric tensor $\boldsymbol{\varepsilon}(\mathbf{v})$ have components

$$\varepsilon_{ij}(\mathbf{v})(x)=\frac{1}{2}\left(\frac{\partial v_j}{\partial x_i}(x)+\frac{\partial v_i}{\partial x_j}(x)\right). \tag{4.9}$$

Given elements $\mathbf{f}=(f_1, f_2)\in[L^2(R^2)]^2$, $\mathbf{P}=(P_1, P_2)\in[H^1(R^2)]^2$, denote $\mathbf{F}_t\in(\mathbf{H}(\Omega_t))'$

$$\langle\mathbf{F}_t,\boldsymbol{\zeta}_t\rangle_t\overset{\text{def}}{=}\int_{\Omega_t}f_i(x)\zeta_t^i(x)\,dx+\int_{\Gamma_t^1}P_i(x)\zeta_t^i(x)\,d\Gamma\quad\forall\boldsymbol{\zeta}_t\in\mathbf{H}(\Omega_t). \tag{4.10}$$

Consider, for $t\in[0,\delta)$, the Signorini variational inequality

$$\begin{gathered}\mathbf{u}_t\in\mathbf{K}(\Omega_t),\\ \mathbf{a}_t(\mathbf{u}_t,\boldsymbol{\zeta}_t-\mathbf{u}_t)\geqq\langle\mathbf{F}_t,\boldsymbol{\zeta}_t-\mathbf{u}_t\rangle\quad\forall\boldsymbol{\zeta}_t\in\mathbf{K}(\Omega_t).\end{gathered} \tag{4.11}$$

THEOREM 3. *Mapping*

$$[0,\delta)\ni t\to u_t\circ T_t\in\mathbf{H}(\Omega) \tag{4.12}$$

is differentiable at $t=0^+$, *i.e.*,

$$\mathbf{u}_t \circ T_t = \mathbf{u}_0 + t\dot{\mathbf{u}} + o(t) \tag{4.13}$$

where $\|o(t)\|_{\mathbf{H}(\Omega)}/t \to 0$ *with* $t \downarrow 0$.

The element $\dot{\mathbf{u}} \in \mathbf{H}(\Omega)$ is a unique solution to the variational inequality

$$\begin{aligned} &\dot{\mathbf{u}} \in \mathbf{S}(\Omega), \\ &\mathbf{a}_0(\dot{\mathbf{u}}, \boldsymbol{\zeta}-\dot{\mathbf{u}}) \geqq \langle \mathbf{F}' - B\mathbf{u}_0, \boldsymbol{\zeta}-\dot{\mathbf{u}}\rangle + \mathbf{a}_0(D\mathbf{V}\cdot\mathbf{u}_0, \boldsymbol{\zeta}-\dot{\mathbf{u}}) \quad \forall \boldsymbol{\zeta} \in \mathbf{S}(\Omega) \end{aligned} \tag{4.14}$$

where

$$\begin{aligned} \mathbf{S}(\Omega) = \{\boldsymbol{\zeta} \in \mathbf{H}(\Omega) \,|\, \mathbf{n}\cdot\boldsymbol{\zeta} \leqq \mathbf{n}\cdot D\mathbf{V}\cdot\mathbf{u}_0 \text{ a.e. on } Z(\mathbf{u}_0) \subset \Gamma^c, \\ \mathbf{a}_0(\mathbf{u}_0, \zeta) - \langle \mathbf{F}_0, \boldsymbol{\zeta}\rangle = \mathbf{a}_0(D\mathbf{V}\cdot\mathbf{u}_0, \mathbf{u}_0)\}, \end{aligned} \tag{4.15}$$

$$Z(\mathbf{u}_0) = \{x \in \Gamma^c \,|\, \mathbf{u}_0(x)\cdot\mathbf{n}(x) = 0\}. \tag{4.16}$$

Element $\mathbf{F}' \in (\mathbf{H}(\Omega))'$ and operator $B \in L(\mathbf{H}(\Omega); (\mathbf{H}(\Omega))')$ are defined by (4.55) and (4.56) respectively.

For the proof of Theorem 3 we need some lemmas. First consider sensitivity analysis of the Signorini variational inequality defined in Ω:

$$\begin{aligned} &\mathbf{w}^t \in \mathbf{K}(\Omega), \\ &\mathbf{b}^t(\mathbf{w}^t, \boldsymbol{\zeta}-\mathbf{w}^t) \geqq \langle \mathbf{G}^t, \boldsymbol{\zeta}-\mathbf{w}^t\rangle \quad \forall \boldsymbol{\zeta} \in \mathbf{K}(\Omega) \end{aligned} \tag{4.17}$$

where

$$\mathbf{b}^t(\mathbf{w}, \boldsymbol{\zeta}) = \int_\Omega \boldsymbol{\varepsilon}(\mathbf{w}) \,..\, \mathbf{b}(t) \,..\, \boldsymbol{\varepsilon}(\boldsymbol{\zeta})\, dx, \tag{4.18}$$

$$\langle \mathbf{G}^t, \boldsymbol{\zeta}\rangle = \int_\Omega g_i(t,x)\zeta_i(x)\,dx + \int_{\Gamma^1} G_i(t,x)\zeta_i(x)\,dx, \tag{4.19}$$

$\mathbf{b}(t)$, $t \in [0, \delta)$, is a tensor function with components $b_{ijkl}(t,x)$—we assume that conditions (4.6) and (4.7) are satisfied by $b_{ijkl}(\cdot,\cdot)$ on $[0,\delta)\times\Omega$, $i, j, k, l = 1, 2$. Furthermore

$$b_{ijkl}(\cdot,\cdot) \in C([0,\delta); L^\infty(\Omega)),$$

$$\frac{\partial b_{ijkl}}{\partial t}(0,\cdot) \in L^\infty(\Omega), \qquad i, j, k, l = 1, 2.$$

Elements $g_i \in C([0,\delta); L^2(\Omega))$, $G_i \in C([0,\delta); L^2(\Gamma^1))$ are assumed to be differentiable with respect to t, at point $t = 0^+$ with

$$\frac{\partial g_i}{\partial t}(0,\cdot) \in \mathrm{L}^2(\Omega),$$

$$\frac{\partial G_i}{\partial t}(0,\cdot) \in L^2(\Gamma^1), \qquad i = 1, 2.$$

We also need the following notation. Let $E \in L(\mathbf{H}(\Omega); (\mathbf{H}(\Omega))')$, $\mathbf{G}' \in (\mathbf{H}(\Omega))'$ be defined as follows

$$\langle E\mathbf{w}, \boldsymbol{\zeta}\rangle = \int_\Omega \frac{\partial b_{ijkl}}{\partial t}(0,x)\varepsilon_{ij}(\mathbf{w})(x)\varepsilon_{kl}(\boldsymbol{\zeta})(x)\,dx \quad \forall \mathbf{w}, \boldsymbol{\zeta} \in \mathbf{H}(\Omega), \tag{4.20}$$

$$\langle \mathbf{G}', \boldsymbol{\zeta}\rangle = \int_\Omega \frac{\partial g_i}{\partial t}(0,x)\zeta_i(x)\,dx + \int_{\Gamma^1} \frac{\partial G_i}{\partial t}(0,x)\zeta_i(x)\,d\Gamma \quad \forall \boldsymbol{\zeta} \in \mathbf{H}(\Omega). \tag{4.21}$$

LEMMA 4. *Mapping*

$$[0,\delta)\ni t\to w^t\in \mathbf{H}(\Omega) \tag{4.22}$$

is differentiable at $t=0^+$, *i.e.*,

$$\mathbf{w}^t=\mathbf{w}^0+t\mathbf{r}+o(t), \tag{4.23}$$

where $\|o(t)\|_{\mathbf{H}(\Omega)}/t\to 0$ *with* $t\downarrow 0$ *and the element* $\mathbf{r}\in\mathbf{H}(\Omega)$ *is a unique solution to the variational inequality*

$$\begin{aligned}&\mathbf{r}\in \mathbf{S}_1(\Omega),\\ &\mathbf{b}^0(\mathbf{r},\boldsymbol{\zeta}-\mathbf{r})\geqq\langle \mathbf{G}'-E\mathbf{w}^0,\boldsymbol{\zeta}-\mathbf{r}\rangle\quad \forall\boldsymbol{\zeta}\in\mathbf{S}_1(\Omega)\end{aligned} \tag{4.24}$$

where

$$\mathbf{S}_1(\Omega)=\{\boldsymbol{\zeta}\in\mathbf{H}(\Omega)\,|\,\boldsymbol{\zeta}\cdot\mathbf{n}\leqq 0 \textit{ a.e. on } Z(\mathbf{w}^0)\subset\Gamma^c,\ \mathbf{b}^0(\mathbf{w}^0,\boldsymbol{\zeta})=\langle\mathbf{G}^0,\boldsymbol{\zeta}\rangle\}. \tag{4.25}$$

Proof. We shall apply Theorem 1. Assumptions (i) and (ii) of Theorem 1 are verified by linear operator $E\in L(\mathbf{H}(\Omega);\ (\mathbf{H}(\Omega))')$ and element $\mathbf{G}'\in(\mathbf{H}(\Omega))'$. We have to verify assumption (iii) of Theorem 1.

Assume that the outward unit normal vector on Γ^c has the form

$$\mathbf{n}(x)=(1,0),\qquad x\in\Gamma^c, \tag{4.26}$$

and denote by $R\in L(\mathbf{H}(\Omega);\ L^2(\Gamma^c))$ the linear mapping

$$(R\boldsymbol{\zeta})(x)=\zeta_1(x)=\mathbf{n}(x)\cdot\boldsymbol{\zeta}(x),\qquad x\in\Gamma^c\quad\forall\boldsymbol{\zeta}\in\mathbf{H}(\Omega). \tag{4.27}$$

Denote by $H(\Gamma^c)\subset L^2(\Gamma^c)$ the linear subspace

$$H(\Gamma^c)=\{\eta\in L^2(\Gamma^c)\,|\,\exists\boldsymbol{\zeta}\in\mathbf{H}(\Omega)\text{ such that }R\boldsymbol{\zeta}=\eta\}. \tag{4.28}$$

Define scalar product $((\cdot,\cdot))$ in space $H(\Gamma^c)$

$$((\eta_1,\eta_2))\overset{\text{def}}{=}\mathbf{b}^0(R^{-1}\eta_1,R^{-1}\eta_2)\quad\forall\eta_1,\eta_2\in H(\Gamma^c). \tag{4.29}$$

It can be verified that space $H(\Gamma^c)$ with scalar product (4.29) is a Hilbert space. Consider the following variational inequality posed on Γ^c, given element $\mathbf{h}\in(\mathbf{H}(\Omega))'$, determine $\Phi(\mathbf{h})$:

$$\begin{aligned}&\Phi(\mathbf{h})\in K(\Gamma^c)=\{\eta\in H(\Gamma^c)\,|\,\eta(x)\geqq 0\text{ a.e. on }\Gamma^c\},\\ &((\Phi(\mathbf{h}),\eta-\Phi(\mathbf{h})))\geqq\langle\mathbf{h},R^{-1}(\eta-\Phi(\mathbf{h}))\rangle\quad\forall\eta\in K(\Gamma^c).\end{aligned} \tag{4.30}$$

Denote by $\mathbf{w}^0=P(\mathbf{G}^0)$ the solution of (4.17) for $t=0$. It follows that

$$P(\mathbf{G}^0)=R^{-1}\Phi(\mathbf{G}^0)+Y(\mathbf{G}^0) \tag{4.31}$$

where the element $Y(\mathbf{G}^0)$ solves the variational equation

$$\begin{aligned}&Y(\mathbf{G}^0)\in\mathbf{H}_1(\Omega)=\ker R,\\ &b^0(Y(\mathbf{G}^0),\boldsymbol{\zeta})=\langle\mathbf{G}^0,\boldsymbol{\zeta}\rangle\quad\forall\boldsymbol{\zeta}\in\mathbf{H}_1(\Omega).\end{aligned} \tag{4.32}$$

Assume for the moment that the following conditions are satisfied:

$$\eta^+,\eta^-\in H(\Gamma^c)\quad\forall\eta\in H(\Gamma^c) \tag{A1}$$

where $\eta^+=\max\{0,\eta\}$,

$$((\eta^+,\eta^-))\leqq 0\quad\forall\eta\in H(\Gamma^c), \tag{A2}$$

$$H(\Gamma^c)\cap C_{\text{comp}}(\Gamma^c)\text{ is dense in }C_{\text{comp}}(\Gamma^c). \tag{A3}$$

Then from a result of Mignot [8] it follows that the solution to (4.30) is conically differentiable with respect to $\mathbf{h} \in (\mathbf{H}(\Omega))'$, i.e., $\forall \mathbf{z} \in (\mathbf{H}(\Omega))'$,

$$\Phi(\mathbf{h}+\tau\mathbf{z}) = \Phi(\mathbf{h}) + \tau w(\mathbf{z}) + o(\tau) \tag{4.33}$$

where $\| o(\tau) \|_{\mathbf{H}(\Gamma^c)}/\tau \to 0$ with $\tau \downarrow 0$ and the element $w(\mathbf{z}) \in H(\Gamma^c)$ is a unique solution of the variational inequality

$$\begin{aligned} & w(\mathbf{z}) \in S(\Gamma^c), \\ & ((w(\mathbf{z}), \eta - w(\mathbf{z}))) \geqq \langle \mathbf{z}, R^{-1}(\eta - w(\mathbf{z}))\rangle \quad \forall \eta \in S(\Gamma^c), \end{aligned} \tag{4.34}$$

where

$$S(\Gamma^c) = \{\eta \in H(\Gamma^c) \mid \eta(x) \leqq 0 \text{ a.e. on } Z(\Phi(\mathbf{h})) \subset \Gamma^c, ((\Phi(\mathbf{h}), \eta)) = \langle \mathbf{h}, R^{-1}\eta\rangle\}, \tag{4.35}$$

$$Z(\Phi(\mathbf{h})) = \{x \in \Gamma^c \mid \Phi(\mathbf{h})(x) = 0\}. \tag{4.36}$$

Hence, taking into account (4.31) we obtain $\forall \mathbf{z} \in (\mathbf{H}(\Omega))'$,

$$P(\mathbf{G}^0 + \tau\mathbf{z}) = P(\mathbf{G}^0) + \tau Q(\mathbf{z}) + o(\tau) \tag{4.37}$$

where

$$Q(\mathbf{z}) \overset{\text{def}}{=} R^{-1}w(\mathbf{z}) + Y(\mathbf{z}), \tag{4.38}$$

which implies that assumption (iii) of Theorem 1 is verified. From (1.11) follows (4.23) with

$$\mathbf{r} = Q(\mathbf{G}' - E\mathbf{w}^0).$$

The form (4.25) of cone $S_1(\Omega)$ can be obtained by simple calculations, taking into account (4.34), (4.35) and (4.36). Assumption (4.26) is not restrictive, since for $\Gamma^c \subset C^{1,1}$ the following linear transformation of displacement field $\boldsymbol{\zeta}$ can be defined

$$\begin{gathered} [H^1(\Omega)]^2 \ni \boldsymbol{\psi} \to \boldsymbol{\zeta} \in [H^1(\Omega)]^2, \\ \psi_1 = N_1\zeta_1 + N_2\zeta_2, \\ \psi_2 = -N_2\zeta_1 + N_1\zeta_2 \end{gathered}$$

where $N(\cdot) \in [W^{1,\infty}(\Omega)]^2$ is an extension of normal field $\mathbf{n}(x)$, $x \in \Gamma$. Since

$$\mathbf{n}(x) \cdot \boldsymbol{\zeta}(x) = \psi_1(x), \qquad x \in \Gamma^c,$$

it follows that for displacement field $\boldsymbol{\psi}$ condition (4.26) is satisfied.

For the proof of (A1)–(A3) see the following Lemma 5.

LEMMA 5. *Assumptions* (A1)–(A3) *are satisfied.*

Proof. Assumption (A1) is obviously verified since space $H(\Gamma^c)$ is a closed, linear subspace of Sobolev space $H^{1/2}(\Gamma^c)$. It follows from a general property of Sobolev spaces $H^{1/2}$ [7] that assumption (A3) is satisfied. We prove that (A2) holds.

Recall that

$$(R\boldsymbol{\zeta})(\cdot) = \zeta_1(\cdot) \in H(\Gamma^c) \quad \forall \boldsymbol{\zeta} = (\zeta_1, \zeta_2) \in \mathbf{H}(\Omega).$$

Let $\eta \in H(\Gamma^c)$ be a given element, and denote

$$\boldsymbol{\zeta}^* = R^{-1}\eta, \qquad \zeta^* = (\zeta_1^*, \zeta_2^*). \tag{4.39}$$

We denote $b(\cdot, \cdot) = \mathbf{b}^0(\cdot, \cdot)$.

Note that

$$((\eta, \eta)) = \inf\{b(\boldsymbol{\zeta}, \boldsymbol{\zeta}) : \boldsymbol{\zeta} \in \mathbf{H}(\Omega), R\boldsymbol{\zeta} = \eta\} = b(\boldsymbol{\zeta}^*, \boldsymbol{\zeta}^*). \tag{4.40}$$

Since we have

$$b(\boldsymbol{\zeta},\boldsymbol{\zeta})=b_1(\zeta_1,\zeta_1)+b_2(\zeta_1,\zeta_2)+b_3(\zeta_2,\zeta_1)+b_4(\zeta_2,\zeta_2) \tag{4.41}$$

with appropriate bilinear forms $b_i(\cdot,\cdot)$, $i=1,\cdots,4$, by taking into account the necessary and sufficient optimality conditions for (4.40), we obtain that

$$b(\boldsymbol{\zeta}^*,\boldsymbol{\zeta}^*)=b_1(\zeta_1^*,\zeta_1^*)-b_4(\zeta_2^*,\zeta_2^*). \tag{4.42}$$

Denote

$$\boldsymbol{\psi}^*=R^{-1}|\eta| \tag{4.43}$$

where

$$|\eta|=\eta^++\eta^- \quad \text{for} \quad \eta\in H(\Gamma^c).$$

Furthermore, since $\zeta_i^*\in H^1(\Omega)$, $i=1,2$ it follows that $|\zeta_i^*|\in H^1(\Omega)$, $i=1,2$, where we denote

$$|\boldsymbol{\zeta}^*|\overset{\text{def}}{=}(|\zeta_1^*|,|\zeta_2^*|)\in\mathbf{H}(\Omega). \tag{4.44}$$

Note that

$$|\boldsymbol{\zeta}^*|-\boldsymbol{\psi}^*\in\mathbf{H}_1(\Omega)=\ker R, \tag{4.45}$$

$$\boldsymbol{\psi}^*\in\mathbf{H}_2(\Omega)=(\ker R)^\perp, \tag{4.46}$$

hence

$$b(\boldsymbol{\psi}^*,|\boldsymbol{\zeta}^*|-\boldsymbol{\psi}^*)=0 \tag{4.47}$$

whence

$$\begin{aligned} b(\boldsymbol{\psi}^*,\boldsymbol{\psi}^*)-b(|\boldsymbol{\zeta}^*|,|\boldsymbol{\zeta}^*|)&=2b(\boldsymbol{\psi}^*,\boldsymbol{\psi}^*-|\boldsymbol{\zeta}^*|)-b(\boldsymbol{\psi}^*-|\boldsymbol{\zeta}^*|,\boldsymbol{\psi}^*-|\boldsymbol{\zeta}^*|)\\ &=-b(\boldsymbol{\psi}^*-|\boldsymbol{\zeta}^*|,\boldsymbol{\psi}^*-|\boldsymbol{\zeta}^*|)\leqq 0. \end{aligned} \tag{4.48}$$

By (4.48) it follows that

$$\begin{aligned} ((|\eta|,|\eta|))=b(\boldsymbol{\psi}^*,\boldsymbol{\psi}^*)&\leqq b(|\boldsymbol{\zeta}^*|,|\boldsymbol{\zeta}^*|)\\ &=b_1(|\zeta_1^*|,|\zeta_1^*|)-b_4(|\zeta_2^*|,|\zeta_2^*|)\\ &=b_1(\zeta_1^*,\zeta_1^*)-b_4(\zeta_2^*,\zeta_2^*)\\ &=((\eta,\eta)). \end{aligned} \tag{4.49}$$

The condition $((\eta^+,\eta^-))\leqq 0$, $\forall\eta\in H(\Gamma^c)$ follows from (4.49).

Proof of Theorem 3. Denote

$$\mathbf{z}^t=DT_t^{-1}\cdot\mathbf{u}_t\circ T_t. \tag{4.50}$$

Since we have

$$\boldsymbol{\zeta}_t\in\mathbf{K}(\Omega_t) \quad \text{iff } \boldsymbol{\zeta}=DT_t^{-1}\cdot\zeta_t\circ T_t\in\mathbf{K}(\Omega) \tag{4.51}$$

it follows that the element $\mathbf{z}^t\in\mathbf{H}(\Omega)$ solves the variational inequality

$$\begin{aligned} &\mathbf{z}^t\in\mathbf{K}(\Omega),\\ &\mathbf{a}^t(\mathbf{z}^t,\boldsymbol{\zeta}-\mathbf{z}^t)\geqq\langle\mathbf{F}(t),\boldsymbol{\zeta}-\mathbf{z}^t\rangle \quad \forall\boldsymbol{\zeta}\in\mathbf{K}(\Omega) \end{aligned} \tag{4.52}$$

with appropriate bilinear forms $\mathbf{a}^t(\cdot,\cdot)$ and linear forms $\mathbf{F}(t)\in(\mathbf{H}(\Omega))'$, $t\in[0,\delta)$. It can be verified that a proper choice of bilinear forms $\mathbf{a}^t(\cdot,\cdot)$ and elements $\mathbf{F}(t)$ is the following:

$$\mathbf{a}^t(\mathbf{u},\boldsymbol{\zeta})=\int_\Omega \varepsilon^t(\mathbf{u})\,..\,\mathbf{c}^t\,..\,\boldsymbol{\varepsilon}^t(\boldsymbol{\zeta})\,dx \quad \forall\mathbf{u},\boldsymbol{\zeta}\in\mathbf{H}(\Omega) \tag{4.53}$$

where

$$\boldsymbol{\varepsilon}^t(\boldsymbol{\zeta})=\tfrac{1}{2}\{D(DT_t\cdot\boldsymbol{\zeta})\cdot DT_t^{-1}+{}^*DT_t^{-1}\cdot{}^*(D(DT_t\cdot\boldsymbol{\zeta}))\}$$

and tensor $\mathbf{c}^t$ has components

$$c^t_{ijkl}=\det(DT_t)c_{ijkl}\circ T_t,\qquad i,j,k,l=1,2,\tag{4.54}$$

$$\langle\mathbf{F}(t),\boldsymbol{\zeta}\rangle=\int_\Omega \mathbf{f}^t\cdot\boldsymbol{\zeta}\,dx+\int_\Gamma \mathbf{P}^t\cdot\boldsymbol{\zeta}\,d\Gamma$$

where

$$\mathbf{f}^t=\det(DT_t)^*DT_t\cdot(\mathbf{f}\circ T_t),$$

$$\mathbf{P}^t=\|M(DT_t)\cdot\mathbf{n}\|_{R^2}{}^*DT_t\cdot\mathbf{P}\circ T_t,$$

$$M(DT_t)=\det(DT_t)^*DT_t^{-1}.$$

We need the following notation:

$$\begin{aligned}\langle\mathbf{F}',\boldsymbol{\zeta}\rangle=&\int_\Omega\left\{\sum_{i=1}^{2}(\operatorname{div}(f_i\mathbf{V}(0))\xi_i)+\mathbf{f}\cdot D\mathbf{V}(0)\cdot\boldsymbol{\xi}\right\}dx\\&+\int_{\Gamma_1}\left\{\sum_{i=1}^{2}(\operatorname{div}(P_i\mathbf{V}(0))\xi_i)-(\mathbf{n}\cdot D\mathbf{V}(0)\cdot\mathbf{n})\mathbf{P}\cdot\boldsymbol{\xi}+\mathbf{P}\cdot D\mathbf{V}(0)\cdot\boldsymbol{\xi}\right\}d\Gamma,\end{aligned}\tag{4.55}$$

$$\begin{aligned}\langle B\mathbf{u},\boldsymbol{\zeta}\rangle=&\int_\Omega\boldsymbol{\varepsilon}(\mathbf{u})\,..\,\mathbf{c}'\,..\,\boldsymbol{\varepsilon}(\boldsymbol{\zeta})\,dx+\int_\Omega\boldsymbol{\varepsilon}'(\mathbf{u})\,..\,\mathbf{c}\,..\,\boldsymbol{\varepsilon}(\boldsymbol{\zeta})\,dx\\&+\int_\Omega\boldsymbol{\varepsilon}(\mathbf{u})\,..\,\mathbf{c}\,..\,\boldsymbol{\varepsilon}'(\boldsymbol{\zeta})\,dx\quad\forall\mathbf{u},\boldsymbol{\zeta}\in\mathbf{H}(\Omega)\end{aligned}\tag{4.56}$$

where

$$\boldsymbol{\varepsilon}'(\boldsymbol{\zeta})=\frac{\partial}{\partial t}\varepsilon^t(\boldsymbol{\zeta})\bigg|_{t=0}=\frac{1}{2}\{D(D\mathbf{V}\cdot\boldsymbol{\zeta})+{}^*(D(D\mathbf{V}\cdot\boldsymbol{\zeta}))-D\boldsymbol{\zeta}\cdot D\mathbf{V}-{}^*D\mathbf{V}\cdot{}^*D\boldsymbol{\zeta}\}.$$

Tensor field $c'=(\partial/\partial t)c^t|_{t=0}$ has components

$$c'_{ijkl}=\frac{\partial}{\partial t}c^t_{ijkl}\bigg|_{t=0}=\operatorname{div}\mathbf{V}c_{ijkl}+\operatorname{grad}c_{ijkl}\cdot\mathbf{V}.$$

It can be verified that element $F'\in(\mathbf{H}(\Omega))'$ and operator $B\in L(\mathbf{H}(\Omega);(\mathbf{H}(\Omega))')$ satisfy

$$\langle\mathbf{F}',\boldsymbol{\zeta}\rangle=\frac{d}{dt}\langle\mathbf{F}(t),\boldsymbol{\zeta}\rangle\bigg|_{t=0^+}\quad\forall\boldsymbol{\zeta}\in\mathbf{H}(\Omega),$$

$$\langle B\mathbf{u},\boldsymbol{\zeta}\rangle=\frac{d}{dt}\mathbf{a}^t(\mathbf{u},\boldsymbol{\zeta})\bigg|_{t=0^+}\quad\forall\mathbf{u},\boldsymbol{\zeta}\in\mathbf{H}(\Omega).$$

Furthermore conditions corresponding to (2.3) and (2.2) are verified. We can apply Lemma 4 to (4.52), hence the mapping

$$[0,\delta)t\to\mathbf{z}^t\in\mathbf{H}(\Omega)\tag{4.57}$$

is differentiable at $t=0^+$, i.e.

$$\mathbf{z}^t=\mathbf{z}^0+t\partial\mathbf{z}+o(t) \tag{4.58}$$

where $\|o(t)\|_{H(\Omega)}/t\to 0$ with $t\downarrow 0$,

$$\begin{gathered}\partial\mathbf{z}\in\mathbf{S}_2(\Omega),\\ a^0(\partial\mathbf{z},\boldsymbol{\zeta}-\partial\mathbf{z})\geqq\langle\mathbf{F}'-B\mathbf{z}^0,\boldsymbol{\zeta}-\partial\mathbf{z}\rangle\quad\forall\boldsymbol{\zeta}\in\mathbf{S}_2(\Omega),\end{gathered} \tag{4.59}$$

$$\mathbf{S}_2(\Omega)=\{\boldsymbol{\zeta}\in\mathbf{H}(\Omega)\,|\,\boldsymbol{\zeta}\cdot\mathbf{n}\leqq 0 \text{ a.e. on } Z(\mathbf{z}^0),\, a^0(\mathbf{z}^0,\boldsymbol{\zeta})=\langle\mathbf{F}(0),\mathbf{z}\rangle\},$$

$$Z(\mathbf{z}^0)=\{x\in\Gamma^c\,|\,\mathbf{z}^0(x)\cdot\mathbf{n}(x)=0\}.$$

On the other hand, since

$$\mathbf{u}_t\circ T_t=DT_t\cdot\mathbf{z}^t \quad\text{and}\quad DT_t|_{t=0}=I, \tag{4.60}$$

it follows that

$$\begin{aligned}\dot{\mathbf{u}}&=\partial\mathbf{z}-DV\cdot\mathbf{z}^0\\ &=\partial\mathbf{z}-DV\cdot\mathbf{u}_0\end{aligned} \tag{4.61}$$

and simple calculations show that $\dot{\mathbf{u}}$ is determined by (4.14).

We use the material derivative $\dot{\mathbf{u}}'$ in order to derive the form of the so-called Eulerian semiderivative $dJ(\Omega;\mathbf{V})$ of a functional $J(\Omega)$ in the direction of a vector field $\mathbf{V}(\cdot,\cdot)$.

Let us consider a functional

$$J(\Omega)=\int_\Omega F(\mathbf{u}(x))\,dx$$

where $F(\cdot)\in C^1(R^2)$ and $\mathbf{u}(\cdot)$ is given by (4.11) for $t=0$.

We denote

$$dJ(\Omega;\mathbf{V})\overset{\text{def}}{=}\lim_{t\downarrow 0}(J(\Omega_t)-J(\Omega))/t.$$

It can be verified that

$$dJ(\Omega;\mathbf{V})=\int_\Omega (DF(\mathbf{u}(x))\cdot\dot{\mathbf{u}}(x)+F(\mathbf{u}(x))\operatorname{div}\mathbf{V}(0,x))\,dx.$$

Let us assume now, that the following regularity condition is verified

$$D\mathbf{u}\cdot\mathbf{V}(0)\in(H^1(\Omega))^2;$$

then there exists [17] the so-called domain derivative for problem (4.11) given by

$$\mathbf{u}'=\dot{\mathbf{u}}-D\mathbf{u}\cdot V(0).$$

Furthermore,

$$dJ(\Omega;\mathbf{V})=\int_\Omega DF(\mathbf{u}(x))\cdot\mathbf{u}'(x)\,dx+\int_{\partial\Omega}F(\mathbf{u}(x))\langle\mathbf{V}(0,x),\mathbf{n}(x)\rangle\,d\Gamma.$$

The form of the domain derivative $\mathbf{u}'$ is given in [14], [17] and [19].

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris 1978.

[2] J. CEA, *Une méthode numérique pour la recherche d'un domaine optimal*, Publication IMAN de l'Université de Nice, 1976.

[3] ———, *Problems of shape optimal design*, in Optimization of Distributed Parameter Structures, E. J. Haug and J. Cea, eds., Sijthoff and Nordhoff, Alphen aan den Rijn, The Netherlands, 1981.

[4] A. DERVIEUX, *Résolution de problèmes à frontière libre*, Thèse d'état, L'Université Paris 6, 1981.

[5] G. DUVANT AND J. L. LIONS, *Les inéquations en méchanique et en physique*, Dunod, Paris, 1972.

[6] G. FICHERA, *Boundary Value Problems of Elasticity with Unilateral Constraints*, Handbuch der Physik, Band 6a/2 Springer-Verlag, Berlin, New York, Heidelberg, 1972.

[7] J. L. LIONS AND E. MAGENES, *Problémes aux limites non homogènes et applications, Vol.* 1, Dunod, Paris, 1968.

[8] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130-185.

[9] F. MURAT AND J. SIMON, *Sur le contrôle par un domaine géométrique*, Publication de l'Université Paris 6 No. 76015, 1977.

[10] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, New York, Heidelberg, 1983.

[11] J. SOKOŁOWSKI, *Optimal control in coefficients of boundary value problems with unilateral Constraints*, Bulletin of the Polish Academy of Sciences, Technical Sciences, Vol. 31, Nos. 1-12 (1983), pp. 71-81.

[12] ———, *Sensitivity analysis for a class of variational inequalities*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1981, pp. 1600-1609.

[13] J. SOKOŁOWSKI AND J. P. ZOLESIO, *Shape sensitivity analysis for variational inequalities*, Lecture Notes in Control and Information Sciences, Vol. 38, System Modelling and Optimization, R. F. Drenick and F. Kozin, eds., Springer-Verlag, Berlin, New York, Heidelberg, 1982, pp. 401-407.

[14] ———, *Derivation par rapport au domaine dans les problèmes unilateraux*, INRIA, Rapport de Recherche No 132, 1982.

[15] J. P. ZOLESIO, *Identification de domaines par déformations*, Thèse d'Etat, l'Université de Nice, 1979.

[16] ———, *The material derivative/or speed/method for shape optimization*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1981.

[17] J. SOKOLOWSKI AND J. P. ZOLESIO, *Shape sensitivity analysis of elastic structures*, The Danish Center for Applied Mathematics and Mechanics, Report No. 289, July 1984.

[18] ———, C.R. Acad. Sci. Paris, Sér. I., 301 (1985).

[19] ———, *Shape sensitivity analysis of plates and plane elastic solids under unilateral constraints*, J. Optim. Theory Appl., 54 (1987).

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

017

THE INITIAL VELOCITY OF THE EMERGING FREE BOUNDARY IN A TWO-PHASE STEFAN PROBLEM WITH IMPOSED FLUX*

A. SOLOMON†, V. ALEXIADES†‡ AND D. G. WILSON†

Abstract. Consider the one-dimensional, two-phase Stefan problem with an imposed flux, in which the phase change front originates at the material boundary. We prove that, subject to suitable assumptions on the initial temperature and imposed flux, the initial speed of the phase change front is equal to the jump between the surface and initial fluxes at the boundary. Similarly, we prove that the front $X(t)$ is continuously differentiable in the closed interval $[0, t^*]$ for some $t^* > 0$.

Key words. two-phase Stefan problem, initial interface speed, onset of melting, L^2-estimates, smoothness of free boundary, maximum principle

AMS(MOS) subject classifications. 35R35, 35K05, 35B65

Introduction. Consider a slab of material occupying the interval $0 \leqq x \leqq L$ which is initially totally solid, with a known temperature distribution (below the critical temperature T_{cr} of the material). Heat is input at $x = 0$ via a known heat flux $q(t) > 0$, while the face $x = L$ is kept insulated. Then at some time t_0 a melt front $x = X(t)$ will emerge from $x = 0$ and move into the material, separating liquid $(x < X(t))$ from solid $(x > X(t))$. The location $X(t)$ of the front as well as the temperature $T(x, t)$ are to be determined.

The mathematical problem here is a one-dimensional two-phase Stefan problem which, however, involves the appearance of a new phase. The onset of melting is a major source of difficulties for both the physical understanding of the phenomenon as well as the mathematical analysis of the problem.

One of the key problems associated with the onset of melting is determining the initial velocity of the melt front, which is the objective of this study. The possibilities range from no melting at all (Solomon, Alexiades and Wilson [14]) to initially infinite propagation speed (when a temperature jump is imposed at $x = 0$ (Carslaw and Jaeger [2])). Hence, finding conditions on the data that guarantee melting is also useful.

The problem we described in the beginning is typical of problems arising in diverse areas of applications ranging from latent heat thermal energy storage (Solomon [10]) to crystal growth (Rubinstein [9]). The need to know the initial interface speed also appears in the study of phase-change problems by either approximation methods (Solomon [11]) or numerical methods (Solomon [12]). In particular, front tracking numerical schemes for an emerging front require the initial front speed to be known (Landau [7]).

Returning to the problem at hand, we note that it consists of two distinct parts: a pure heat conduction problem until melting begins, and afterwards a Stefan problem which is the one of interest. Taking as initial time $(t = 0)$ the instant at which the front emerges, we are interested in finding $X'(0)$ when we know the input flux $q(t) > 0$ and the initial temperature $T(x, 0) = f(x) \leqq T_{cr}$. This temperature $f(x)$ is the result of the premelting heat conduction problem and as such it has certain natural properties like

* Received by the editors April 3, 1985; accepted for publication June 19, 1986. This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC05-840R21400 with Martin Marietta Energy Systems, Inc.

† Mathematics and Statistics Research, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831.

‡ Department of Mathematics, The University of Tennessee, Knoxville, Tennessee 37996.

$f(0)=T_{cr}, f'(x)\leqq 0, f''(x)\geqq 0$, which we take as assumptions for the data of the melting problem (see § 1).

At first glance, it appears reasonable that the initial speed should be determined by the jump in the flux at the surface $x=0$ at $t=0$, namely

$$\rho HX'(0)=q(0)-[-k_s f'(0)] \tag{0.1}$$

(see § 1 for notation). In fact, (0.1) is merely the limiting value, as $t\downarrow 0$, of the Stefan condition along the interface. Also, (0.1) is obeyed by a number of asymptotic results for the first and third boundary conditions in Rubinstein [9] and for a problem with internal heating and $q(0)=-k_s f'(0)=0$ in Lacey and Shillor [6]. Thus, the question of whether or not (0.1) is valid amounts to whether or not the fluxes in the liquid and in the solid are continuous down to $t=0$. Unfortunately, the existing literature on the two-phase problem with imposed flux, for an initially one-phase state, does not answer this smoothness question. Indeed, despite the very extensive literature on Stefan-type problems (Wilson, Solomon and Trent [15], Niezgodka [8]), very few papers actually allow a new phase to appear and we are not aware of any works establishing the well-posedness of the classical formulation of our problem. Nevertheless, the methods of Cannon and Primicerio [1] and of Ishii [5], dealing with Dirichlet data on $x=0$, can also be applied to the case of imposed flux and thus answer the well-posedness question. Furthermore, under certain restrictions on the flux $q(t)$, the work of Fasano and Primicerio [3] also applies.

In this paper we prove the continuity of the fluxes down to $t=0$ under realistic assumptions on the data $f(x)$ and $q(t)$. It then follows that $X(t)$ is differentiable down to $t=0$ and (0.1) holds. These results and the notations are stated in § 1. The steps necessary to complete the proof are divided into four parts. Crucial bounds on the interface location are derived in § 2, and needed L^2-estimates of T_x and T_{xx} are derived in § 3. In § 4 we generalize a version of the maximum principle (Solomon, Alexiades and Wilson [13]) to the two-phase problem and establish the global boundedness of T_x. The proof of the main results is finally completed in § 5.

1. Notation and main results. Consider the following one-dimensional two-phase Stefan problem with imposed flux:

For a given $t_\infty>0$, find a function $X(\cdot)\in C[0,t_\infty]\cap C^1(0,t_\infty)$, $X(0)=0$, and a function $T(\cdot,\cdot)\in C(\bar{D}_l\cup\bar{D}_s)$, where

$$D_l=\{(x,t)\colon 0<x<X(t), 0<t<t_\infty\}\quad \text{(liquid)},$$

$$D_s=\{(x,t)\colon X(t)<x<L, 0<t<t_\infty\}\quad \text{(solid)}$$

such that: for each $t\in(0,t_\infty)$, $T_x(\cdot,t)$ is continuous on $[0,X(t))$ and $(X(t),L]$ and on each of the closed intervals can be extended to a continuous function with a possible jump discontinuity at $X(t)$; T_t, T_{xx} are continuous in D_l and D_s;

$$T_t=\alpha_i T_{xx}\quad \text{in } D_i,\quad i=l,s, \tag{1.1}$$

$$T(x,0)=f(x),\qquad 0\leqq x\leqq L, \tag{1.2}$$

$$-k_l T_x(0,t)=q(t),\qquad 0<t<t_\infty, \tag{1.3}$$

$$T_x(L,t)=0,\qquad 0<t<t_\infty, \tag{1.4}$$

$$T(X(t),t)=T_{cr},\qquad 0\leqq t\leqq t_\infty, \tag{1.5}$$

$$\rho HX'(t)=-k_l T_x^-[t]+k_s T_x^+[t],\qquad 0<t<t_\infty. \tag{1.6}$$

Here, $T_x^{\pm}[t] := \lim_{x\to X(t)\pm} T_x(x,t)$; $k=k_l$ or k_s is the conductivity; $\alpha=\alpha_l$ or α_s is the diffusivity, $\alpha_i = k_i/\rho c_i$, $i=l,\ s$, ρ being the density and c_l, c_s the specific heats; H is the latent heat, T_{cr} the critical temperature and L the slab thickness. All these thermophysical parameters are positive constants.

On the data $q(t)$ and $f(x)$ we assume

$$q\in C^1[0,t_\infty),\quad q(t)\geqq 0,\quad 0\leqq t\leqq t_\infty, \tag{1.7}$$

$$f\in C^2[0,L]\cap H^3(0,L), \tag{1.8}$$

$$f(x)\leqq T_{cr},\quad f'(x)\leqq 0,\quad f''(x)\geqq 0\quad \text{for } 0\leqq x\leqq L, \tag{1.9}$$

$$f(0)=T_{cr},\qquad f'(L)=0. \tag{1.10}$$

As mentioned in the Introduction, the assumptions on $f(x)$ are natural for the temperature of an initially cold solid, insulated at $x=L$, and heated by a smooth flux at $x=0$. In fact, in that case f will also satisfy

$$-k_s f'(0)=q(0). \tag{1.11}$$

The well-posedness of this Stefan problem we take as known (see the Introduction). Precisely, we assume there exists a solution $X(t)$, $T(x,t)$ with

$$X\in C^{\mu}[0,t_\infty]\cap C^\infty(0,t_\infty)\quad \text{for some } 0<\mu<1, \tag{1.12}$$

$$T\in C(\overline{D_l\cup D_s})\cap C^\infty(D_l\cup D_s). \tag{1.13}$$

The paper is devoted to proving the following.

MAIN THEOREM 1.1. *Under the assumptions* (1.7)-(1.10), *the solution of* (1.1)-(1.6) *has the following properties*:

T_x *is continuous on each of* D_l *and* D_s *and has a continuous extension to each of* $\bar{D}_l$ *and* $\bar{D}_s$; *these continuous extensions differ along the line* $\{(t,x)\,|\,x=X(t)\}$ *by a finite jump*;

$$X\in C^1[0,t_\infty),$$

$$\rho H X'(0)=q(0)+k_s f'(0). \tag{1.14}$$

COROLLARY 1.2. *At the onset of melting of an initially solid slab* $[0,L]$, *insulated at* $x=L$ *and heated by a flux* $q(t)\geqq 0$, *the initial velocity of the phase-change front is zero.*

2. Bounds on the interface. We begin with the derivation of bounds on the interface location.

By the maximum principle (Friedman [4]), the assumptions $q(t)\geqq 0$, $f(x)\leqq T_{cr}$, and $f''(x)\leqq 0$ imply the following.

LEMMA 2.1. $f(x)\leqq T(x,t)\leqq T_{cr}$ *in the solid* D_s *and* $T(x,t)\geqq T_{cr}$ *in the liquid* D_l. *Hence*

$$T_x^-[t]\equiv T_x(X(t)-,t)\leqq 0$$

and

$$T_x^+[t]\equiv T_x(X(t)+,t)\leqq 0,\qquad 0<t\leqq t_\infty.$$

Proof. Let $z(x,t)=T(x,t)-f(x)$ in $\bar{D}_s$. Then $z\in C(\bar{D}_s)$ satisfies $z_t-\alpha_l z_{xx}=\alpha_l f''(x)\leqq 0$ in D_s; $z(X(t),t)=T_{cr}-f(X(t))\geqq 0$, $z_x(L,t)=0$, $0\leqq t\leqq t_\infty$; $z(x,0)=0$, $0\leqq x\leqq L$. Hence its minimum value must be $\geqq 0$. The remaining inequalities are proved similarly.

Next we obtain a crucial upper bound for the front location.

THEOREM 2.2. *There exists a constant* $M_0>0$ *such that*

$$X(t)\leqq M_0 t,\qquad 0<t\leqq t_\infty. \tag{2.1}$$

Proof. For any $0<t\leqq t_\infty$, using (1.1)-(1.6) we have

$$\begin{aligned}\frac{d}{dt}\int_0^L \rho c[T(x,t)-T_{cr}]\,dx&=\frac{d}{dt}\int_0^{X(t)}\rho c_l[T-T_{cr}]\,dx+\frac{d}{dt}\int_{X(t)}^L \rho c_s[T-T_{cr}]\,dx\\&=k_lT_x^-[t]-k_lT_x(0,t)+k_sT_x(L,t)-k_sT_x^+[t]\\&=-\rho HX'(t)+q(t).\end{aligned}$$

Integrating over (τ, t) with $0\leqq\tau<t\leqq t_\infty$, we obtain

$$\int_0^L \rho c[T(x,t)-T_{cr}]\,dx-\int_0^L \rho c[T(x,\tau)-T_{cr}]\,dx+\rho H[X(t)-X(\tau)]=\int_\tau^t q(s)\,ds.$$

Taking $\tau\to 0$, using $X(0)=0$ and (1.2), and splitting $\int_0^L$ into $\int_0^{X(t)}+\int_{X(t)}^L$, we obtain the following heat balance:

$$\begin{aligned}(2.2)\qquad &\int_0^{X(t)}\rho c_l[T-T_{cr}]\,dx+\int_{X(t)}^L \rho c_s[T-f(x)]\,dx+\rho HX(t)\\&=\int_0^t q(s)\,ds+\int_0^{X(t)}\rho c_l[f(x)-T_{cr}]\,dx.\end{aligned}$$

By Lemma 2.1, the first two terms on the left are nonnegative and the last term on the right is nonpositive. We conclude that

$$\rho HX(t)\leqq\int_0^t q(s)\,ds\leqq q_{\max}\cdot t,$$

where $q_{\max}:=\max q(t)$ over $0\leqq t\leqq t_\infty$. This establishes (2.1) with $M_0:=q_{\max}/\rho H$.

To obtain a lower bound on $X(t)$, we compare T with the solution of the following heat conduction problem without phase change:

$$\begin{aligned}&w_t=\alpha_l w_{xx}, && x>0,\quad t>0,\\&w(x,0)\equiv T_{cr}, && x>0,\\&-k_lw_x(0,t)=q(t)\geqq 0, && t>0,\end{aligned}$$

the solution of which is given explicitly (Carslaw and Jaeger [2, p. 76]) by

$$(2.3)\qquad w(x,t)=T_{cr}+k_l^{-l}\left(\frac{\alpha_l}{\pi}\right)^{1/2}\int_0^t q(t-s)\exp\{-x^2/4\alpha_l(t-s)\}s^{-1/2}\,ds.$$

It is easy to see that $T(x,t)\leqq w(x,t)$ in $D_l\cup D_s$ and that $w(x,t)\geqq T_{cr}\geqq f(x)$, $0\leqq x\leqq L$, $t>0$. Hence, (2.2) implies the following.

PROPOSITION 2.3.

$$\rho HX(t)\geqq\int_0^t q(s)\,ds-\rho c_{\max}\int_0^L [w(x,t)-f(x)]\,dx,\qquad 0<t<t_\infty,$$

where $c_{\max}:=\max\{c_l, c_s\}$ *and* $w(x,t)$ *is given by* (2.3).

This says that melting may not necessarily begin immediately, but since the right-hand side will eventually become positive, the liquid will have to expand with increasing time.

Another useful relation obtainable from the heat balance (2.2) is

$$\rho c_{\min} \int_0^t [T(x,t)-f(x)]\,dx \leqq \int_0^t q(s)\,ds, \qquad 0<t\leqq t_\infty,$$

where $c_{\min} := \min\{c_l, c_s\}$.

3. Estimates on T_x and T_{xx}.

Lemma 3.1. *For any* $0<t\leqq t_\infty$,

$$0 \leqq \int_0^t -k_l T_x^-[s]\,ds \leqq \int_0^t q(s)\,ds. \tag{3.1}$$

Proof. This is obtained from the heat balance in the liquid as follows:

$$\frac{d}{dt}\int_0^{X(t)} \rho c_l [T(x,t)-T_{cr}]\,dx = k_l T_x^-[t] + q(t).$$

Integrating over $(0, t)$ and since $T \geqq T_{cr}$ (see Lemma 2.1), we have

$$\int_0^t q(s)\,ds + \int_0^t k_l T_x^-[s]\,ds = \int_0^{X(t)} \rho c_l[T-T_{cr}]\,dx \geqq 0.$$

Theorem 3.2. *There exists a constant* $M_1>0$, *independent of time, such that*

$$\int_0^t \int_0^L |T_x(x,s)-f'(x)|^2\,dx\,ds \leqq M_1 \cdot t^2, \qquad 0<t\leqq t_\infty. \tag{3.2}$$

Proof. For $t>0$,

$$\begin{aligned} \frac{d}{dt}\int_0^L \frac{1}{2}\rho c[T(x,t)-f(x)]^2\,dx &= \frac{1}{2}\rho(c_l-c_s)[T_{cr}-f(X)]^2 X'(t) \\ &\quad + \int_0^L k[T-f(x)][T-f(x)]_{xx}\,dx \\ &\quad + \int_0^L k[T-f(x)]f''(x)\,dx. \end{aligned} \tag{3.3}$$

By integration by parts, the first integral in the right-hand side becomes

$$\begin{aligned} \int_0^L k[T-f][T_x-f']_x\,dx = \int_0^L k([T-f][T_x-f'])_x - k[T_x-f']^2 \\ = k_l[T_{cr}-f(X)][T_x^-[t]-f'(X)] \\ - k_l[T(0,t)-f(0)][T_x(0,t)-f'(0)] \\ + k_s[T(L,t)-f(L)][T_x(L,t)-f'(L)] \\ - k_s[T_{cr}-f(X)][T_{cr}-f(X)][T_x^+[t]-f'(X)] \\ - \int_0^L k[T_x-f')]^2\,dx \end{aligned}$$

and using (1.6), (1.10), (1.3) and $f'(x)\leqq 0$,

$$=[f(0)-f(X)]\{-\rho HX'(t)-(k_l-k_s)f'(X)\}$$
$$+[T(0,t)-T_{cr}][q(t)+k_lf'(0)]-\int_0^L k[T_x-f']^2\,dx.$$

Substituting this into (3.3) and integrating over (τ, t), $0<\tau<t\leqq t_\infty$, we find

$$\int_0^L \frac{\rho c}{2}[T(x,t)-f(x)]^2\,dx-\int_0^L \frac{\rho c}{2}[T(x,\tau)-f(x)]^2\,dx+\int_\tau^t\int_0^L k[T_x-f']^2\,dx$$

$$\begin{aligned} &=\int_\tau^t \frac{\rho(c_l-c_s)}{2}X'(s)[f(0)-f(X(s))]^2\,ds \\ &\quad-\int_\tau^t \{\rho HX'(s)+(k_l-k_s)f'(X(s))\}[f(0)-f(X(s))]\,ds \\ &\quad+\int_\tau^t [T(0,s)-T_{cr}][q(s)+k_lf'(0)]\,ds \\ &\quad+\int_\tau^t\int_0^L k[T-f(x)]f''(x)\,dx\,ds. \end{aligned} \tag{3.4}$$

Now, applying (2.1) and (1.7)-(1.10), we estimate each term separately as follows.

$$\left|\int_\tau^t \frac{\rho(c_l-c_s)}{2}X'(s)[f(0)-f(X(s))]^2\,ds\right|=\left|\int_{X(\tau)}^{X(t)}\frac{\rho(c_l-c_s)}{2}[f(0)-f(\xi)]^2\,d\xi\right|$$
$$\leqq \rho c_{\max}|f'_{\max}|^2\frac{X(t)^3-X(\tau)^3}{3}\leqq M_{12}\cdot t^3,$$

$$\left|\int_\tau^t \rho HX'(s)[f(X(s))-f(0)]\,ds\right|=\left|\int_{X(\tau)}^{X(s)}\rho H[f(\xi)-f(0)]\,d\xi\right|$$
$$\leqq \rho H|f'_{\max}|\frac{X(t)^2}{2}\leqq M_{13}\cdot t^2,$$

$$\begin{aligned} &\left|\int_\tau^t (k_l-k_s)f'(X(s))[f(0)-f(X(s))]\,ds\right| \\ &\leqq k_{\max}|f'_{\max}|^2\int_\tau^t X(s)\,ds \\ &\leqq M_{14}\cdot t^2, \end{aligned}$$

$$\begin{aligned} &\left|\int_\tau^t [T(0,s)-T_{cr}][q(s)+k_lf'(0)]\,ds\right| \\ &=\left|\int_\tau^t\left\{\int_0^{X(s)}[T_x(x,s)-f'(x)]\,dx+[f(0)-f(X(s))]\right\}[q(s)+k_lf'(0)]\,ds\right| \\ &\leqq[q_{\max}+k_l|f'(0)|]\left\{\int_\tau^t X(s)^{1/2}\int_0^{X(s)}X(s)^{-1/2}|T_x(x,s)-f'(x)|\,dx\,ds\right. \\ &\qquad\left.+\int_\tau^t|f(0)-f(X(s))|\,ds\right\} \end{aligned}$$

$$\leqq M_{15}\left\{\int_\tau^t\left[\frac{X(s)}{2\varepsilon}+\frac{\varepsilon}{2X(s)}\left(\int_0^{X(s)}1\cdot|T_x(x,s)-f'(x)|\,dx\right)^2\right]ds+|f'_{\max}|\int_\tau^t X(s)\,ds\right\},\qquad \varepsilon>0, \tag{3.5}$$

$$\leqq \frac{M_{16}}{\varepsilon} t^2 + \frac{\varepsilon M_{15}}{2} \int_\tau^t \int_0^L |T_x - f'|^2 \, dx \, ds,$$

$$0 \leqq \int_\tau^t \int_0^L k[T(x,s) - f(x)] f''(x) \, dx \, ds \leqq f''_{\max} \int_\tau^t \int_0^L k[T(x,s) - T(x,0)] \, dx \, ds$$

$$\leqq k_{\max} f''_{\max} \int_\tau^t \int_0^L \left\{ \int_0^s T_t(x,\sigma) \, d\sigma \right\} dx \, ds = M_{17} \int_\tau^t \int_0^s \int_0^L T_t(x,\sigma) \, dx \, d\sigma \, ds$$

$$= M_{17} \int_\tau^t \int_0^s \left\{ \frac{1}{\rho c_l} \int_0^{X(\sigma)} k_l T_{xx}(x,\sigma) \, dx + \frac{1}{\rho c_s} \int_{X(\sigma)}^L k_s T_{xx}(x,\sigma) \, dx \right\} d\sigma \, ds$$

$$= M_{18} \int_\tau^t \int_0^s \{-\rho H X'(\sigma) + q(\sigma)\} \, d\sigma \, ds \leqq 0 + M_{18} q_{\max} \frac{t^2}{2} = M_{19} \cdot t^2.$$

We incorporate these estimates into the right-hand side of (3.4), take $\tau \downarrow 0$ (which makes the negative term on the left-hand side vanish) and discard the first (positive) term. We thus obtain

$$\left[k_{\min} - \frac{\varepsilon M_{15}}{2} \right] \int_0^t \int_0^L [T_x - f']^2 \, dx \leqq M_{12} t^3 + M_{13} t^2 + M_{14} t^2 + \frac{M_{16}}{\varepsilon} t^2 + M_{19} t^2.$$

Choosing $\varepsilon M_{15}/2 = \frac{1}{2} k_{\min}$, we finally establish (3.2). Note that the constant M_1 is essentially known explicitly.

By the Mean Value Theorem we immediately obtain the following.

COROLLARY 3.3. *For each $t \in (0, t_\infty)$, there exists $t^* \in (t/2, t)$ such that*

$$\int_0^L |T_x(x, t^*) - f'(x)|^2 \, dx \leqq M_2 t. \tag{3.6}$$

Next we estimate the second derivative.

THEOREM 3.4. *There exists a constant $M_3 > 0$, independent of time, such that*

$$\int_0^t \int_0^L |T_{xx}(x,s) - f''(x)|^2 \, dx \, ds \leqq M_3 \cdot t, \qquad 0 < t < t_\infty. \tag{3.7}$$

Proof. We begin with

$$\begin{aligned} \frac{d}{dt} \int_0^L \frac{k^2}{2} [T_x(x,t) - f'(x)]^2 \, dx &= \frac{X'(t)}{2} \{ k_l^2 [T_x^-[t] - f'(X(t))]^2 - k_s^2 [T_x^+[t] - f'(X(t))]^2 \} \\ &\quad + \int_0^L k^2 [T_x(x,t) - f'(x)] \alpha T_{xxx}(x,t) \, dx. \end{aligned} \tag{3.8}$$

The last integral can be rewritten as

$$\int_0^L \alpha k^2 [T_x - f'][T_{xxx} - f'''] \, dx + \int_0^L \alpha k^2 [T_x - f'] f''' \, dx$$

$$= \int_0^L \alpha k^2 \{ [(T_x - f')(T_{xx} - f'')]_x - [T_{xx} - f'']^2 + f'''[T_x - f'] \} \, dx$$

$$= \alpha_l k_l^2 [T_x^-[t] - f'(X)][T_{xx}^-[t] - f''(X)] - \alpha_l k_l^2 [T_x(0,t) - f'(0)][T_{xx}(0,t) - f''(0)]$$

$$+0-\alpha_s k_s^2[T_x^+[t]-f'(X)][T_{xx}^+[t]-f'(X)]$$
$$-\int_0^L \alpha k^2\{[T_{xx}-f'']^2-f'''[T_x-f']\}\,dx.$$

By (1.12), (1.13) the front is C^∞ for $t>0$ and the heat equation is valid along it from each side. Moreover, $T(X(t),t)=T_{cr}$ implies $T_t^\pm+X'(t)T_x^\pm=0$, so we have

$$\alpha T_{xx}^\pm[t]=T_t^\pm=-X'(t)T_x^\pm[t], \qquad t>0.$$

Then (3.8) becomes

$$\begin{aligned}
&\frac{d}{dt}\int_0^L \frac{k^2}{2}[T_x-f']^2\,dx\\
&\quad=\frac{X'(t)}{2}\{k_l^2[T_x^-[t]-f'(X)]^2-k_s^2[T_x^+[t]-f'(X)]^2\}\\
&\qquad+k_l^2[T_x^--f'(X)][-X'T_x^--\alpha_l f''(X)]\\
&\qquad-k_l^2[T_x(0,t)-f'(0)][T_t(0,t)-\alpha_l f''(0)]\\
&\qquad-k_s^2[T_x^+-f'(X)][-X'T_x^+-\alpha_s f''(X)]\\
&\qquad-\int_0^L \alpha k^2[T_{xx}-f'']^2\,dx+\int_0^L \alpha k^2 f'''[T_x-f']\,dx\\
&\quad=\frac{X'}{2}(k_s^2T_x^{+2}-k_l^2T_x^{-2})+\frac{1}{2}(k_l^2-k_s^2)X'f'(X)^2+(k_s^2\alpha_sT_x^+-k_l^2\alpha_lT_x^-)f''(X)\\
&\qquad+(k_l^2\alpha_l-k_s^2\alpha_s)f'(X)f''(X)+k_l[q(t)+k_lf'(0)][T_t(0,t)-\alpha_lf''(0)]\\
&\qquad-\int_0^L \alpha k^2[T_{xx}-f'']^2\,dx+\int_0^L \alpha k^2 f'''[T_x-f']\,dx.
\end{aligned}$$

Integrating over (τ,t), $0<\tau<t<t_\infty$, we obtain

$$\begin{aligned}
&\int_0^L \frac{k^2}{2}|T_x(x,t)-f'(x)|^2\,dx-\int_0^L \frac{k^2}{2}|T_x(x,\tau)-f'(x)|^2\,dx\\
&\qquad+\int_\tau^t\int_0^L \alpha k^2|T_{xx}-f''|^2\,dx\,ds\\
&\quad=\frac{1}{2}\int_\tau^t X'(s)(k_s^2T_x^+[s]^2-k_l^2T_x^-[s]^2)\,ds+\frac{1}{2}(k_l^2-k_s^2)\int_\tau^t X'(s)f'(X(s))^2\,ds\\
&\qquad+\int_\tau^t (k_s^2\alpha_sT_x^+[s]-k_l^2\alpha_lT_x^-[s])f''(X(s))\,ds\\
&\qquad+(k_l^2\alpha_l-k_s^2\alpha_s)\int_\tau^t f'(X(s))f''(X(s))\,ds\\
&\qquad+k_l\int_\tau^t [q(s)+k_lf'(0)][T_t(0,s)-\alpha_lf''(0)]\,ds\\
&\qquad+\int_\tau^t\int_0^L \alpha k^2f'''(x)[T_x(x,s)-f'(x)]\,dx\,ds.
\end{aligned} \tag{3.9}$$

We now estimate each term separately as follows:

$$\frac{1}{2}\int_\tau^t X'(s)(k_sT_x^+-k_lT_x^-)(k_sT_x^++k_lT_x^-)\,ds=\frac{1}{2}\int_\tau^t \rho HX'(s)^2(k_sT_x^++k_lT_x^-)\,ds\leqq 0,$$

because of Lemma 2.1;

$$\int_\tau^t f'(X(s))^2 X'(s)\,ds=\int_{X(\tau)}^{X(t)} f'(\xi)^2\,d\xi \leqq |f'_{\max}|^2 X(t)\leqq M_{31}\cdot t,$$

$$\int_\tau^t (k_s^2\alpha_s T_x^+ - k_l^2\alpha_l T_x^-) f''(X)\,ds \leqq 0+k_l^2\alpha_l \int_\tau^t -k_l T_x^-[s] f''(X(s))\,ds$$

$$\leqq M_{32}\int_\tau^t -k_l T_x^-[s]\,ds \leqq M_{32}\int_\tau^t q(s)\,ds \leqq M_{33}\cdot t,$$

where we have used Lemmas 2.1 and 3.1;

$$\int_\tau^t f'(X(s)) f''(X(s))\,ds \leqq M_{34}\cdot t,$$

$$\int_\tau^t [q(s)+k_l f'(0)][T_t(0,s)-\alpha_l f''(0)]\,ds$$

$$(3.10)\qquad = [q(s)+k_l f'(0)][T(0,s)-T_{cr}]\Big|_\tau^t$$

$$-\int_\tau^t q'(s)[T(0,s)-T_{cr}]\,ds - \int_\tau^t [q(s)+k_l f'(0)]\alpha_l f''(0)\,ds$$

$$\leqq M_{35}|T(0,t)-T_{cr}|+|q'_{\max}|\int_\tau^t |T(0,s)-T_{cr}|\,ds + M_{36}\cdot t,$$

where now the second term can be made $<M_{35}\cdot t$ by choosing τ small enough. The other terms are estimated as follows:

$$|T(0,t)-T_{cr}|\leqq \int_0^{X(t)} |T_x(x,t)-f'(x)|\,dx + |f(0)-f(X(t))|$$

$$\leqq \int_0^{X(t)} \left\{\frac{1}{2\varepsilon}+\frac{\varepsilon}{2}|T_x-f'|^2\right\} dx + |f'_{\max}|\cdot X(t)\quad (\varepsilon>0)$$

$$\leqq M_{37}\cdot t+\frac{\varepsilon}{2}\int_0^L |T_x-f'|^2\,dx,$$

$$\int_\tau^t |T(0,s)-T_{cr}|\,ds \leqq M_{38}t^2+\frac{1}{2}\int_\tau^t\int_0^L |T_x-f'|^2\,dx\,ds \quad \text{(as in (3.5))}$$

$$\leqq M_{38}t^2+\frac{M_1}{2}t^2,\quad \text{by (3.2).}$$

Hence the right-hand side of (3.10) is bounded by

$$M_{39}\cdot t+M_{40}\cdot\varepsilon\int_0^L |T_x-f'|^2\,dx + M_{35}\cdot t + M_{41}t^2 + M_{36}\cdot t$$

$$=M_{42}\cdot t + M_{40}\cdot\varepsilon\int_0^L |T_x-f'|^2\,ds.$$

Finally, for the last term in (3.9) we have

$$\int_\tau^t \int_0^L \alpha k^2 f'''(x)[T_x(x,s)-f'(x)]\,dx\,ds \leqq \int_\tau^t \int_0^L \alpha k^2\left\{\frac{1}{2}|f'''|^2+\frac{1}{2}|T_x-f'|^2\right\}dx\,ds$$
$$\leqq M_{43}\cdot t+M_{44}\cdot t^2.$$

Incorporating all these bounds into (3.9) we obtain

$$\int_\tau^t \int_0^L \alpha k^2|T_{xx}-f''|^2\,dx\,ds \leqq M_{45}\cdot t+\int_0^L\left(\varepsilon M_{40}-\frac{k^2}{2}\right)|T_x-f'|^2\,dx$$
$$+\int_0^L \frac{k^2}{2}|T_x(x,\tau)-f'(x)|^2\,dx.$$

By Corollary 3.3 there exists $\tau<t$ such that the last term is $\leqq M_2 t$. Thus, choosing $\varepsilon M_{40}>k^2_{\max}/2$, we finally obtain (3.7).

COROLLARY 3.5. *There exists a constant $M_4>0$, independent of x and t, such that*

$$(3.11)\qquad \int_0^t |T_x(x,s)|^2\,ds \leqq M_4\cdot t \quad \textit{for } 0\leqq x\leqq L,\quad 0<t<t_\infty.$$

Proof. Fix (x_0,t), $0<x_0\leqq L$, $0<t<t_\infty$. We have

$$|T_x(x_0,t)-f'(x_0)|^2 \leqq 2\left|\int_0^{x_0}[T_{xx}(x,t)-f''(x)]\,dx\right|^2+2|T_x(0,t)-f'(0)|^2$$
$$\leqq 2x_0\int_0^{x_0}|T_{xx}-f''|^2\,dx+2\left|-\frac{q(t)}{k_l}-f'(0)\right|^2$$
$$\leqq 2L\int_0^L |T_{xx}-f''|^2\,dx+M_{46}.$$

Integrating and using (3.7) we obtain

$$|T_x(x_0,\cdot)|_{L^2(0,t)}\leqq\left(\int_0^t |f'(x_0)|^2\,ds\right)^{1/2}+(M_{47}t)^{1/2}\leqq(M_{48}t)^{1/2};$$

hence

$$\int_0^t |T_x(x_0,s)|^2\,ds\leqq M_{48}\cdot t,\qquad 0<x_0\leqq L,\quad 0<t<t_\infty.$$

For $x_0=0$, $|T_x(0,t)|=|-q(t)/k_l|\leqq M_{49}$ and (3.11) still holds.

COROLLARY 3.6. *There exists a constant $M_5>0$ and a sequence of times $\{t_n\}\to 0$ such that*

$$(3.12)\qquad \int_0^L |T_{xx}(x,t_n)|^2\,dx\leqq M_5.$$

Proof. By the Mean Value Theorem applied to (3.7), for any $\tau_n>0$ there exists $0<t_n<\tau_n$ such that

$$\tau_n\int_0^L |T_{xx}(x,t_n)-f''(x)|^2\,dx\leqq M_3\tau_n,$$

i.e.,

$$|T_{xx}(\cdot, t_n) - f''(\cdot)|^2_{L^2(0,L)} \leqq M_3;$$

hence

$$|T_{xx}(\cdot, t_n)|_{L^2(0,L)} \leqq \sqrt{M_3} + |f''|_{L^2(0,L)} =: \sqrt{M_5}.$$

4. Maximum principle and boundedness of the fluxes. Boundedness of the fluxes, $-k_l T_x$ and $-k_s T_x$, can be deduced from Corollary 3.5 once we make certain that unboundedness could only occur at the origin. This is shown in Theorem 4.2, the proof of which requires the following form of the Corner Point Maximum Principle, proved in Solomon, Alexiades and Wilson [13, p. 205].

LEMMA 4.1. *Let D be a simply connected domain in the x, t plane and $P_0 = (x_0, t_0)$ a point of its boundary. Let N be a disk of radius $\delta > 0$ centered at P_0, and set $G^0 = D \cap N \cap \{t < t_0\}$, $\hat{G}^0 = \bar{G}^0 - \partial D$. If* (a) *$u \in C(\bar{D})$, u_x, u_t, $u_{xx} \in C(D)$, and $u_t - \alpha u_{xx} \leqq 0$ in D, and* (b) *$u(P) < u(P_0)$ for $P \in \hat{G}^0$, $u(P) \leqq u(P_0)$ for $P \in \partial D \cap N$ and $\partial D \cap N$ is a C^2-curve representable as $x = X(t)$, then*

$$\limsup_{\substack{P \to P_0 \\ P \in \hat{G}^0}} \frac{u(P) - u(P_0)}{|P - P_0|} < 0,$$

where P tends to P_0 in any nontangential direction.

THEOREM 4.2. *The maximum and minimum values of the flux over any strip $\sum_{t_2}^{t_1} = \{(x, t) \colon 0 \leqq x \leqq L,\ t_1 \leqq t \leqq t_2\}$, $0 < t_1 < t_2 < t_\infty$, are attained either on the line $x = 0$, $t_1 \leqq t \leqq t_2$ or on the bottom $0 \leqq x \leqq L$, $t = t_1$.*

Proof. Let $u(x, t) := -kT_x(x, t)$ where, as before $k = k_l$ in liquid and $k = k_s$ in the solid. Since u satisfies the heat equation in $D_l \cap \sum_{t_2}^{t_1}$ and in $D_s \cap \sum_{t_2}^{t_1}$, its extrema over $\sum_{t_2}^{t_1}$ must be attained on the parabolic boundaries of these domains. By Lemma 4.1, they cannot be assumed on $x = L$, $t_1 \leqq t \leqq t_2$ due to the boundary condition $u(L, t) = -k_s T_x(L, t) \equiv 0$ there. Thus the result will be proved once we show that the extrema cannot be attained on the interface $x = X(t)$, $t_1 < t \leqq t_2$.

Let $u^{\pm}[t] := -kT_x^{\pm}[t]$ be the values of the flux on $x = X(t)$ from the solid (+) and the liquid (−) respectively. Suppose $u(x, t)$ attains its maximum over $\sum_{t_2}^{t_1}$ at a point $(X(t^*), t^*)$, $t_1 < t^* \leqq t_2$. Note that, by Lemma 2.1, $u^{\pm}[t^*] \geqq 0$.

If $X'(t^*) \geqq 0$ then $\rho H X'(t^*) = u^-[t^*] - u^+[t^*] \geqq 0$ implies the maximum of u must be $u^-[t^*]$. But $T(X(t), t) = T_{cr} \Rightarrow T_t^- + T_x^- X'(t) = 0 \Rightarrow k_l T_{xx}^-[t] = \rho c_l T_t^-[t] = -\rho c_l T_x^-[t] X'(t) \Rightarrow u_x^-[t^*] = \rho c_l T_x^-[t^*] X'(t^*) = -(\rho c_l / k_l) u^-[t^*] X'(t^*) \leqq 0$. This contradicts Lemma 4.1 and we conclude that the maximum cannot occur at time t^* but at an earlier time, down to $t = t_1$.

If $X'(t^*) < 0$ then, similarly to the previous case we obtain $u_x^+[t^*] = -(\rho c_s / k_s) u^+[t^*] X'(t^*) \geqq 0$, whereas Lemma 4.1 requires <0. Again, the maximum must be at an earlier time, down to $t = t_1$.

At a minimum, all the inequalities reverse and we draw the same conclusion.

THEOREM 4.3. *The flux is bounded uniformly in $\overline{D_l \cup D_s}$.*

Proof. Suppose $u(x, t) = -kT_x(x, t)$ is unbounded in $\overline{D_l \cup D_s}$. By Theorem 4.2, it can only become unbounded at (0, 0) because on $x = 0$, $u = q$ is bounded and at $t = 0$, $0 < x \leqq L$, $u = -k_s f'(x)$ is also bounded. Therefore, near (0, 0), $T_x(x, t)$ is either uniformly large or experiences unbounded oscillations. But the former possibility contradicts (3.11) while the latter one contradicts (3.12). We conclude that there exists a constant $M_6 > 0$ such that

$$|T_x(x, t)| \leqq M_6 \quad \text{in } \overline{D_l \cup D_s}, \tag{4.1}$$

in the sense that

$$-M_6 \leqq \liminf T_x \leqq \limsup T_x \leqq M_6 \quad \text{as } (x, t) \to (0, 0)$$

in case $T_x(0, 0)$ does not exist.

COROLLARY 4.4. *The interface speed is uniformly bounded for* $0 \leqq t \leqq t_\infty$, *i.e., there exists a constant* $M_7 > 0$ *such that*

$$|X'(t)| \leqq M_7, \qquad 0 \leqq t \leqq t_\infty, \tag{4.2}$$

in the sense

$$-M_7 \leqq \liminf_{t \to 0} X'(t) \leqq \limsup_{t \to 0} X'(t) \leqq M_7.$$

(*Note that we have not yet shown the existence of* $X'(0)$.)

5. Continuity of the fluxes. The uniform boundedness of the fluxes established in the previous section allows us to estimate the third space derivation of T and improve (3.12) at the same time, leading to our goal.

THEOREM 5.1. *There exists a constant* $M_8 > 0$, *independent of time, such that*

$$\int_0^L |T_{xx}(x, t)|^2 \, dx \leqq M_8 \quad \textit{for any } 0 < t < t_\infty. \tag{5.1}$$

Proof. We begin with

$$\begin{aligned} \frac{d}{dt} \int_0^L \frac{1}{2} T_{xx}(x, t)^2 \, dx &= \frac{1}{2} X'(t) \{ T_{xx}^-[t]^2 - T_{xx}^+[t]^2 \} + \int_0^L T_{xx} T_{xxt} \, dx \\ &= \frac{1}{2} X'(t) \{ T_{xx}^-[t]^2 - T_{xx}^+[t]^2 \} + T_{xx} T_{xt}|_{x=L} - T_{xx} T_{xt}|_{x=0} \\ &\quad + \{ T_{xx}^-[t] T_{xt}^-[t] - T_{xx}^+[t] T_{xt}^+[t] \} - \int_0^L \alpha T_{xxx}(x, t)^2 \, dx. \end{aligned} \tag{5.2}$$

On $x = L$, $T_x(L, t) = 0$ implies $T_{xt} \equiv 0$. On $x = 0$, by parabolic regularity theory, the heat equation is valid up to the boundary so $-T_{xx} T_{xt}|_{x=0} = (1/\alpha_l) T_t(0, t)(q'(t)/k_l)$, $t > 0$. Similarly along the front, but separately from each side

$$\begin{aligned} T_{xx}^\pm[t] T_{xt}^\pm[t] &= \frac{1}{\alpha} T_t^\pm[t] T_{xt}^\pm[t] = -\frac{1}{\alpha} X'(t) T_x^\pm[t] T_{xt}^\pm[t] \\ &= -\frac{X'(t)}{\alpha} \left\{ \frac{d}{dt} \left[\frac{1}{2} (T_x^\pm)^2 \right] - X'(t) T_{xx}^\pm[t] \right\} \\ &= -\frac{X'(t)}{2a} \frac{d}{dt} (T_x^\pm[t]^2) + \frac{X'(t)^2}{\alpha} T_{xx}^\pm[t] \\ &= -\frac{1}{2\alpha} \frac{d}{dt} [X'(T_x^\pm)^2] + \frac{1}{2\alpha} X''(t) T_x^\pm[t]^2 + \frac{X'(t)^2}{\alpha} T_{xx}^\pm[t], \qquad t > 0. \end{aligned}$$

With these modifications, we integrate (5.2) over (t_n, t), where $t_n \to 0$ is the sequence in (3.12) and $t > 0$ is arbitrary, to obtain

$$\begin{aligned}
&\int_{t_n}^{t}\int_0^L \alpha T_{xxx}(x, s)^2\,dx\,ds + \frac{1}{2}\int_0^L T_{xx}(x, t)^2\,dx \\
&\quad = \frac{1}{2}\int_0^L T_{xx}(x, t_n)^2\,dx + \frac{1}{2}\int_{t_n}^{t} X'(s)\{T_{xx}^-[s]^2 - T_{xx}^+[s]^2\}\,ds + \frac{1}{\alpha_l k_l}\int_{t_n}^{t} T_t(0, s)q'(s)\,ds \\
&\qquad + \int_{t_n}^{t}\left\{-\frac{1}{2\alpha_l}\frac{d}{ds}(X'(s)T_x^-[s]^2) + \frac{1}{2\alpha_l}X''(s)T_x^-[s]^2 + \frac{1}{\alpha_l}X'(s)^2T_{xx}^-[s]\right\}ds \\
&\qquad - \int_{t_n}^{t}\left\{-\frac{1}{2\alpha_s}\frac{d}{ds}(X'(s)T_x^+[s]^2) + \frac{1}{2\alpha_s}X''(s)T_x^+[s]^2 + \frac{1}{\alpha_s}X'(s)^2T_{xx}^+[s]\right\}ds.
\end{aligned} \tag{5.3}$$

Each term on the right can be bounded as follows.

By the choice of t_n as in (3.12), the first integral is bounded by M_5. Along the front, thanks to (4.1) and (4.2),

$$T_{xx}^{\pm}[t] = \frac{1}{\alpha}T_t^{\pm}[t] = \frac{1}{\alpha}(-X'(t)T_x^{\pm}[t]) \leqq \frac{1}{\alpha}M_7M_6; \tag{5.4}$$

hence the second integral is bounded by (constant) · t. If we apply the second Mean Value Theorem, there exists $\tilde{t}_n \in (t_n, t)$ such that

$$\left|\int_{t_n}^{t} T_t(0, s)q'(s)\,ds\right| = |q'(\tilde{t}_n)[T(0, t) - T(0, t_n)] \leqq M_{51},$$

and there exists $\tilde{t}_n \in (t_n, t)$ such that

$$\int_{t_n}^{t} X''(s)T_x^-[s]^2\,ds = T_x^-[\tilde{t}_n]^2(X'(t) - X'(t_n)) \leqq 2M_6^2M_7.$$

Also,

$$\left|\int_{t_n}^{t} \frac{d}{ds}(X'(s)T_x^-[s]^2)\,ds\right| \leqq 2M_6M_7,$$

$$\left|\int_{t_n}^{t} X'(s)^2T_{xx}^-[s]\,ds\right| \leqq \frac{1}{\alpha_l}M_7^3M_6 \quad \text{as in (5.4).}$$

It is similar for the last three terms (from the solid side).

With its right-hand side thus bounded, (5.3) yields

$$\int_{t_n}^{t}\int_0^L \alpha T_{xxx}(x, s)^2\,dx\,ds + \frac{1}{2}\int_0^L T_{xx}(x, t)^2\,dx \leqq M_{52}, \qquad 0 < t_n < t \leqq t_\infty.$$

Taking $t_n \to 0$, we obtain

$$\int_0^t\int_0^L |T_{xxx}(x, s)|^2\,dx\,ds \leqq M_{53} \quad \text{for } 0 < t < t_\infty, \tag{5.5}$$

as well as estimate (5.1).

COROLLARY 5.2. *For any points (x_1, t), (x_2, t) in $\bar{D}_s$, $t>0$,*

$$|T_x(x_2, t)-T_x(x_1, t)| \leqq M_8^{1/2}|x_2-x_1|^{1/2}, \tag{5.6}$$

uniformly in $0<t<t_\infty$. Hence the family $\{T_x(\cdot, t)\colon 0<t<t_\infty\}$ is equicontinuous.

Proof. $|T_x(x_2, t)-T_x(x_1, t)|^2=|\int_{x_1}^{x_2} T_{xx}(x, t)\,dx|^2 \leqq |x_2-x_1| \int_0^L |T_{xx}|^2\,dx \leqq M_8|x_2-x_1|$, independently of $t \in (0, t_\infty)$.

THEOREM 5.3. *The solid flux $-k_sT_x(x, t)$ is continuous on $\bar{D}_s$. Hence,*

$$T_x(x, t) \to f'(0) \quad \textit{as } (x, t) \to (0, 0) \quad \textit{inside } \bar{D}_s,$$

and in particular,

$$T_x^+[t] \to f'(0) \quad \textit{as } t \downarrow 0.$$

Proof. The family of functions $\{T_x(\cdot, t)\colon 0<t<t_\infty\}$ is equicontinuous on $[0, L]$ by (5.6) and equibounded by (4.1). The Ascoli-Arzela lemma implies that there exists a sequence of times $\{t_n^*\} \to 0$ such that $T_x(x, t_n^*)$ converges to its limit $f'(x)$ uniformly in $x \in [0, L]$.

Suppose now that $T_x(x, t)$ is not continuous at $(0, 0)$, i.e., that there exists $M>0$ and a sequence of points $(x_n, t_n) \to (0, 0)$ such that

$$|T_x(x_n, t_n)-f'(0)| \geqq M.$$

Then, by Ascoli-Arzela, there exists a subsequence $\{t_n^*\} \to 0$ such that

$$M<|T_x(x_n, t_n^*)-f'(0)| \leqq |T_x(x_n, t_n^*)-T_x(x_n, 0)|+|f'(x_n)-f'(0)|,$$

the right-hand side of which can be made arbitrarily small by choosing x_n sufficiently close to $x=0$, a contradiction.

THEOREM 5.4. *The liquid flux $-k_lT_x(x, t)$ is continuous on $\bar{D}_l$. Hence,*

$$-k_lT_x(x, t) \to q(0) \quad \textit{as } (x, t) \to (0, 0) \textit{ inside } \bar{D}_l,$$

and in particular,

$$-k_lT_x^-[t] \to q(0) \quad \textit{as } t \downarrow 0.$$

Proof. For $(x, t) \in \bar{D}_l$, $t>0$, we have

$$\begin{aligned}
|-k_lT_x(x, t)-q(0)| &\leqq k_l|T_x(x, t)-T_x(0, t)|+|q(t)-q(0)| \\
&= k_l\left|\int_0^x T_{xx}(\xi, t)\,d\xi\right|+|q(t)-q(0)| \\
&\leqq k_lx^{1/2}\left(\int_0^L |T_{xx}(\xi, t)|^2\,d\xi\right)^{1/2}+|q(t)-q(0)| \\
&\leqq k_lM_8^{1/2}\sqrt{x}+|q(t)-q(0)|,
\end{aligned}$$

which $\to 0$ as $(x, t) \to (0, 0)$.

COROLLARY 5.5. *The speed $X'(t)$ of the phase change front is continuous for $0 \leqq t \leqq t_\infty$, and*

$$\begin{aligned}
\rho HX'(t) &= -k_lT_x^-[t]+k_sT_x^+[t] \\
&\to q(0)+k_sf'(0) \quad \textit{as } t \downarrow 0.
\end{aligned}$$

REFERENCES

[1] J. CANNON AND M. PRIMICERIO, *A Stefan problem involving the appearance of a phase*, this Journal, 4 (1973), pp. 141-148.

[2] H. CARSLAW AND J. JAEGER, *Conduction of Heat in Solids*, 2nd edition, Oxford Univ. Press, London-Oxford, 1959.

[3] A. FASANO AND M. PRIMICERIO, *General free-boundary problems for the heat equation*, II, J. Math. Anal. Appl., 58 (1977), pp. 202-231.

[4] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[5] H. ISHII, *On a certain estimate of free boundary in the Stefan problem*, J. Differential Equations, 42 (1981), pp. 106-115.

[6] A. LACEY AND M. SHILLOR, *The existence and stability of regions with superheating in the classical two-phase one-dimensional Stefan problem with heat sources*, preprint.

[7] H. G. LANDAU, *Heat conduction in a melting solid*, Quart. Appl. Math., 8 (1950), pp. 81-94.

[8] M. NIEZGODKA, *Stefan-like problems*, in Free Boundary Problems—Theory and Applications, Vol. 2, A. Fasano and M. Primicerio, eds., Pitman Advanced Publishing Program, 1983, pp. 321-348.

[9] L. RUBINSTEIN, *The Stefan Problem*, AMS Translation, American Mathematical Society, Providence, RI, 1971.

[10] A. D. SOLOMON, *Some aspects of the computer simulation of conduction heat transfer and phase change processes*, in Computing Methods in Applied Sciences and Engineering V, R. Glowinski and J. Lions, eds., North-Holland, Groningen, The Netherlands, 1982, pp. 457-470.

[11] ———, *The applicability and extendability of Megerlin's method for solving parabolic free boundary problems*, in Moving Boundary Problems, D. G. Wilson, Nan D. Solomon and Paul T. Boggs, eds., Academic Press, New York, 1978, pp. 187-202.

[12] ———, *Numerical methods for solving Stefan-type problems*, Res. Mechanica, to appear.

[13] A. D. SOLOMON, V. ALEXIADES AND D. G. WILSON, *The Stefan problem with a convective boundary condition*, Quart. Appl. Math., 40 (1982), pp. 203-217.

[14] ———, *Explicit solutions to phase change problems*, Quart. Appl. Math., 41 (1983), pp. 237-243.

[15] D. G. WILSON, A. D. SOLOMON AND J. S. TRENT, *A bibliography on moving boundary problems with key word index*, Union Carbide Corp., Report No. ORNL/CSD-44, 1979.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

018

ON THE SINGULAR LIMIT FOR A CLASS OF PROBLEMS MODELLING PHASE TRANSITIONS*

NICHOLAS D. ALIKAKOS† AND K. C. SHAING‡

Abstract. The total variation is a measure of the complexity of a given solution to $(\varepsilon)/\varepsilon^2 u_\varepsilon'' - f(x, u_\varepsilon) = 0$. For stable solutions, under appropriate conditions on f, the total variation can be bounded uniformly in ε. This estimate provides sufficient compactness for relating the stable solutions of (ε) to those of the limiting equation $(\varepsilon = 0)$.

Key words. singular perturbations, phase transitions, complexity of stable equilibria, total variation

AMS(MOS) subject classifications. 34C35, 34D15, 35B45

1. Introduction. Consider

$$\varepsilon^2 u_\varepsilon'' - f(x, u_\varepsilon) = 0, \quad u_\varepsilon'(0) = u_\varepsilon'(1) = 0, \quad 0 \leqq x \leqq 1. \tag{1}$$

Let

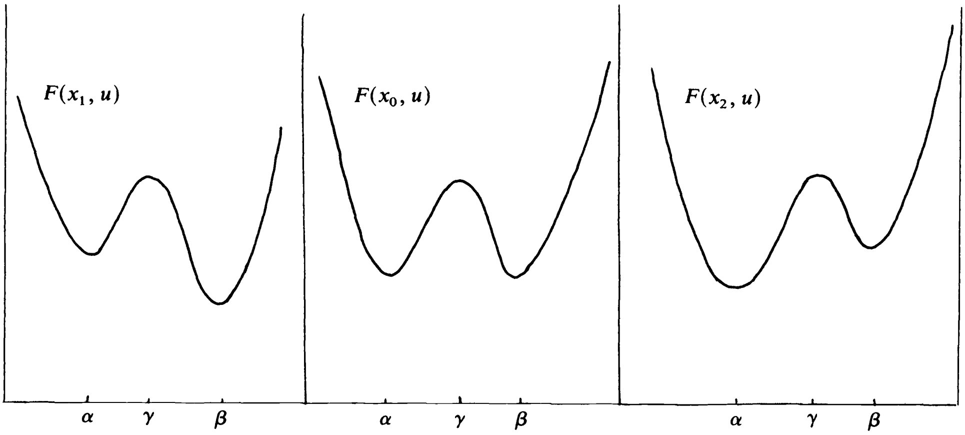
$$F(x, u) = \int^u f(x, z)\, dz, \qquad \text{an antiderivative of } f,$$

and assume that for $x_1 < x_0 < x_2$ we have (see Fig. A): The energy functional associated to (1) is

$$J_\varepsilon(u) = \frac{\varepsilon^2}{2} \int (u')^2 + \int F(x, u). \tag{2}$$

For $\varepsilon = 0$ the global minimizer is given by

$$\bar{u}(x) = \begin{cases} \beta, & 0 \leqq x < x_0, \\ \alpha, & x_0 < x \leqq 1. \end{cases} \tag{3}$$



A-I A-II A-III

FIG. A

* Received by the editors May 29, 1985; accepted for publication (in revised form) November 15, 1985. This research was partially supported by a grant from the National Science Foundation.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300.

‡ Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831.

The question we investigate concerns the convergence of any sequence $\{u_\varepsilon\}$ of global minimizers of (2) to $\bar{u}$ as $\varepsilon \to 0$. Since $\bar{u}$ is discontinuous, one expects a sharp transition for u_ε near $x = x_0$ and so a penalty in the energy in terms of a substantial contribution from $\int (u'_\varepsilon)^2$, a fact that makes the problem from the mathematical point of view not entirely clear (note that α, β are local minimizers of J_ε for all ε). Setting $\varepsilon = 0$ and so obtaining a profile by locally minimizing the energy is the natural approach in applications. As a rule the equation is coupled to a system and the limiting case has the advantage of simplicity and therefore it is of interest to have rigorous justification of the approximation. A typical physical setting where the question of the relation between local and nonlocal models is raised can be found in Shaing [S].

Fife [F] constructs solutions to (1) that converge as $\varepsilon \to 0$ to a given solution of the limiting equation by developing singular perturbation techniques based ultimately on a suitable implicit function theorem. No evaluation of the energy of the approximating family is considered in that work.

Recently Carr, Gurtin and Slemrod [CGS1][1] studied phase transitions on a finite interval. They minimize the functional

$$J_\varepsilon(u) = \frac{\varepsilon^2}{2}\int (u'(x))^2 + \int W(u(x)) \tag{4}$$

with the constraint

$$\int u = M \tag{5}$$

where $W(u)$ has a sigmoid shape. Utilizing that the Euler–Lagrange equation to (4), (5) is autonomous, they employ elaborate phase plane analysis arguments and establish under appropriate conditions that the minimizers converge as $\varepsilon \to 0$ to a two-phase solution of the $\varepsilon = 0$ problem.

Our approach is different from any of the above. It is based on the observation that if $\{u_\varepsilon\}$ is a family of *stable* equilibria of (1), bounded uniformly in ε,

$$|u_\varepsilon|_{L^\infty} \leqq C,$$

then the total variations, $V(u_\varepsilon)$, are uniformly bounded,

$$V(u_\varepsilon) \leqq C, \qquad C \text{ independent of } \varepsilon. \tag{6}$$

The obvious counterexample, $u_\varepsilon(x) = \cos(x/\varepsilon)$, for $\varepsilon^2 u''_\varepsilon + u_\varepsilon = 0$, $u'_\varepsilon(0) = u'_\varepsilon(1) = 0$, shows that the hypothesis of stability is essential. By employing Helly's compactness theorem we can pass to the limit in the weak formulation of the ε-problem along a subsequence of minimizers and conclude its convergence to a solution of the $\varepsilon = 0$ problem. The rest of the argument turns out not to be difficult. This method, being a compactness argument, has the advantage of not distinguishing between autonomous and nonautonomous equations and it is potentially applicable to systems relevant, for example, to the study of phase transitions of fluids which differ in more than one scalar parameter (Cahn and Hilliard [CH, p. 260]), and to equations in more than one space dimension. Estimates of the variation under general circumstances will be pursued elsewhere. Here we describe the main idea in the simplest of circumstances. In § 2 under appropriate hypotheses on f the pointwise convergence of the global minimizers to $\bar{u}$ is established. Finally in an appendix we give a quick self-contained proof of the main result in [CGS1].

[1] See also Novick-Cohen and Segel [NS].

2. We will assume for simplicity that α, β, γ are independent of x and that $F(\cdot, u)$ changes in a monotone way from A-I to A-III, i.e. the left branch goes down while the right goes up and that otherwise F is as in Fig. A. More precisely we introduce the hypotheses:

(H1) F is C^1 in both arguments.

(H2) $f(x, \alpha) = f(x, \beta) = f(x, \gamma) = 0$, $0 \leqq x \leqq 1$, $u \to f(x, u)$ changes sign as u crosses the roots, $f(x, u) < 0$ for $u < \alpha$, $[F(0, \alpha) - F(0, \beta)][F(1, \alpha) - F(1, \beta)] < 0$.

(H3) $f_x(x, u) > 0$ for $u \notin \{\alpha, \beta, \gamma\}$, $0 \leqq x \leqq 1$.

We take

$$F(x, u) = \int_\gamma^u f(x, z)\, dz + \text{const} \geqq 0.$$

Note that $F_x(x, u) \leqq 0$ $(\geqq 0)$ for $u \leqq \gamma$ $(u \geqq \gamma)$. A concrete example would be

$$f(x, u) = u(u^2 - 1) \int_{-\infty}^{(x-1/2)u(u^2-1)} g(t)\, dt$$

where $g(t) > 0$ with $\int_{-\infty}^{\sigma} g(t)\, dt < +\infty$, $\forall \sigma \in \mathbb{R}$, $\alpha = -1$, $\beta = 1$, $\gamma = 0$, $x_0 = \frac{1}{2}$.

LEMMA 1. *There is a global minimizer u_ε of $J_\varepsilon(u)$ over $W^{1,2}$ that is a classical solution to* (1) *and satisfying*

$$\alpha \leqq u_\varepsilon \leqq \beta \tag{7}$$

with the inequalities being strict if u_ε is not identically constant.

Proof. This is standard. We sketch the argument for the convenience of the reader.

Take $\varepsilon = 1$ for definiteness. First note that for a given u in $W^{1,2}$ the truncated $v = \min\{u, \beta\}$ has energy less or equal to u. Clearly

$$\int v_x^2 \leqq \int u_x^2.$$

Also,

$$\int F(x, u) = \int_{u \leqq \beta} F(x, u) + \int_{u \geqq \beta} F(x, u)$$
$$\geqq \int_{u \leqq \beta} F(x, u) + \int_{u \geqq \beta} F(x, \beta)$$

since $\partial F(x, u)/\partial u = f(x, u) \geqq 0$ for $u \geqq \beta$ by (H2). Therefore

$$J_1(v) \leqq J_1(u).$$

Applying an analogous argument to v, we conclude that $\max\{\alpha, \min\{u, \beta\}\}$ has energy less than or equal to u. Consider J_1 on

$$X = \{u \in W^{1,2} \mid \alpha \leqq u \leqq \beta\}. \tag{8}$$

Clearly J_1 is coercive on X (with the obvious norm) since $F(x, u) \geqq 0$ and weakly lower semicontinuous. Therefore J_1 attains its infimum over X. By the observation above this minimizer u_1 minimizes J_1 over $W^{1,2}$ as well. From

$$\left.\frac{d}{d\lambda}\right|_{\lambda=0} J_1(u_1 + \lambda h) = 0 \tag{9}$$

we obtain that u_1 is a weak solution to (1). A standard regularity argument shows that u_1 is a classical solution. By applying a familiar variant of the strong maximum principle we obtain strict inequalities when the minimizer is not identically equal to a constant. Q.E.D.

LEMMA 2. *Let u_ε be a global minimizer. Then we have the estimate*[2]

$$\int |u'_\varepsilon|^2 \leqq \frac{C}{\varepsilon^2}. \tag{10}$$

Proof. It follows from (1) by multiplying with u_ε, integrating, and using the Schwarz inequality and the uniform bound in (7). Q.E.D.

LEMMA 3. *Stable equilibria of* (1) *are necessarily monotone decreasing in x.*

Proof. We will show that if u_ε is not monotone then the second variation of J_ε

$$J''_\varepsilon(u_\varepsilon)(h, h) = \varepsilon^2 \int (h')^2 + \int f_u(x, u_\varepsilon)h^2$$

is strictly negative for some $h \in W^{1,2}$. We drop the ε for convenience. Differentiating (1) we obtain for $u' = w$

$$\varepsilon^2 w'' - f_u(x, u)w = f_x(x, u), \qquad w(0) = w(1) = 0. \tag{11}$$

Since w is in $W^{1,2}$ (cf. Lemma 1), so is $w^+ = \max\{w, 0\}$. Multiplying (11) by w^+ and integrating, we obtain

$$-\varepsilon^2 \int |(w^+)|^2 - \int f_u(x, u)(w^+)^2 = \int f_x(x, u)w^+. \tag{12}$$

By (H3) the right-hand side is positive. Hence stability implies $(u_x)^+ = 0$. Q.E.D.

LEMMA 4. *Let $\{u_{\varepsilon_n}\}$ be a sequence of global minimizers of J_{ε_n} over $W^{1,2}$. Then there is a subsequence, denoted again by $\{u_{\varepsilon_n}\}$ such that*

$$u_{\varepsilon_n} \underset{\varepsilon_n \to 0}{\longrightarrow} u \quad \textit{pointwise on } [0, 1] \tag{13}$$

where u is monotone decreasing

$$\alpha \leqq u \leqq \beta \tag{14}$$

and

$$f(x, u(x)) = 0 \quad \textit{on } [0, 1] \tag{15}$$

and thus $u(x) \in \{\alpha, \gamma, \beta\}$.

Proof. By Lemma 3 the elements of the sequence $\{u_{\varepsilon_n}\}$ are monotone decreasing functions of x. We write u_ε in the place of u_{ε_n} for convenience. Since

$$\alpha \leqq u_\varepsilon \leqq \beta$$

$V(u_\varepsilon) \leqq \beta - \alpha$ and so by Helly's theorem [N, p. 220] there is a subsequence, denoted again by $\{u_\varepsilon\}$ such that

$$u_\varepsilon(x) \to u(x), \qquad x \in [0, 1] \tag{16}$$

where $u(x)$ necessarily is monotone decreasing and satisfies (14). Multiplying (1) by

[2] The stronger estimate $\int |u'_\varepsilon|^2 \leqq C/\varepsilon$ holds. See [AS].

ϕ, a smooth function and integrating by parts,

$$\varepsilon^2 \int u'_\varepsilon \phi' + \int f(x, u_\varepsilon)\phi = 0. \tag{17}$$

The first term in (17) goes to zero as $\varepsilon \to 0$ by the estimate (10). By (14) and the dominated convergence theorem we obtain from (16), (17)

$$\int f(x, u)\phi = 0 \tag{18}$$

and therefore (15) follows. Q.E.D.

Remark 1. Since $u(x)$ is monotone decreasing, the only left continuous possibilities are:

(i) $$u(x) \equiv \text{constant},$$

(ii) $$u(x) = \begin{cases} \beta, & 0 \leqq x \leqq \bar{x}, \\ \alpha, & \bar{x} < x \leqq 1, \end{cases}$$

(iii) $$u(x) = \begin{cases} \beta, & 0 \leqq x \leqq \bar{x}, \\ \gamma, & \bar{x} < x < 1, \end{cases}$$

(iv) $$u(x) = \begin{cases} \gamma, & 0 \leqq x \leqq \bar{x}, \\ \alpha, & \bar{x} < x \leqq 1, \end{cases}$$

(v) $$u(x) = \begin{cases} \beta, & 0 \leqq x \leqq \bar{x}, \\ \gamma, & \bar{x} \leqq x \leqq \bar{x}, \\ \alpha, & \bar{x} < x \leqq 1. \end{cases}$$

LEMMA 5. *Let $\{u_{\varepsilon_n}\}$ be a sequence of global minimizers. Then*

$$J_{\varepsilon_n}(u_{\varepsilon_n}) \to J_0(\bar{u}) = \int F(x, \bar{u}(x))$$

as $\varepsilon_n \to 0$.

Proof. Replace ε_n by ε for simplicity. Consider the functions

$$\tilde{u}_\varepsilon(x) = \begin{cases} \beta, & 0 \leqq x \leqq x_0 - \varepsilon, \\ \alpha, & x_0 + \varepsilon \leqq x \leqq 1, \\ \text{linear}, & x_0 - \varepsilon \leqq x \leqq x_0 + \varepsilon, \end{cases}$$

$\tilde{u}_\varepsilon \in W^{1,2}$ and

$$J_\varepsilon(\tilde{u}_\varepsilon) = \frac{\varepsilon^2}{2} \int_{x_0-\varepsilon}^{x_0+\varepsilon} \left(\frac{\beta - \alpha}{2\varepsilon}\right)^2 + \int_0^{x_0-\varepsilon} F(x, \beta) + \int_{x_0+\varepsilon}^1 F(x, \alpha) + \int_{x_0-\varepsilon}^{x_0+\varepsilon} F(x, \tilde{u}_\varepsilon(x)).$$

Clearly $J_\varepsilon(\tilde{u}_\varepsilon) \to J_0(\bar{u})$, and since $J_0(\bar{u}) \leqq J_\varepsilon(u_\varepsilon) \leqq J_\varepsilon(\tilde{u}_\varepsilon)$, we are done. Q.E.D.

Remark 2. The weaker statement $\overline{\lim}_{\varepsilon \to 0} J_\varepsilon(u_\varepsilon) \leqq J_0(\bar{u})$ would suffice.

THEOREM 6. *Let $\{u_{\varepsilon_n}\}$ be a sequence of global minimizers corresponding to J_{ε_n} over $W^{1,2}$. Then*

$$u_{\varepsilon_n} \to \bar{u} \quad \textit{pointwise on } [0, 1]$$

as $\varepsilon_n \to 0$.

Proof. Assume that the limit u of some subsequence, denoted again by $\{u_{\varepsilon_n}\}$, is different from $\bar{u}$. By Remark 1 it has to be one of the remaining four possibilities. Since the global minimizer for J_0 is unique, we can choose δ such that

$$\int F(x, \bar{u}) = J_0(\bar{u}) < J_0(u) - \delta = \int F(x, u) - \delta. \tag{19}$$

On the other hand by Lemma 5 above for ε sufficiently small

$$J_\varepsilon(u_\varepsilon) = \frac{\varepsilon^2}{2}\int (u'_\varepsilon)^2 + \int F(x, u_\varepsilon) \leqq J_0(\bar{u}) + \frac{\delta}{2} = \int F(x, \bar{u}) + \frac{\delta}{2}.$$

Therefore,

$$\frac{\varepsilon^2}{2}\int (u'_\varepsilon)^2 \leqq \int F(x, \bar{u}) - \int F(x, u_\varepsilon) + \frac{\delta}{2}. \tag{20}$$

Taking now the limit and using that $u_\varepsilon \to u$ pointwise, we obtain via (19) that the right-hand side of (20) becomes negative for ε sufficiently small, a contradiction. Q.E.D.

Remark 3. The degree of complexity of stable equilibria for the case in which F does not depend explicitly on x has been investigated in one space dimension by Chafee [C] and in higher space dimensions by Casten and Holland [CH2] and independently by Matano [M]. That the extent of complexity of stable equilibria in the general case ($\mathscr{F} = F(x, u)$) can be estimated uniformly with respect to the diffusion coefficient is a point that apparently has not been exploited in settings similar to ours.

Remark 4. How small ε has to be taken so that u_ε is a qualitatively good approximation to $\bar{u}$ is determined by the difference between the energy levels. It can be shown via the implicit function theorem that if ε is outside this range then u_ε in general will not exhibit a sharp transition and so it will be a poor (qualitatively) approximation. For example, $u_\varepsilon \equiv$ constant for ε large enough is a possibility for appropriate $\mathscr{F}$. This question seems to be intimately related to uniqueness for the minimizer of J_ε [AS].

Appendix. We refer the reader to [CGS1] and to the references therein for the physical background of the problem.

Consider the functional

$$J_\varepsilon(u) = \int_{-1}^{1} [W(u(x)) + \varepsilon^2(u'(x))^2] \tag{21}$$

with the constraint

$$\int_{-1}^{1} u = M. \tag{22}$$

(P_ε): Minimize J_ε under the constraint (22) for $u > 0$, $u \in W^{1,2}$.

(P_0): Minimize J_0 under the constraint (22) for $u > 0$ such that $W(u) \in L^1$. We take

(i) $W \in C^2(0, \infty)$,

(ii) $W'' > 0$ on $(0, \bar{\alpha}) \cup (\bar{\beta}, \infty)$, $W'' < 0$ on $(\bar{\alpha}, \bar{\beta})$,

(iii) $W'(0) < W'(\bar{\beta})$, $W'(\infty) > W'(\bar{\alpha})$,

(iv) $\alpha_0 < r < \beta_0$, $r = M/2$.

(See Fig. B.)

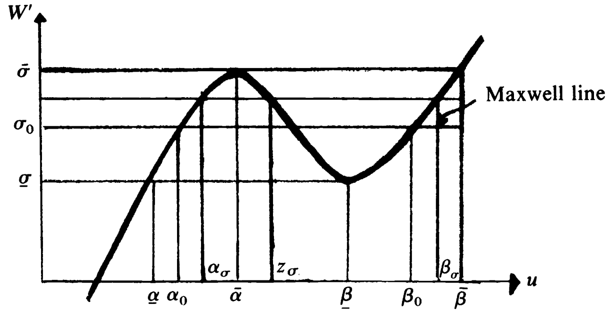


FIG. B

The problem (P_0) is solved with the auxiliary functional

$$\int_{-1}^{1} [W(u(x)) - \sigma u(x)],$$

σ = constant, a Lagrange multiplier. At a minimum the Weierstrass-Erdmann corner conditions have to be satisfied:

$$W'(u) = \sigma \text{ at points of continuity of } u,$$

$$W(u) = \sigma u \text{ continuous across jumps of } u.$$

These last conditions force the solution to be either constant or piecewise constant:

$$\bar{u}(x) = \begin{cases} \alpha_0, & x \in S_1, \\ \beta_0, & x \in S_2. \end{cases} \tag{23}$$

S_1, S_2 are disjoint measurable sets whose union is $[-1, 1]$ and

$$W(\beta_0) - W(\alpha_0) = \sigma_0(\beta_0 - \alpha_0),$$

$$\sigma_0 = W'(\alpha_0) = W'(\beta_0),$$

$$l_i = |S_i|, \quad l_1 = \frac{2(\beta_0 - r)}{\beta_0 - \alpha_0}, \quad l_2 = \frac{2(r - \beta_0)}{\beta_0 - \alpha_0}.$$

The condition $\alpha_0 < r < \beta_0$ forces the two-phase solution as the only possibility. Since only the measures l_i are determined but not the sets themselves there are infinitely many global minimizers for (P_0). The following theorem shows that from these infinitely many two-phase solutions the single-interface solution

$$\bar{u}(x) = \begin{cases} a_0, & -1 \leqq x \leqq -1 + l_1, \\ \beta_0, & -1 + l_1 < x \leqq 1 \end{cases}$$

or its reversal, $\bar{u}(-x)$, are preferred. In fact we have the following.

THEOREM A1 [CGS1]. *Let $\{u_{\varepsilon_n}\}$ be a sequence of global minimizers of* (21), (22). *Then u_{ε_n} or its reversal converges as $\varepsilon_n \to 0$, pointwise, to $\bar{u}(x)$.*

Remark A1. In [CGS1] additional information about the asymptotic shape (in ε) of the global minimizers is obtained as well as uniqueness of the global minimizer for ε small enough.

Fix ε and consider the Euler-Lagrange equation corresponding to (21), (22):

$$2\varepsilon^2 u_\varepsilon'' = W'(u_\varepsilon) - \sigma_\varepsilon, \qquad u_\varepsilon'(-1) = u_\varepsilon'(1) = 0. \tag{24}$$

LEMMA A2. *There is a smooth minimizer to* (21), (22) *that satisfies classically* (24) *and moreover*

$$\underset{\sim}{\sigma} \leqq \sigma_\varepsilon \leqq \bar{\sigma}, \qquad \underset{\sim}{\alpha} \leqq u_\varepsilon \leqq \bar{\beta} \tag{25}$$

with strict inequalities for nonconstant solutions.

Proof. The existence of a smooth minimizer is classical and is omitted. To establish (25), note that it follows from (24) if u_ε attains its minimum and its maximum in the interior. Indeed, in this case $W'(\max u_\varepsilon) \leqq \sigma_\varepsilon \leqq W'(\min u_\varepsilon)$ and the conclusion follows from Fig. B. In general by reflecting we can extend u_ε periodically on $\mathbb{R}$ so that the extension satisfies (24) everywhere and thus we reduce the situation to that previously considered. Q.E.D.

LEMMA A3. *Let u_ε be a global minimizer. Then we have the estimate*

$$\int |u_\varepsilon'|^2 \leqq \frac{C}{\varepsilon^2}. \tag{26}$$

Proof. The argument is identical to that in Lemma 2, § 2.

LEMMA A4. *Nonmonotonic solutions of* (24) *are not stable.*

Proof. The argument is similar to the proof of Lemma 3. See also [CGS1, p. 348].

LEMMA A5. *Let $\{u_{\varepsilon_n}\}$ be a sequence of* (*monotone increasing*) *global minimizers. Then there is a subsequence, denoted again by $\{u_{\varepsilon_n}\}$ that converges pointwise to a monotone function u which satisfies at points of continuity*

$$W'(u) = \sigma \quad \textit{for some } \sigma \textit{ in } [\underset{\sim}{\sigma}, \bar{\sigma}].$$

Proof. Multiplying (24) by ϕ, a smooth function, and integrating by parts, we obtain ($\varepsilon_n \equiv \varepsilon$)

$$-2\varepsilon^2 \int u_\varepsilon' \phi' = \int W'(u_\varepsilon)\phi - \int \sigma_\varepsilon \phi. \tag{27}$$

By Lemma A3 and the Schwarz inequality we obtain that the left-hand side of (27) converges to zero as $\varepsilon \to 0$. By (24A) we may assume that

$$\sigma_\varepsilon \to \sigma, \qquad \underset{\sim}{\sigma} \leqq \sigma \leqq \bar{\sigma}.$$

By Helly's theorem and the dominated convergence theorem we conclude from (26)

$$\int W'(u)\phi - \int \sigma\phi = 0 \tag{28}$$

and by (22),

$$\int_{-1}^{1} u = M = 2r. \tag{29}$$

The lemma follows from (28). Q.E.D.

Since u is monotone increasing, the only continuous possibilities left in the light of $a_0 < r < \beta_0$ are (see Fig. B):

$$\text{(i)} \qquad u(x) = \begin{cases} \alpha_\sigma & \text{on } [-1, x_1]. \\ z_\sigma & \text{on } (x_1, x_2], \\ \beta_\sigma & \text{on } (x_2, 1], \end{cases}$$

$$\text{(ii)} \qquad u(x) = \begin{cases} \alpha_\sigma & \text{on } [-1, x_1], \\ z_\sigma & \text{on } (x_1, 1], \end{cases}$$

$$\text{(iii)} \qquad u(x) = \begin{cases} \alpha_\sigma & \text{on } [-1, x_1], \\ \beta_\sigma & \text{on } (x_1, 1], \end{cases}$$

$$\text{(iv)} \qquad u(x) = \begin{cases} z_\sigma & \text{on } [-1, x_1], \\ \beta_\sigma & \text{on } (x_1, 1]. \end{cases}$$

By the Weierstrass-Erdmann criterion from these the only one that is a global minimizer for J_0 under the constraint (22) is given in (23).

LEMMA A6. *Let u_ε be as in Lemma* A5. *Then*

$$\overline{\lim_{\varepsilon \to 0}}\, J_\varepsilon(u_\varepsilon) \leqq J_0(\bar{u}). \tag{30}$$

Proof. Consider the function

$$\tilde{u}_\varepsilon(x) = \begin{cases} \alpha_0 & \text{on } [-1, -1 + l_1 - \varepsilon], \\ \beta_0 & \text{on } (1 - l_2 + \varepsilon, 1], \\ \text{linear} & \text{on } (-1 + l_1 - \varepsilon,\ 1 - l_2 + \varepsilon]. \end{cases}$$

Note that $\int_{-1}^{1} \tilde{u}_\varepsilon = M$. A simple computation reveals that $J_\varepsilon(\tilde{u}_\varepsilon) \to \int_{-1}^{1} W(\bar{u}) = J_0(\bar{u})$. Since $J_\varepsilon(u_\varepsilon) \leqq J_\varepsilon(\tilde{u}_\varepsilon)$ we are done. Q.E.D.

Proof of Theorem A1 (*conclusion*). Assume that for some subsequence the limiting state u given by Lemma A5 is different from $\bar{u}$. In the light of the remark following that lemma we may assume that

$$J_0(\bar{u}) < J_0(u) - \delta, \qquad \delta > 0 \quad \text{fixed.} \tag{31}$$

On the other hand by (30) for ε sufficiently small we have

$$J_\varepsilon(u_\varepsilon) \leqq J_0(\bar{u}) + \frac{\delta}{2}. \tag{32}$$

Therefore

$$\varepsilon^2 \int_{-1}^{1} (u_\varepsilon'(x))^2 \leqq \int W(\bar{u}) - \int W(u_\varepsilon) + \frac{\delta}{2}. \tag{33}$$

Since $\int W(u_\varepsilon) \to \int W(u) = J_0(u)$ we conclude via (31) that the right-hand side of (33) becomes negative in the limit, a contradiction. Q.E.D.

Remark A2. In a similar way one can establish the analogue of Theorem A1 for

$$J_\varepsilon(u) = \int_{-1}^{1} [W(u'(x)) + \varepsilon^2 (u''(x))^2]$$

subject to

$$u(-1) = u(1) = 0$$

that models exchanges of phase of a one-dimensional elastic material, originally studied with phase plane techniques by Carr, Gurtin and Slemrod [CGS2].

Acknowledgments. We are indebted to P. Fife for a valuable communication; also we would like to acknowledge a conversation with B. Nicolaenko in the course of which [CGS1] was brought to our attention. Finally we wish to thank H. Simpson for useful discussions on the contents of this paper.

REFERENCES

[C] N. Chafee, *Asymptotic behavior for solutions of a one-dimensional parabolic equation with homogeneous Neumann boundary conditions*, J. Differential Equations, 18 (1975), pp. 111-134.

[CH] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system. I. Interfacial free energy*, J. Chem. Phys., 28, pp. 258-267.

[CH2] G. R. Casten and J. R. Holland, *Instability results for reaction diffusion equations with Neumann boundary conditions*, J. Differential Equations, 27 (1978), pp. 260-273.

[CGS1] J. Carr, M. E. Gurtin and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Rat. Mech. Anal. (1985), pp. 317-351.

[CGS2] ———, *One-dimensional structured phase transformations under prescribed loads*, J. Elast., to appear.

[F] P. C. Fife, *Transition layers in singular perturbation problems*, J. Differential Equations, 15 (1974), pp. 77-106.

[M] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. RIMS, Kyoto Univ., 15 (1979), pp. 401-454.

[N] I. P. Natanson, *Theory of Functions of a Real Variable*, VI, L. Boron, transl., Frederick Unger, New York, 1955.

[NS] A. Novick-Cohen and L. A. Segel, *Nonlinear aspects of the Cahn-Hilliard equation*, Physica D., pp. 277-298.

[S] K. C. Shaing, *Stability of the radial electric field in a nonaxisymmetric torus*, Phys. Fluids, 27 (1984), pp. 1567-1569.

[AS] N. D. Alikakos and H. Simpson, *A variational approach for a class of singular perturbation problems and applications*, Proc. Roy. Soc. Edinburgh, to appear.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

019

FILIPPOV SOLUTIONS TO SINGULAR DIFFERENTIAL EQUATIONS*

SHIVA SHANKAR†

Abstract. This paper considers solutions to singular differential equations in the sense of Filippov. Such singular systems occur in the order reduced models of singularly perturbed systems as well as in the semistate description of nonlinear circuits.

Key words. singular differential equations, Filippov solutions

AMS(MOS) subject classifications. 34E15, 34A08, 34A34

Introduction. A problem often encountered in control and system theory is the modelling of dynamical systems exhibiting a multi-time-scale behaviour. A model that is often employed is the singular perturbation model, where the derivatives of some of the state vectors are multiplied by a small parameter ε, namely

$$(1_\varepsilon)\qquad \begin{aligned} &\dot{x}-f(t,x,y)=0,\\ &\varepsilon\dot{y}-g(t,x,y)=0,\\ &x(t_0)=x_\varepsilon, \qquad y(t_0)=y_\varepsilon \end{aligned}$$

where x is in $\mathbb{R}^n$, y in $\mathbb{R}^p$, and f and g are functions defined on some open subset of $\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^p$ with values in $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. A first step in the analysis and design of controllers for such systems consists of "order reduction" where ε is formally equated to 0. This reduced model

$$(2)\qquad \begin{aligned} &\dot{x}-f(t,x,y)=0,\\ &g(t,x,y)=0,\\ &x(t_0)=x_0, \qquad y(t_0)=y_0, \end{aligned}$$

represents the slow response or the "quasi-steady-state." Note that while the initial conditions in (1_ε) can be arbitrary, in (2) they must satisfy $g(t_0,x_0,y_0)=0$. For more details see Kokotovic, O'Malley and Sannuti [6]. Systems such as (2) also occur in the semistate description of nonlinear circuits [7].

The analysis of the reduced model (2) usually begins with the assumption that at the initial point (t_0,x_0,y_0) satisfying $g(t_0,x_0,y_0)=0$, the partial derivative $(\partial g/\partial y)$ is nonsingular. The implicit function theorem is then employed to determine the (locally) unique manifold $y=y(t,x)$ on which the system (2) is constrained to evolve. Of late there have been attempts to develop a theory when the above assumption of nonsingularity of $(\partial g/\partial y)$ is not satisfied (Rabier and Shankar [8]), or even when g fails to be of class C^1 (Dolezal and Shankar [2]). This note presents another approach to this problem and is motivated by the following considerations.

An important difference between systems (1_ε) and (2), apart from the discrepancy noted above in the assignment of initial values, is that while solutions to (1_ε) are guaranteed for an arbitrary continuous function g, this is not the case as regards system (2). Given that $g(t_0,x_0,y_0)=0$, it may not be that g vanishes on a set whose projection

* Received by the editors November 27, 1983; accepted for publication (in revised form) September 17, 1986.

† Tata Institute of Fundamental Research, Bangalore 560012, India.

to the $\mathbb{R}\times\mathbb{R}^n$ space contains a neighbourhood, say V, of (t_0, x_0). And even if so, the zero set of g may not admit appropriate sections (that is a function $y=y(t,x)$ for (t,x) in V with $g(t,x,y(t,x))=0$) whose substitution in (2) allows the existence of a solution to the resulting system, namely to

$$\dot{x}-f(t,x,y(t,x))=0, \qquad x(t_0)=x_0.$$

This is certainly undesirable, for we would like to consider system (2) as a limit of system (1_ε) as ε tends to 0, and hence would not like to impose any greater regularity on g than is strictly necessary. However minimising regularity requirements on g usually results in a loss of regularity of the solution. In this note we consider solutions to system (2) in the sense of Filippov.

Definition. For D an open set in $\mathbb{R}\times\mathbb{R}^n$, let $h:D\to\mathbb{R}^n$ map (t,x) in D to $h(t,x)$ in $\mathbb{R}^n$. The function $x(t)$ defined on an interval I is called a solution in the sense of Filippov (or a Filippov solution) of the differential equation $\dot{x}-h(t,x)=0$ if $x(t)$ is absolutely continuous and if for almost all t in I and for arbitrary $\delta>0$, the vector $\dot{x}(t)$ belongs to the smallest closed convex set (of $\mathbb{R}^n$) containing all the values of the function $h(t,x)$ where x ranges over almost all of the δ-neighbourhood of the point $x(t)$ (with t fixed). Denoting by konv(E) the convex closure of E, $x(t)$ is then a Filippov solution if for almost all t

$$\dot{x}(t)\in\bigcap_{\delta>0}\ \bigcap_{\mu M=0}\operatorname{konv} f(t, N_\delta(x(t))-M)$$

where $N_\delta(x(t))$ is the δ-ball about $x(t)$ in $\mathbb{R}^n$ and μ is Lebesgue measure.

Filippov solutions to $\dot{x}(t)-h(t,x)=0$ are guaranteed by the following.

THEOREM 1 (Filippov [3]). *Let D be an open set in $\mathbb{R}\times\mathbb{R}^n$ and $h:D\to\mathbb{R}^n$ a locally integrable measurable function. Then for any (t_0,x_0) in D, there exists a Filippov solution to $\dot{x}-h(t,x)=0$ satisfying $x(t_0)=x_0$.*

PROPOSITION. *For X a Polish space and V an open subset of $\mathbb{R}^p$, let $g:X\times V\to\mathbb{R}^p$ be a continuous function such that for some x_0 in X, the partial map $g_{x_0}=g(x_0,\cdot):V\to\mathbb{R}^p$ is injective and whose image contains the origin. Then there exists a neighbourhood N_{x_0} of x_0 in X and a Borel function $h:N_{x_0}\to V$ such that $g(x,h(x))=0$ for all x in N_{x_0}.*

Proof. By the invariance of domain theorem $g(x_0,V)$ is open in $\mathbb{R}^p$ and g_{x_0} is a homeomorphism from V onto $g(x_0,V)$. Choose a closed ball B centered at the origin and contained in $g(x_0,V)$. Let δ be small enough such that for all points v in $B_\delta(v_1)$ for any v_1 in $Bd(B)$—where $B_\delta(v_1)$ is the ball of radius δ with centre v_1 and $Bd(B)$ is the boundary of B—the straight line from 0 through v intersects $Bd(B)$ at a point w with $\|w-v_1\|<$ diameter (B).

Let y_0 and F be the pre-images in V of 0 and B respectively. F is compact and contains y_0. Choose any y in F. Since g is continuous at (x_0,y) there exist neighbourhoods M_y of y and $N_y\subset X$ of x_0 such that

$$\|g(x,y')-g(x_0,y)\|<\frac{\delta}{2} \tag{3}$$

for all x in N_y and y' in M_y. The collection $\{M_y: y\in F\}$ is then an open cover of F, and consequently there exists a finite subcollection $\{M_{y_i}: y_i\in F,\ 1\leqq i\leqq n\}$ which also covers F. Let $N_{x_0}=\bigcap_{i=1}^n N_{y_i}$. Then N_{x_0} is a neighbourhood of x_0 in X.

Given any x in N_{x_0} and y in F, there is an i, $1\leqq i\leqq n$ such that y belongs to M_{y_i}. As x belongs to $N_{x_0}\subset N_{y_i}$, (3) yields

$$\|g(x,y)-g(x_0,y_i)\|<\frac{\delta}{2}.$$

Further as x_0 belongs to N_{y_i},

$$\|g(x_0, y) - g(x_0, y_i)\| < \frac{\delta}{2}.$$

Hence by the triangle inequality

$$\|g(x, y) - g(x_0, y)\| < \delta.$$

Now suppose that 0 is not contained in $g(x, V)$ for some x in N_{x_0}. Let $G: B \to Bd(B)$ be the function which maps u in B to the intersection of $Bd(B)$ with the straight line from 0 through $g(x, g_{x_0}^{-1}(u))$. As 0 is not in $g(x, V)$, G is well defined. Clearly G is continuous. Moreover the choice of δ ensures that $\|G(u) - u\|$ is strictly less than the diameter of B.

G restricted to the boundary of B is homotopic to the identity where the homotopy H moves $G(u)$ to u along the shorter great circle path. Consider $H \circ G: B \to Bd(B)$. This map is a retraction of the ball B onto its boundary, which is absurd (Brouwer's fixed point theorem). Thus for all x in N_{x_0}, $g(x, V)$ contains the origin.

Let $Z_x = \{y \in V: g(x, y) = 0\}$. By the above Z_x is nonempty for each x in N_{x_0}. By the continuity of g_x, each Z_x is closed. The union of the Z_x for x in N_{x_0} is then clearly a Polish subspace of $X \times V$. A standard result now guarantees the existence of a Borel section (Arveson, [1, p. 75, Thm. 3.4.1]) that is a Borel function $h: N_{x_0} \to V$ such that $h(x)$ is in Z_x. But then $g(x, h(x)) = 0$ for all x in N_{x_0}. □

THEOREM 2. *Let $f: D \to \mathbb{R}^n$, $g: D \to \mathbb{R}^p$ be functions defined on an open set $D \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^p$ such that*

(i) *f is a locally integrable Borel map,*

(ii) *g is continuous and for some (t_0, x_0, y_0) in D the partial map $g(t_0, x_0, \cdot)$ is injective with y_0 the (unique) point such that $g(t_0, x_0, y_0) = 0$.*

Then there exists an open interval containing t_0, an absolutely continuous function $x: I \to \mathbb{R}^n$ and a Borel measurable function $y: I \to \mathbb{R}^p$ such that for all t in I $(t, x(t), y(t))$ is in D with

$$\begin{aligned} &\dot{x}(t) - f(t, x(t), y(t)) = 0 \qquad \textit{(in the sense of Filippov)},\\ &g(t, x(t), y(t)) = 0,\\ &x(t_0) = x_0, \qquad y(t_0) = y_0. \end{aligned}$$

Proof. By the above proposition, the projection of the zero set of g onto a subset of $\mathbb{R} \times \mathbb{R}^n$ contains an open neighbourhood N of (t_0, x_0). Further there exists a Borel function $h: N \to \mathbb{R}^p$ such that $g(t, x, h(t, x)) = 0$ for every (t, x) in N. Clearly $h(t_0, x_0) = y_0$.

Define the function $\underset{\sim}{f}: N \to \mathbb{R}^n$ which maps (t, x) in N to $f(t, x, h(t, x))$. Obviously $\underset{\sim}{f}$ is Borel. By replacing D by its intersection with a sufficiently large closed ball, it can be assumed that $h(N)$ is bounded. Hence $\underset{\sim}{f}$ is locally integrable. By Filippov's Theorem, there is an absolutely continuous $x: I \to \mathbb{R}^n$ that is a Filippov solution to

$$\dot{x} - \underset{\sim}{f}(t, x) = 0, \qquad x(t_0) = x_0.$$

Let $y(t) = h(t, x(t))$. Then $y(t)$ is Borel and

$$\begin{aligned} &\dot{x}(t) - f(t, x(t), y(t)) = 0 \quad \text{(in the sense of Filippov)},\\ &g(t, x(t), y(t)) = 0,\\ &x(t_0) = x_0, y(t_0) = y_0. \end{aligned}$$

□

Remark. Although Filippov's Theorem requires only Lebesgue measurability it is necessary to assume that f be Borel measurable in the above theorem. This is because Lebesgue measurability of f and even continuity of h does not imply measurability of $f(t, x, h(t, x))$ (Halmos [4, p. 83]).

The above result can be somewhat sharpened for scalar constrained systems, namely when p is equal to 1, that is, when g is a real valued function.

THEOREM 3. *Let $f: D \to \mathbb{R}^n$, $g: D \to \mathbb{R}$ be functions defined on an open set $D \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$* such that

(i) *f is a locally integrable Borel map*;

(ii) *g is continuous in the (t, x) and y variables separately. For some (t_0, x_0), the image of the partial map $g(t_0, x_0, \cdot)$ contains a neighbourhood of the origin.*

Then given y_0 in $\mathbb{R}$ such that $g(t_0, x_0, y_0) = 0$, there exists an open interval containing t_0, an absolutely continuous function $x: I \to \mathbb{R}^n$ and a Lebesgue measurable function $y: I \to \mathbb{R}$ such that for all t in I $(t, x(t), y(t))$ is in D with

$$\dot{x}(t) - f(t, x(t), y(t)) = 0 \qquad (\textit{in the sense of Filippov}),$$

$$g(t, x(t), y(t)) = 0,$$

$$x(t_0) = x_0, \qquad y(t_0) = y_0.$$

Proof. Choose compact sets T and X in $\mathbb{R}$ and $\mathbb{R}^n$ with nonempty interior that contain the points t_0 and x_0 respectively. Let Y be a compact interval in $\mathbb{R}$ such that $g(t_0, x_0, Y)$ contains a neighbourhood of the origin. Consider g restricted to $T \times X \times Y$. By a result of B. E. Johnson (namely Proposition 2.1 in [5]), $Z = g^{-1}(0)$ is a Borel subset of $T \times X \times Y$. It can be easily verified that the projection of Z onto a subset of $T \times X$ contains a neighbourhood N of (t_0, x_0). By Theorem 3.4.3 in Arveson [1] there exists a Lebesgue measurable function $h: N \to \mathbb{R}$ such that $g(t, x, h(t, x)) = 0$. Also $h(t_0, x_0)$ can be chosen to equal y_0 (as the value at a point can be changed without effecting the measurability of h).

The remainder of the proof now follows in a manner identical to that of the second half of the proof of Theorem 2. □

Acknowledgments. These results appeared in [9]. I am grateful to Professor V. Dolezal for many discussions as well as to Professor A. H. Zemanian for his careful reading of that manuscript.

REFERENCES

[1] W. B. ARVESON, *An invitation to C^*-Algebras*, Graduate Texts in Math., Springer, Berlin-New York, 1976.

[2] V. DOLEZAL AND S. SHANKAR, *A local implicit function theorem and application to systems of differential equations*, Math. Systems Theory, 18 (1985), pp. 251-256.

[3] A. F. FILIPPOV, *Differential equations with discontinuous righthand side*, Amer. Math. Soc. Transl., 2 (1964), pp. 199-231.

[4] P. R. HALMOS, *Measure Theory*, Graduate Texts in Math., Springer, Berlin-New York, 1950.

[5] B. E. JOHNSON, *Separate continuity and measurability*, Proc. Amer. Math. Soc., 20 (1969), pp. 420-422.

[6] P. KOKOTOVIC, R. O'MALLEY AND P. SANNUTI, *Singular perturbations and order reduction in control theory—an overview*, Automatica-J. IFAC, 12 (1976), pp. 123-132.

[7] R. W. NEWCOMB, *The semistate description of nonlinear time variable circuits*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 62-71.

[8] P. RABIER AND S. SHANKAR, *On the number of C^1 quasi-steady-states of singularly perturbed systems near a singular point*, Appl. Anal., to appear.

[9] S. SHANKAR, *Singular ordinary differential equations*, Ph.D. thesis, State Univ. of New York, Stony Brook, NY, 1983.

SIAM J. MATH. ANAL.
Vol. 18, No. 5, September 1987

020

ASYMPTOTIC ANALYSIS OF A SINGULAR PERTURBATION PROBLEM*

SHAGI-DI SHIH† AND R. BRUCE KELLOGG‡

Abstract. We study, in a rectangle $0 < x < a$ and $0 < y < b$, the Dirichlet problem for an elliptic differential equation of the form

$$-\epsilon\Delta u_\epsilon + p\frac{\partial u_\epsilon}{\partial x} + qu_\epsilon = f(x,y)$$

where ϵ is a small parameter $0 < \epsilon \ll 1$, Δ is the Laplacian operator, p is a positive number, q is a nonnegative number and all of the input data are smooth. We establish a constructive procedure for obtaining an asymptotic approximation of arbitrary order with respect to ϵ of this singular perturbation problem, and also give a proof of its uniform validity in the closed rectangle by the use of the maximum principle and exponential estimates of all boundary or corner layer functions. The corner singularities of parabolic boundary layer functions are removed by introducing elliptic boundary layer functions along the characteristic boundaries $y = 0$ and $y = b$. Both ordinary corner layer functions and elliptic corner layer functions are employed at the outflow corners $(a, 0)$ and (a, b).

An application is made to settle a long-standing problem in the magnetohydrodynamic flow in a rectangular duct.

Key words. singular perturbation, outer approximation, ordinary boundary layer, elliptic boundary layer, parabolic boundary layer, ordinary corner layer, elliptic corner layer, maximum principle, magnetohydrodynamics

AMS(MOS) subject classifications. Primary 35B25, 35C20, 35J25; secondary 76W05

1. Introduction.

We study, in a rectangle $\Omega = (0, a) \times (0, b)$, the Dirichlet boundary value problem for an elliptic partial differential equation of the form

$$L_\epsilon u_\epsilon \equiv -\epsilon\Delta u_\epsilon + p\frac{\partial u_\epsilon}{\partial x} + qu_\epsilon = f(x,y) \tag{1.1}$$

with boundary conditions

$$u_\epsilon(0,y) = g_1(y), \qquad u_\epsilon(a,y) = g_2(y), \qquad 0 < y < b, \tag{1.2a,b}$$

$$u_\epsilon(x,0) = g_3(x), \qquad u_\epsilon(x,b) = g_4(x), \qquad 0 < x < a, \tag{1.2c,d}$$

* Received by the editors May 5, 1986; accepted for publication July 30, 1986.

† Department of Mathematics, University of Wyoming, Laramie, Wyoming 82071.

‡ Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

where ϵ is a small parameter $0 < \epsilon \ll 1$, Δ is the Laplacian operator, p is a positive number, q is a nonnegative number, the remaining input data $f(x,y)$, $g_1(y)$, $g_2(y)$, $g_3(x)$, and $g_4(x)$ are assumed to be smooth. We also suppose that the assigned boundary functions are continuous at the corners. That is,

$$g_1(0) = g_3(0), \qquad g_1(b) = g_4(0), \tag{1.3a,b}$$

$$g_2(0) = g_3(a), \qquad g_2(b) = g_4(a). \tag{1.3c,d}$$

The distinguishing feature of a singular perturbation problem is that a small parameter multiplies some terms in the differential equation which, if absent, would change the character of the equation. Often these contain the highest derivatives in the equation and the approximation as this parameter tends to zero is therefore governed by a lower order equation which cannot satisfy all the boundary conditions prescribed. Hence the solution converges nonuniformly in the domain as the parameter tends to zero. Problems of this type frequently arise in fluid dynamics [14], [29], [34], [44], heat transfer [1], [3], theory of plates and shells [30], oil reservoir simulation [38], and magnetohydrodynamic flow [41]. The specific character of the problem (1.1), (1.2a,b,c,d) is brought about by the presence of the four corners of right angle for the domain and by the fact that the parts of the boundary, $y = 0$ and $y = b$, coincide with the characteristic curves of the reduced equation

$$p\frac{\partial u_0}{\partial x} + qu_0 = f(x,y), \tag{1.4}$$

which is obtained from (1.1) by putting $\epsilon = 0$. The boundary $x = 0$ is called the inflow boundary while the boundary $x = a$ is called the outflow boundary.

The purpose of this paper is to establish a constructive procedure for obtaining the asymptotic approximation of arbitrary order with respect to ϵ of this singular perturbation problem, and to give a proof of its uniform validity in the closed rectangle by use of the maximum principle and exponential estimates of all boundary or corner layer functions, which will be defined in Section 3. It is well known [13] that ordinary boundary layer functions appear along the outflow boundary $x = a$ while parabolic boundary layer functions occur along the characteristic boundaries $y = 0$ and $y = b$. In 1966, W. Eckhaus and E. M. De Jager [11] discovered the singularities of parabolic boundary layer functions near the inflow corners $(0,0)$ and $(0,b)$. These corner singularities of parabolic boundary layer functions are removed by introducing elliptic boundary layer functions along the characteristic boundaries. Both ordinary corner layer functions and elliptic corner layer functions are employed at the outflow corners $(a,0)$ and (a,b).

An application is made to settle a long-standing problem in the magnetohydrodynamic flow in a rectangular duct.

We give a brief historical survey of studies on the elliptic singular perturbation problems of the type considered in this work. In providing references we have tried to give those that might be of interest for further reading, rather than presenting a comprehensive bibliography of the subject. We apologize to those who feel neglected.

2. Historical survey. In 1944, W. Wasow [46] studied the problem of the following type

$$-\epsilon\Delta u + \frac{\partial u}{\partial x} = f(x,y) \tag{2.1}$$

over a finite plane domain B with the smooth boundary C under the prescribed boundary condition

$$u = g(x,y) \tag{2.2}$$

on C. Both f and g are smooth functions. It is shown that, as $\epsilon \to 0+$, the solution of the problem (2.1), (2.2) converges to the solution of the reduced equation

$$\frac{\partial u}{\partial x} = f(x,y),$$

assuming the prescribed boundary values along the inflow boundary of C, not the whole boundary.

In 1950, N. Levinson [32] considered the Dirichlet boundary value problem for the equation of the following type:

$$-\epsilon\Delta u + A(x,y)u_x + B(x,y)u_y + C(x,y)u = D(x,y) \tag{2.3}$$

over an open simply or multiply connected region R whose boundary S consists of a finite number of simple closed curves. Let $R \cup S$ be contained in an open connected region R' and suppose all of the data of the problem are smooth in R'. Furthermore, we require that $A^2(x,y) + B^2(x,y) > 0$ in R' and that either R' is simply connected or $C(x,y) > 0$ in R'. Either hypothesis suffices to establish a maximum principle in $R \cup S$. Under these conditions, the Dirichlet boundary value problem for (2.3) has a unique solution in $R \cup S$ for each $\epsilon > 0$.

Let S_1 be a segment of one of curves of S such that $(A,B)\cdot \boldsymbol{n} < 0$, where $\boldsymbol{n}$ is an outward normal vector of S at point (x,y). Let the characteristic curves of the reduced equation corresponding to (2.3) emanating from S_1 pass out of R on the segment S_2 of a curve of S. The closed simply connected region in $R \cup S$ bounded by S_1 and S_2 and by the two characteristics of the reduced equation of (2.3) joining the endpoints of S_1 and S_2 is called a "regular quadrilateral". Then Levinson proved the following result.

THEOREM 2.1. *In a regular quadrilateral Q in $R \cup S$ we have*

$$u(x,y) = u_0(x,y) + w(x,y;\epsilon) + O(\sqrt{\epsilon})$$

uniformly in the quadrilateral. The function $u_0(x,y)$ is the solution of the reduced equation of (2.3) *which takes on the given boundary value of u on S_1. The ordinary boundary layer term $w(x,y;\epsilon)$ is defined as follows:*

$$w(x,y;\epsilon) = \begin{cases} h(x,y)\exp\left[\dfrac{-g(x,y)}{\epsilon}\right], & \textit{near } S_2, \\ \exp\left(-\dfrac{\delta}{\epsilon}\right) & \textit{for some } \delta > 0 \textit{ elsewhere in } Q, \end{cases}$$

where $h = u - u_0$ and $g = 0$ on S_2 and g is positive away from S_2. The function g satisfies the nonlinear equation

$$g_x^2 + g_y^2 + Ag_x + Bg_y = 0$$

(g exists and is uniquely determined in a neighborhood of S_2).

In terminology that we will establish below, Levinson treated the "ordinary boundary layer" part of the solution.

In 1957, M. I. Vishik and L. A. Lyusternik [45] studied the equation of the following type

$$-\epsilon\Delta u + \frac{\partial u}{\partial x} + u = f(x,y) \tag{2.4}$$

in the rectangle Ω, $0 < x < a$ and $0 < y < b$, under the homogeneous boundary condition

$$u = 0$$

on the boundary of Ω. They gave an expansion of the solution of the form for some $\delta > 0$

$$u(x,y) = u_0(x,y) + z_0(x,Y)\,\psi\Big(\frac{y}{\delta}\Big) + z_0^T(x,Y^T)\,\psi\Big(\frac{b-y}{\delta}\Big) + w_0(X_1,y)\,\psi\Big(\frac{a-x}{\delta}\Big) + R_0(x,y;\epsilon)$$

where $Y = y/\sqrt{\epsilon}$, $Y^T = (b-y)/\sqrt{\epsilon}$, $X_1 = (a-x)/\epsilon$ and the infinitely differentiable smoothing function $\psi(x)$ is identically equal to 1 for $x \le \frac{1}{3}$ and equal to zero for $x \ge \frac{2}{3}$. The function u_0 satisfies the reduced equation under the condition $u_0(0,y) = 0$. The function $z_0(x,Y)$ is a boundary layer near the boundary $y = 0$, which satisfies the parabolic equation

$$-\frac{\partial^2 z_0}{\partial Y^2} + \frac{\partial z_0}{\partial x} + z_0 = 0$$

over the semi-infinite region $0 < x < a$, $0 < Y < \infty$ under the conditions

$$z_0(0,Y) = 0, \quad z_0(x,0) = -u_0(x,0).$$

The function z_0^T is a boundary layer near the part of the boundary given by $y = b$, $0 \le x \le a$, which has the same structure as z_0. The function $w_0(X_1,y)$ is a boundary layer along the boundary $x = a$, $0 \le y \le b$, which satisfies the ordinary differential equation

$$\frac{\partial^2 w_0}{\partial {X_1}^2} + \frac{\partial w_0}{\partial X_1} + \epsilon w_0 = 0$$

on the unbounded interval $0 < X_1 < \infty$ under the boundary condition

$$w_0(0,y) = -u_0(a,y) - z_0\Big(a, \frac{y}{\sqrt{\epsilon}}\Big)\,\psi\Big(\frac{y}{\delta}\Big) - z_0^T\Big(a, \frac{b-y}{\sqrt{\epsilon}}\Big)\,\psi\Big(\frac{b-y}{\delta}\Big).$$

From this Vishik and Lyusternik deduced on the basis of the maximum principle [39] that

$$R_0 = O(\epsilon)$$

everywhere except in the neighborhood of the points $(a,0)$ and (a,b), under the assumption that the smooth function $f(x,y)$ vanishes at the points $(0,0)$ and $(0,b)$. They also asserted that on applying an iteration process, one can obtain, as above, an asymptotic formula of arbitrary order if the parameters of the problem are sufficiently smooth. In 1966, W. Eckhaus and E. M. De Jager [11] made the remark that the extension of the theory involving parabolic boundary layers to higher order approximation should not be considered a trivial matter.

In our terminology, Vishik and Lyusternik treated the "parabolic boundary layers" along the characteristic boundaries of the region.

In 1964, J. K. Knowles and R. E. Messick [30] discussed a class of singular perturbation problems arising in the theory of thin elastic plates and shells. An important feature of these problems is that the boundary of the domain coincides either partially or entirely with portions of characteristic curves of the reduced equations. In order to understand this exceptional character of a characteristic boundary, Knowles and Messick considered the equation of the following type

$$-\epsilon\Delta u + \frac{\partial u}{\partial x} = 0 \tag{2.5}$$

in the semi-infinite strip $R: 0 < x < a$ and $0 < y < \infty$, under the boundary conditions

$$u(0,y) = 0, \quad u(a,y) = 0, \quad 0 < y < \infty, \tag{2.6a, b}$$

$$u(x,0) = g(x), \quad 0 \leq x \leq a, \tag{2.6c}$$

where the function u and its partial derivatives are required to be bounded as $y \to \infty$. They obtained as an approximation to the solution the function

$$z_0(x,Y) + W_0(X_1,Y) + V_0(X_1,Y_1)$$

where $Y = y/\sqrt{\epsilon}$, $X_1 = (a-x)/\epsilon$, and $Y_1 = y/\epsilon$. Note that for this problem, the solution of the reduced equation under the condition $u_0(0,y) = 0$ is $u_0(x,y) = 0$, which also satisfies the boundary condition (2.6b). Therefore the ordinary boundary layer along the boundary $x = a$ is identical to zero. The function $z_0(x,Y)$ is the parabolic boundary layer function along the boundary $y = 0$, defined by the equation

$$-\frac{\partial^2 z_0}{\partial Y^2} + \frac{\partial z_0}{\partial x} = 0$$

over the domain $0 < x < a$, $0 < Y < \infty$ under the conditions

$$z_0(x,0) = g(x), \quad z_0(0,Y) = 0,$$

and the condition that $z_0(x,Y)$ decays exponentially in Y. The function $W_0(X_1,Y)$ is a "corner layer" at the outflow corner $(a,0)$, defined by the ordinary differential equation

$$\frac{\partial^2 W_0}{\partial {X_1}^2} + \frac{\partial W_0}{\partial X_1} = 0$$

on the unbounded interval $0 < X_1 < \infty$ under the conditions

$$W_0(0,Y) = -z_0(a,Y)$$

and

$W_0(X_1,Y)$ has the exponential decay property in both X_1 and Y.

Then it follows that

$$W_0(X_1,Y) = -z_0(a,Y)\exp(-X_1).$$

The function $V_0(X_1,Y_1)$ is a "corner layer" at the outflow corner $(a,0)$, defined by the elliptic equation

$$-\Big(\frac{\partial^2 V_0}{\partial {X_1}^2} + \frac{\partial^2 V_0}{\partial {Y_1}^2}\Big) + \frac{\partial V_0}{\partial X_1} = 0$$

over the quarter-plane $0 < X_1 < \infty$ and $0 < Y_1 < \infty$ under the conditions

$$\begin{aligned} &V_0(X_1,0) = -W_0(X_1,0) = g(a)\exp(-X_1),\\ &V_0(0,Y_1) = 0, \end{aligned}$$

and

$V_0(X_1,Y_1)$ decays exponentially in X_1 and Y_1.

Knowles and Messick mentioned that using a representation of the solution to the original boundary value problem (2.5), (2.6a,b,c), which involves the expansion of the Green's function in a series of modified Bessel functions, it is possible to prove the statements listed below. First let us define some notation: Let S be the boundary of R. δ denotes an arbitrary small positive number. The symbols ${}_\delta D$ and D_δ represent quarter-discs at $(0,0)$ and $(a,0)$, respectively, and are defined as follows:

$$\begin{aligned} {}_\delta D &= \{(x,y) : x \geq 0,\ y \geq 0,\ x^2 + y^2 < \delta^2\},\\ D_\delta &= \{(x,y) : x \geq 0,\ y \leq a,\ x^2 + (y-a)^2 < \delta^2\}. \end{aligned}$$

It can be shown that

i) There exist positive constants $M(\delta)$, $c(\delta)$ such that

$$|u(x,y)| \leq M(\delta)\exp\Big(-\frac{c(\delta)}{\epsilon}\Big)$$

for all $y \geq \delta$, $0 \leq x \leq a$.

ii) There exists a positive constant $M(\delta)$ such that

$$\Big|u(x,y) - z_0\Big(x, \frac{y}{\sqrt{\epsilon}}\Big)\Big| < M(\delta)\epsilon \tag{2.7}$$

for all $(x,y) \in R \cup S - {}_\delta D - D_\delta$.

iii) If $g(x) = O(x)$ as $x \to 0+$, then there exists a constant $M(\delta) > 0$ such that (2.7) holds for $(x,y) \in R \cup S - D_\delta$.

iv) If $g(x) = O(x)$ as $x \to 0+$ and $g(x) = O(a-x)$ as $x \to a-$, then there exists a positive constant M such that

$$\left|u(x,y) - z_0\left(x, \frac{y}{\sqrt{\epsilon}}\right) - W_0\left(\frac{a-x}{\epsilon}, \frac{y}{\sqrt{\epsilon}}\right)\right| \leq M\epsilon$$

for all $(x,y) \in R \cup S$.

In our terminology, Knowles and Messick have thus treated the "ordinary corner layer" and the "elliptic corner layer" at the outflow corner of the region.

In 1966, W. Eckhaus and E. M. De Jager [11] investigated the problem of the following type

$$L_\epsilon u \equiv -\epsilon\Delta u + \frac{\partial u}{\partial x} = 0 \tag{2.8}$$

in the domain Ω, $0 < x < a$ and $0 < y < b$, with the boundary conditions

$$\begin{array}{llll} \text{(2.9a,b)} & u(0,y) = g_1(y), & u(a,y) = g_2(y), & 0 < y < b, \\ \text{(2.9c,d)} & u(x,0) = g_3(x), & u(x,b) = g_4(x), & 0 < x < a. \end{array}$$

The boundary data are assumed to be smooth and to be continuous at the corner points. The solution of the reduced problem is $g_1(y)$, which does not satisfy the given boundary conditions (2.9b,c,d) generally. The local variable $Y = y/\sqrt{\epsilon}$ along the characteristic boundary $y = 0$ transforms (2.8) into

$$-\frac{\partial^2 u}{\partial Y^2} + \frac{\partial u}{\partial x} = \epsilon\frac{\partial^2 u}{\partial x^2}.$$

Define the function $z_0(x,Y)$ as the solution of the reduced equation in local coordinates, that is, we have

$$-\frac{\partial^2 z_0}{\partial Y^2} + \frac{\partial z_0}{\partial x} = 0$$

in the domain $0 < x < a$ and $0 < Y < \infty$. Let z_0 satisfy boundary conditions

$$\begin{aligned} &z_0(0,Y) = 0, \\ &z_0(x,0) = g_3(x) - g_1(0) \equiv \gamma(x). \end{aligned}$$

An explicit form of this function is easily obtained as follows:

$$z_0(x,Y) = \sqrt{\frac{2}{\pi}}\int_{Y/\sqrt{2x}}^{\infty} \exp\left(-\frac{t^2}{2}\right)\gamma\left(x - \frac{Y^2}{2t^2}\right)dt.$$

Furthermore, we have

$$L_\epsilon z_0 = -\epsilon\frac{\partial^2 z_0}{\partial x^2}. \tag{2.10}$$

Due to the assumed continuity of the boundary data, i.e., $\gamma(0) = 0$, we obtain

$$\frac{\partial^2 z_0(x,Y)}{\partial x^2} = \sqrt{\frac{2}{\pi}}\Big[\int_{Y/\sqrt{2x}}^{\infty} \exp\Big(-\frac{t^2}{2}\Big)\,\gamma''\Big(x - \frac{Y^2}{2t^2}\Big)\,dt$$
$$+ \gamma'(0)\,Y\,(2x)^{-3/2}\exp\Big(-\frac{Y^2}{4x}\Big)\Big],$$

where the primes indicate derivatives with respect to the argument. We see that $\partial^2 z_0/\partial x^2$ is uniformly bounded in the closure of Ω if and only if $\gamma'(0) = 0$, which was assumed to be true in the analysis of Vishik and Lyusternik [45]. In the case of general boundary conditions, where $\gamma'(0) \neq 0$, the right-hand side of (2.10) has a singularity at the origin $x = 0$, $y = 0$. The nature of the singularity is most clearly revealed if in (2.10) the origin is approached along any curve $Y = mx^n$, where m and n are constants. The presence of this "corner singularity" indicates that in attempting a proof of the asymptotic properties of the parabolic boundary layer, difficulties should be expected. Also this corner singularity gives more singular functions in the course of the construction of a high order approximation to the parabolic boundary layer.

Eckhaus and De Jager constructed a regularized parabolic boundary layer function $\bar{z}_0$ by replacing $\gamma(x)$ in z_0 by

$$\bar{\gamma}(x) = \gamma(x) - \gamma'(0)\,x\exp\Big(-\frac{x}{\sqrt{\epsilon}}\Big).$$

In other words,

$$\bar{z}_0(x,Y) = \sqrt{\frac{2}{\pi}}\int_{Y/\sqrt{2x}}^{\infty} \exp\Big(-\frac{t^2}{2}\Big)\,\bar{\gamma}\Big(x - \frac{Y^2}{2t^2}\Big)\,dt.$$

The important properties of this function $\bar{\gamma}(x)$ are

i) $\bar{\gamma}(0) = \bar{\gamma}'(0) = 0$, which implies that $\partial^2 \bar{z}_0/\partial x^2$ is bounded uniformly in the closure of Ω, and moreover,

$$\frac{\partial^2 \bar{z}_0}{\partial x^2} = O(1) + O\Big(\epsilon^{-1/2}\Big)$$

in the closure of Ω.

ii) $\bar{\gamma}(x) = \gamma(x) + O(\sqrt{\epsilon})$, which implies that

$$\bar{z}_0(x,Y) = z_0(x,Y) + O\big(\sqrt{\epsilon}\big)$$

uniformly in the closure of Ω.

Define another regularized parabolic boundary layer function $\bar{z}_0^T$ along the boundary $y = b$ similarly. Introduce the local coordinate $X_1 = (a - x)/\epsilon$ along the boundary $x = a$ and define the boundary layer function $\tilde{w}_0(X_1, y)$ as the solution of the ordinary differential equation

$$\frac{\partial^2 \tilde{w}_0}{\partial {X_1}^2} + \frac{\partial \tilde{w}_0}{\partial X_1} = 0$$

over the unbounded interval $0 < X_1 < \infty$ under the conditions

$$\begin{aligned}\tilde{w}_0(0,y) &= g_2(y) - g_1(y) - \bar{z}_0\Big(a, \frac{y}{\sqrt{\epsilon}}\Big) - \bar{z}_0^T\Big(a, \frac{b-y}{\sqrt{\epsilon}}\Big)\\ &\equiv \psi(y),\end{aligned}$$

and

$$\lim_{X_1\to\infty} \tilde{w}_0(X_1, y) = 0.$$

Then it follows that

$$\tilde{w}_0(X_1, y) = \psi(y)\exp(-X_1).$$

Eckhaus and De Jager first showed by using the maximum principle [39] that the solution u of the boundary value problem (2.8) with the boundary conditions (2.9a,b,c,d) has the asymptotic expansion

$$u(x,y) = g_1(y) + \bar{z}_0\Big(x, \frac{y}{\sqrt{\epsilon}}\Big) + \bar{z}_0^T\Big(x, \frac{b-y}{\sqrt{\epsilon}}\Big) + \tilde{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) + R_0(x, y; \epsilon),$$

where the remainder R_0 satisfies

$$R_0 = O(\sqrt{\epsilon})$$

uniformly in the closure of Ω with the exception of a neighborhood of the two corner points $(a, 0)$ and (a, b). Note that at $y = 0$ and $y = b$, $\psi''(y)$ is bounded but of order of $1/\epsilon$. To get a better estimate for the remainder term, they introduced the expansion

$$u(x,y) = g_1(y) + \bar{z}_0\Big(x, \frac{y}{\sqrt{\epsilon}}\Big) + \bar{z}_0^T\Big(x, \frac{b-y}{\sqrt{\epsilon}}\Big) + \bar{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) + \bar{R}_0(x, y; \epsilon),$$

with

$$\bar{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) = [\psi(y) + \epsilon(a-x)\,\psi''(y)]\exp\Big(-\frac{a-x}{\epsilon}\Big)$$

and arrived at the conclusion that

$$\bar{R}_0 = O(\sqrt{\epsilon})$$

uniformly in the closure of Ω on the basis of the maximum principle. Furthermore they showed that

$$\bar{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) = \tilde{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) + O(\epsilon)$$

uniformly in the closure of Ω. Summarizing these results, Eckhaus and De Jager have established the following theorem.

THEOREM 2.2. *The asymptotic approximation for* $u(x,y)$

$$u(x,y) = g_1(y) + z_0\Big(x, \frac{y}{\sqrt{\epsilon}}\Big) + z_0^T\Big(x, \frac{b-y}{\sqrt{\epsilon}}\Big) + \tilde{w}_0\Big(\frac{a-x}{\epsilon}, y\Big) + O(\sqrt{\epsilon})$$

holds uniformly in the closure of Ω*, including the four corner points.*

We remark that the order of the asymptotic error in the above theorem is determined by the presence of corner singularities. If one studies the exceptional case $g_3'(0) = 0$, $g_4'(0) = 0$ in which the corner singularities are absent, one finds along the lines of the preceding analysis that the asymptotic error is no longer $O(\sqrt{\epsilon})$ but $O(\epsilon)$. J. Mauss [35], [36] claimed to improve the estimate of the above theorem and to obtain an $O(\epsilon)$ remainder by using a rather special inequality for the maximum principle when the right-hand side of the differential operator L_ϵ has the singular behavior. The result asserted by Mauss is inconsistent with higher order expansions obtained in this paper.

All of the ideas discussed here need to be modified in order to develop higher order asymptotic expansions. The difficulties involved are

i) Corner singularities appear at the inflow corners of the region in the construction of the parabolic boundary layers along the characteristic boundaries $y = 0$ and $y = b$.

ii) The parabolic boundary layer and ordinary boundary layer overlap in the vicinities of the outflow corners of the region.

The proper way to treat the item ii) is to follow the construction of the ordinary corner layer and the elliptic corner layer done by Knowles and Messick [30] after the modification of the ordinary boundary layer. More precisely, the function $\bar{w}_0(X_1, y)$ in the construction due to Eckhaus and De Jager may be decomposed as

$$\begin{aligned}\bar{w}_0(X_1, y) &= w_0(X_1, y) + W_0(X_1, Y) + W_0^T(X_1, Y^T) \\ &\quad + \epsilon[W_1(X_1, Y) + W_1^T(X_1, Y^T)] + \epsilon^2 w_2(X_1, y),\end{aligned}$$

where the ordinary boundary layer functions $w_0(X_1, y)$ and $w_2(X_1, y)$ satisfy the equations

$$\frac{\partial^2 w_i}{\partial X_1{}^2} + \frac{\partial w_i}{\partial X_1} = \begin{cases} 0, & i = 0, \\ -\dfrac{\partial^2 w_{i-2}}{\partial y^2}, & i = 2, \end{cases}$$

over the unbounded interval $0 < X_1 < \infty$ under the conditions

$$w_i(0, y) = \begin{cases} g_2(y) - g_1(y), & i = 0, \\ 0, & i = 2, \end{cases}$$

and

$$w_i(X_1, y) \text{ decays exponentially as } X_1 \to \infty;$$

the ordinary corner layer functions $W_0(X_1, Y)$ and $W_1(X_1, Y)$ satisfy the equations

$$\frac{\partial^2 W_i}{\partial X_1{}^2} + \frac{\partial W_i}{\partial X_1} = \begin{cases} 0, & i = 0, \\ -\dfrac{\partial^2 W_{i-1}}{\partial Y^2}, & i = 1, \end{cases}$$

over the unbounded interval $0 < X_1 < \infty$ under the conditions

$$W_i(0, Y) = \begin{cases} -\bar{z}_0(a, Y), & i = 0, \\ 0, & i = 1, \end{cases}$$

and

$$W_i(X_1, Y) \text{ decays exponentially as } X_1 \to \infty.$$

The functions W_0^T and W_1^T are defined similarly. It is clear that the term w_2 plays no role in this improvement of estimate.

Eckhaus and De Jager investigated the singularities of parabolic boundary layers near the inflow corners, obtained a formal approximation, and proved the uniform validity of the asymptotic approximation to the solution with the estimate of $O(\sqrt{\epsilon})$. Since then, the problem of dealing with these singularities has attracted much attention from mathematicians, e.g., J. Grasman [15] – [18]; L. P. Cook, G. S. S. Ludford and J. S. Walker [9]; L. P. Cook and G. S. S. Ludford [10]; A. M. Il'in and E. F. Lelikova [24]; V. F. Butuzov [4]; and D. J. Temperley [43]. This type of difficulty also takes place in the problems of the interior layers due to the nonsmooth boundary conditions [8], [20], [36], the domains with corners and noncharacteristic boundary [20], and the nonconvex domains [20], [36].

A similar phenomenon occurs in the problem

$$-(\epsilon u_{yy} + u_{xx}) + u_y = f(x, y), \quad 0 < \epsilon \ll 1,$$

defined on the unit square T, $0 < x < 1$ and $0 < y < 1$, with the prescribed Dirichlet boundary conditions. In the event that the solution of this problem converges to the solution of the reduced problem, we can anticipate the ordinary boundary layer behavior in the vicinity of the upper edge of T, since the solution $u_0(x, y)$ of the reduced equation is uniquely determined throughout T by the boundary values assumed on the other three edges. Note that the reduced equation is a partial differential equation of parabolic type. Thus when one attempts to determine an asymptotic solution, one has to treat the corner singularities of the derivative $\partial^2 u_0/\partial y^2$ first. Such singularities always exist no matter how smooth the prescribed boundary conditions may be unless certain compatibility conditions hold. Furthermore, the construction of the ordinary boundary layer requires the smoothness of the derivative $\partial^2 u_0/\partial x^2$ at point $(x, 1)$. This type of problem had been studied to obtain an asymptotic approximation of arbitrary order with respect to ϵ by G. E. Latta [31], R. E. O'Malley, Jr. [37], and Peng-Cheng Lin and Fa-Wang Liu [33]. None of these authors have treated the difficulties mentioned above.

J. Grasman [15] studied the problem of the following type

$$L_\epsilon u \equiv -\epsilon \Delta u + \frac{\partial u}{\partial x} = 0 \tag{2.11}$$

over the quarter-plane $0 < x < \infty$ and $0 < y < \infty$ under the boundary conditions

$$u(x, 0) = g(x), \qquad g(0) = 0, \tag{2.12a}$$

$$u(0, y) = 0, \tag{2.12b}$$

by means of the Green's theorem, which yields the exact solution. The asymptotic expansion of the integral representation of the solution is shown to be

$$\begin{aligned}u(x,y) &= \sqrt{\frac{2}{\pi}}\int_{y/\sqrt{2\epsilon x}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\, g\Big(x-\frac{y^2}{2\epsilon t^2}\Big)\,dt\\ &\quad+\epsilon\Big\{\tilde v_0\Big(\frac{x}{\epsilon},\frac{y}{\epsilon}\Big)-g'(0)\,\epsilon^{-1}\sqrt{\frac{2}{\pi}}\int_{y/\sqrt{2\epsilon x}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\cdot\Big(x-\frac{y^2}{2\epsilon t^2}\Big)\,dt\\ &\quad+\sqrt{\frac{2}{\pi}}\int_{y/\sqrt{2\epsilon x}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\frac{t^2-1}{2}\Big[g'\Big(x-\frac{y^2}{2\epsilon t^2}\Big)-g'(0)\Big]\,dt\Big\}\\ &\quad+O(\epsilon^2)\end{aligned}$$

uniformly for $x \geq 0$, $y \geq 0$, where the term $\epsilon\tilde v_0(x/\epsilon, y/\epsilon)$ represents the solution of boundary value problem (2.11), (2.12a,b) in the case that $g(x) = xg'(0)$. In order to understand this uniformly valid expansion from another viewpoint, let us define the functions $v_0(x/\epsilon, y/\epsilon;\epsilon)$, $z_0(x, y/\sqrt{\epsilon})$, and $z_1(x, y/\sqrt{\epsilon})$ by

$$v_0\Big(\frac{x}{\epsilon},\frac{y}{\epsilon};\epsilon\Big)=\epsilon\tilde v_0\Big(\frac{x}{\epsilon},\frac{y}{\epsilon}\Big),$$

$$z_0\Big(x,\frac{y}{\sqrt{\epsilon}}\Big)=\sqrt{\frac{2}{\pi}}\int_{y/\sqrt{2\epsilon x}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\Big[g\Big(x-\frac{y^2}{2\epsilon t^2}\Big)-g'(0)\cdot\Big(x-\frac{y^2}{2\epsilon t^2}\Big)\Big]\,dt,$$

and

$$z_1\Big(x,\frac{y}{\sqrt{\epsilon}}\Big)=\sqrt{\frac{2}{\pi}}\int_{y/\sqrt{2\epsilon x}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\frac{t^2-1}{2}\Big[g'\Big(x-\frac{y^2}{2\epsilon t^2}\Big)-g'(0)\Big]\,dt,$$

respectively. In other words, it follows that

$$u(x,y)=z_0\Big(x,\frac{y}{\sqrt{\epsilon}}\Big)+\epsilon\, z_1\Big(x,\frac{y}{\sqrt{\epsilon}}\Big)+v_0\Big(\frac{x}{\epsilon},\frac{y}{\epsilon};\epsilon\Big)+O(\epsilon^2)\,.$$

From a direct computation, one finds that

i) The functions $z_0(x,Y)$ and $z_1(x,Y)$ are the solutions of the equations

$$-\frac{\partial^2 z_k}{\partial Y^2}+\frac{\partial z_k}{\partial x}=\begin{cases}0, & k=0,\\ \dfrac{\partial^2 z_{k-1}}{\partial x^2}, & k=1,\end{cases}$$

over the quarter plane $0 < x < \infty$ and $0 < Y < \infty$ under the conditions

$$z_k(0,Y)=0,$$

$$z_k(x,0)=\begin{cases}g(x)-g'(0)\,x, & k=0,\\ 0, & k=1,\end{cases}$$

and

$$z_k(x,Y)\to 0 \qquad \text{as } x^2+Y^2\to\infty \text{ (and } x>0 \text{ when } k=0).$$

ii) The function $v_0(X, Y_1; \epsilon)$ is the solution of the elliptic partial differential equation

$$-\left(\frac{\partial^2 v_0}{\partial X^2} + \frac{\partial^2 v_0}{\partial {Y_1}^2}\right) + \frac{\partial v_0}{\partial X} = 0$$

over the quarter-plane $0 < X < \infty$ and $0 < Y_1 < \infty$ under the conditions

$$v_0(0, Y_1; \epsilon) = 0, \quad v_0(X, 0; \epsilon) = g'(0)X\epsilon,$$

and

$$v_0(X, Y_1; \epsilon) \to 0 \qquad \text{as } X^2 + Y_1^2 \to \infty \text{ and } X > 0.$$

This means that a uniformly valid approximation with an accuracy $O(\epsilon^2)$ can be constructed by the usual perturbation method even though Grasman claimed that this cannot be done [15]. The detailed construction of such functions v_k, which will be called as the "elliptic boundary layer" functions, is in Section 3.3. It will be clear later that $v_k(x/\epsilon, y/\epsilon; \epsilon) \equiv 0$, for $k \geq 1$ in the problem (2.11), (2.12a,b). For the sake of convenience, let us write the function z_0 as

$$z_0 = z_0^1 - z_0^2$$

where z_0^1 denotes the integral which contains the function $g(\cdot)$ as part of integrand and z_0^2 is the remaining part of z_0. Grasman [15], [16] asserted that

i) the function z_0^1 is a uniformly valid approximation of the solution of problem (2.11), (2.12a,b) with a remainder term $O(\epsilon)$ for $x \geq 0$, $y \geq 0$, and

ii) the estimate $v_0 - z_0^2 = O(\epsilon)$ holds uniformly for $x \geq 0$, $y \geq 0$.

We cannot draw these conclusions because of the singularity of z_0^2 at the origin $x = 0$, $y = 0$. For it is clear that $v_0 - z_0^2$ vanishes when $x = 0$ or $y = 0$, but

$$L_\epsilon(v_0 - z_0^2) = \epsilon\frac{\partial^2 z_0}{\partial x^2} = g'(0)\sqrt{\frac{\epsilon}{4\pi}}\, y\, x^{-3/2}\exp\left(-\frac{y^2}{4\epsilon x}\right)$$

is singular in the neighborhood of the origin.

Cook and Ludford [10] studied the problem on a semi-infinite strip and analyzed the asymptotic approximation from an exact representation of the solution obtained by means of the Fourier sine transforms. Il'in and Lelikova [24] used the asymptotic behavior of the solution at the inflow corners to obtain the uniqueness of the parabolic boundary layer functions by matching two different asymptotic expansions. Butuzov [4] imposed certain compatibility conditions on the input data so that the corner singularities of the parabolic boundary layers disappear, and employed the corner layers at the outflow corners to obtain the asymptotic approximations of arbitrary order with respect to ϵ, which is proved to be valid uniformly in the closure of the rectangular region.

3. Main results.

The main tool used in this paper for estimating solutions of elliptic boundary value problems is furnished by the so-called maximum principle and the concept of barrier function. For the proof of the maximum principle see Eckhaus and De Jager [11]. We now repeat the formulation of the boundary layer problem:

$$L_\epsilon u_\epsilon \equiv -\epsilon \Delta u_\epsilon + p\frac{\partial u_\epsilon}{\partial x} + qu_\epsilon = f(x,y) \quad \text{in } \Omega, \tag{1.1}$$

with boundary conditions

$$u_\epsilon(0,y) = g_1(y), \quad u_\epsilon(a,y) = g_2(y), \quad 0 < y < b, \tag{1.2a, b}$$

$$u_\epsilon(x,0) = g_3(x), \quad u_\epsilon(x,b) = g_4(x), \quad 0 < x < a, \tag{1.2c, d}$$

where ϵ is a small parameter $0 < \epsilon \ll 1$, Δ is the Laplacian operator, p is a positive number, q is a nonnegative number, Ω is the rectangular region $0 < x < a$ and $0 < y < b$, and the remaining input data $f(x,y)$, $g_1(y)$, $g_2(y)$, $g_3(x)$, and $g_4(x)$ are assumed to be smooth. We also suppose that the assigned boundary functions are continuous at the corners. We are ready to state the maximum principle.

MAXIMUM PRINCIPLE. *Let Φ and Ψ be twice continuously differentiable functions in Ω such that*

$$|L_\epsilon[\Phi]| \le L_\epsilon[\Psi] \quad \text{in } \Omega,$$

$$|\Phi| \le \Psi \quad \text{on } \partial\Omega.$$

Then

$$|\Phi| \le \Psi \quad \text{in } \bar{\Omega}.$$

Remark 3.1. For an elliptic differential operator of second order in an unbounded domain a maximum principle holds if the solution satisfies a certain growth condition at infinity. We will discuss this in Section 3.3.

In this work we investigate an asymptotic approximation of the solution of the elliptic boundary value problem given by (1.1), (1.2a,b,c,d). From the assumption that q is nonnegative in Ω, it follows that the solution is unique. Under the conditions assumed, it is well known that for a fixed value of ϵ, $u_\epsilon(x,y)$ is continuous in $\bar{\Omega}$ and is smooth in Ω_1 where Ω_1 is any compact subregion of $\bar{\Omega}$ with positive distance from the corners. In 1979, A. Azzam [2] improved this result to obtain the following:

THEOREM 3.1. *For any fixed value of ϵ, there exists a number $\nu \in (1,2)$ such that the solution u_ϵ of the Dirichlet boundary value problem* (1.1), (1.2a,b,c,d) *with the assumptions stated at the beginning of this section satisfies $u_\epsilon(x,y) \in C_\nu(\bar{\Omega})$. Moreover, in a sufficiently small neighborhood of the corner, $r^\tau D^2 u_\epsilon \in C_\mu$ for some τ, $\mu \in (0,1)$, where r is the distance from the corner point to (x,y) and $D^2 u_\epsilon$ is any second partial derivative of u_ϵ.*

Remark 3.2. The maximum principle implies that the solution $u_\epsilon(x,y)$ is bounded uniformly with respect to ϵ in $\bar{\Omega}$.

3.1. Outer approximation.

In order to obtain the first rough approximation of the solution $u_\epsilon(x,y)$ for small values of the parameter ϵ, we consider a function $u_0(x,y)$ which satisfies the reduced equation

$$p\frac{\partial u_0}{\partial x} + qu_0 = f(x,y). \tag{1.4}$$

The function $u_0(x,y)$ can satisfy only one of the prescribed boundary conditions

$$u_0(0,y) = g_1(y), \tag{1.2a}$$

and

$$u_0(a,y) = g_2(y). \tag{1.2b}$$

THEOREM 3.2. *There exists a positive constant C independent of ϵ such that the inequality*

$$\left|u_\epsilon(x,y) - g_1(y)\right| \le Cx \tag{3.1}$$

holds uniformly in the closure of Ω for all values of ϵ.

Proof. The function $\Phi_\epsilon(x,y)$, defined by

$$\Phi_\epsilon(x,y) \equiv u_\epsilon(x,y) - g_1(y),$$

satisfies in Ω the differential equation

$$L_\epsilon[\Phi_\epsilon] = f(x,y) + \epsilon g_1''(y) - qg_1(y)$$

with the boundary conditions

$$\begin{aligned}
&\Phi_\epsilon(0,y) = 0,\\
&\Phi_\epsilon(a,y) = g_2(y) - g_1(y),\\
&\Phi_\epsilon(x,0) = g_3(x) - g_1(0) = g_3(x) - g_3(0),\\
&\Phi_\epsilon(x,b) = g_4(x) - g_1(b) = g_4(x) - g_4(0).
\end{aligned}$$

We now introduce the barrier function $\Psi(x) = Cx$, where C is some positive constant independent of ϵ. By taking C sufficiently large it follows that the inequalities

$$\left|\Phi_\epsilon(x,y)\right| \le \Psi(x)$$

on the boundary of Ω and

$$\left|L_\epsilon[\Phi_\epsilon]\right| \le L_\epsilon[\Psi]$$

in Ω can be satisfied for all values of ϵ. Applying the maximum principle, we get the desired inequality (3.1) uniformly valid in the closure of Ω for all values of ϵ. This completes the proof.

According to Theorem 3.2, as ϵ tends to zero, we are led to the inequality

$$\left|u_0(x,y) - g_1(y)\right| \le Cx.$$

Therefore (1.2a) is the proper condition for the solution $u_0(x,y)$. Now the function u_0 is easily determined, and the result is

$$u_0(x,y) = g_1(y)\exp\Bigl(-\frac{qx}{p}\Bigr) + p^{-1}\int_0^x f(s,y)\exp\Bigl[-\frac{(x-s)q}{p}\Bigr]\,ds.$$

Since the remaining boundary conditions (1.2b,c,d) are not satisfied by the function $u_0(x,y)$ and the difference $u_\epsilon - u_0$ satisfies in Ω the differential equation

$$L_\epsilon[u_\epsilon - u_0] = \epsilon \Delta u_0,$$

it is quite evident that this approximation for $u_\epsilon(x,y)$ is not valid in a neighborhood of three parts of the boundary of Ω, $x = a$, $y = 0$, and $y = b$, and is valid in the remaining subregion of Ω including the neighborhood of the inflow boundary $x = 0$ up to the order $O(\epsilon)$.

First of all, a better approximation of the solution $u_\epsilon(x,y)$ in this subregion is to be obtained by adding to $u_0(x,y)$ a sum $\sum_{k=1}^{n} \epsilon^k u_k(x,y)$ which equals zero along the inflow boundary $x = 0$ and has the property that $\sum_{k=0}^{n} \epsilon^k u_k(x,y)$ satisfies (1.1) up to the order $O(\epsilon^{n+1})$ for some positive integer n. The functions $u_k(x,y)$ are determined by iteration from the differential equations

$$p\frac{\partial u_k}{\partial x} + q u_k = \Delta u_{k-1}, \tag{3.2}$$

with the boundary conditions

$$u_k(0,y) = 0, \tag{3.3}$$

for $k = 1, 2, \ldots, n$. Therefore it follows that the functions $u_k(x,y)$ are given by the expressions

$$u_k(x,y) = p^{-1} \int_0^x \Delta u_{k-1}(s,y) \exp\left[-\frac{(x-s)q}{p}\right] ds$$

for $k = 1, 2, \ldots, n$.

We will call the series

$$u(x,y;\epsilon) = \sum_{k=0}^{n} \epsilon^k u_k(x,y) \tag{3.4}$$

the outer asymptotic approximation (OA). (Other names that are given are the asymptotic approximation or the interior asymptotic approximation of $u_\epsilon(x,y)$ in Ω.) Applying the differential operator L_ϵ to the function $u^*(x,y;\epsilon)$, defined by

$$u^* = u_\epsilon - u, \tag{3.5}$$

yields the differential equation in Ω

$$L_\epsilon[u^*] = \epsilon^{n+1} \Delta u_n \tag{3.6}$$

with the boundary conditions

$$u^*(0,y;\epsilon) = 0, \tag{3.7a}$$

$$u^*(a,y;\epsilon) = g_2(y) - u_0(a,y) - \sum_{k=1}^{n} \epsilon^k u_k(a,y), \tag{3.7b}$$

$$u^*(x,0;\epsilon) = g_3(x) - u_0(x,0) - \sum_{k=1}^{n} \epsilon^k u_k(x,0), \tag{3.7c}$$

$$u^*(x,b;\epsilon) = g_4(x) - u_0(x,b) - \sum_{k=1}^{n} \epsilon^k u_k(x,b). \tag{3.7d}$$

Now the outer approximation $u(x, y; \epsilon)$ satisfies the boundary conditions (1.2a) and introduces discrepancies in the boundary conditions (1.2b,c,d) on the remaining parts of the boundary of Ω. To obtain a "uniform" approximation of the solution u_ϵ in Ω, we eliminate these discrepancies along the boundaries $x = a$, $y = 0$, and $y = b$ by introducing other functions, called boundary layer functions and corner layer functions, along the three boundaries and the two outflow corners $(a, 0)$ and (a, b). These functions will have the property that when acted on by L_ϵ, the result will be of order $O(\epsilon^{n+1})$ uniformly in the closure of Ω. Also the boundary layer functions have the property of being asymptotically equal to zero everywhere in Ω except for a small neighborhood of one of these three boundaries while the corner layer functions have the property of being asymptotically equal to zero everywhere in Ω except for a small neighborhood of one of the outflow corners of Ω. The boundary layer functions along the outflow boundary $x = a$ satisfy ordinary differential equations and define the ordinary boundary layer (OBL). The boundary layer functions along the characteristic boundaries $y = 0$ and $y = b$ are of two types. One type of function satisfies a parabolic differential equation and is the parabolic boundary layer (PBL) function. The other type of boundary layer functions along the characteristic boundaries $y = 0$ and $y = b$ satisfies an elliptic differential equation and is designed to remove the corner singularities of PBL; it is called the elliptic boundary layer (EBL). There are two types of corner layers at each of the outflow corners. One type of the corner layer function satisfies an ordinary differential equation and is employed to remove the discrepancy in the vicinity of the corner due to the PBL. This function is called an ordinary corner layer (OCL) function. The other type of corner layer function satisfies an elliptic differential equation and is used to remove the discrepancy in the vicinity of the corner due to the EBL, OBL and OCL; this function is called an elliptic corner layer (ECL) function. The detailed construction of these functions will be investigated in the subsequent sections. The location of all functions is indicated in Figure 3.1 and the order of construction is shown in Figure 3.2.

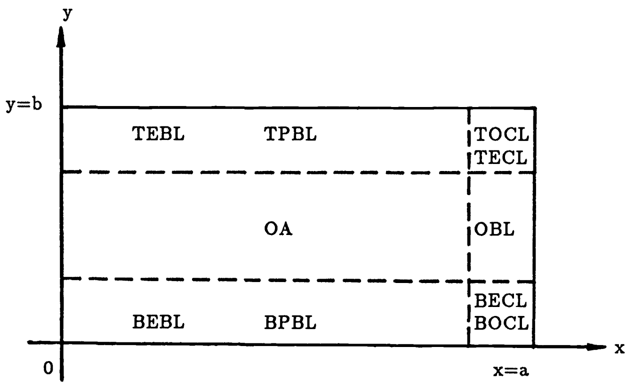


FIG. 3.1. *Location of all functions.*

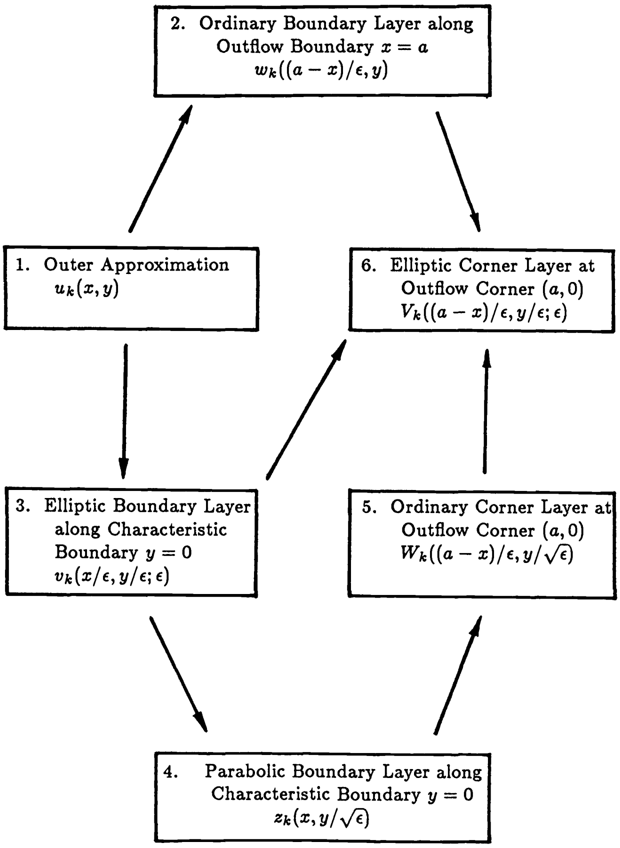

2. Ordinary Boundary Layer along Outflow Boundary $x = a$
$w_k((a-x)/\epsilon, y)$
1. Outer Approximation
$u_k(x,y)$
6. Elliptic Corner Layer at Outflow Corner $(a,0)$
$V_k((a-x)/\epsilon, y/\epsilon; \epsilon)$
3. Elliptic Boundary Layer along Characteristic Boundary $y = 0$
$v_k(x/\epsilon, y/\epsilon; \epsilon)$
5. Ordinary Corner Layer at Outflow Corner $(a,0)$
$W_k((a-x)/\epsilon, y/\sqrt{\epsilon})$
4. Parabolic Boundary Layer along Characteristic Boundary $y = 0$
$z_k(x, y/\sqrt{\epsilon})$

FIG. 3.2. *Order of constructions.*

The uniform asymptotic approximation of the solution $u_\epsilon(x,y)$ in the closure of Ω is expressed as follows:

$$\begin{aligned} u_\epsilon(x,y) = \mathrm{OA} + \mathrm{OBL} + \mathrm{BEBL} + \mathrm{BPBL} + \mathrm{BOCL} + \mathrm{BECL} \\ + \mathrm{TEBL} + \mathrm{TPBL} + \mathrm{TOCL} + \mathrm{TECL} \\ + \mathrm{REMAINDER}, \end{aligned} \tag{3.8}$$

where BEBL and TEBL stand for the EBL along the characteristic boundaries $y = 0$ and $y = b$, respectively, etc.

In order to estimate the term REMAINDER in (3.8), we need the following result, which is a consequence of the maximum principle.

THEOREM 3.3. *If* $\Phi_\epsilon(x,y)$ *is the solution of the boundary value problem*

$$L_\epsilon[\Phi_\epsilon] = h_\epsilon(x,y),$$

valid in Ω *with*

$$\Phi_\epsilon(x,y)\big|_\Gamma = \Psi_\epsilon(x,y)\big|_\Gamma$$

along the boundary Γ *of* Ω, *and if*

$$h_\epsilon(x,y) = O(\epsilon^\mu) \ \ \textit{in } \bar{\Omega}, \quad \mu \geq 0,$$

and

$$\Psi_\epsilon(x,y) = O(\epsilon^\nu) \ \ \textit{along } \Gamma, \quad \nu \geq 0,$$

then at most

$$\Phi_\epsilon(x,y) = O\left(\epsilon^{\min(\mu,\nu)}\right) \ \ \textit{in } \bar{\Omega}.$$

From this theorem, we conclude that if it is possible to have L_ϵ[REMAINDER] being of order $O(\epsilon^{n+1})$ in the closure of Ω and the term REMAINDER being of order $O(\epsilon^{n+1})$ on the boundary of Ω, then it follows that the estimate

$$\text{REMAINDER} = O(\epsilon^{n+1})$$

holds uniformly in the closure of Ω.

3.2. Ordinary boundary layer along the outflow boundary $x = a$. Let the stretched variable X_1 along the outflow boundary $x = a$ be defined by $x = a - \epsilon X_1$. The ordinary boundary layer along the outflow boundary $x = a$ is defined by the series

$$w(X_1, y; \epsilon) = \sum_{k=0}^{n+1} \epsilon^k \, w_k(X_1, y) \tag{3.9}$$

where the functions $w_k(X_1, y)$ are defined iteratively by the ordinary differential equations

$$\frac{\partial^2 w_k}{\partial {X_1}^2} + p \frac{\partial w_k}{\partial X_1} = \pi_k(X_1, y) \tag{3.10}$$

over the unbounded interval $0 < X_1 < \infty$, where y is a parameter $0 \leq y \leq b$, and the functions $\pi_k(X_1, y)$ have the expressions

$$\pi_0(X_1, y) = 0,$$
$$\pi_1(X_1, y) = q w_0,$$

and for $2 \leq k \leq n+1$,

$$\pi_k(X_1, y) = -\frac{\partial^2 w_{k-2}}{\partial y^2} + q w_{k-1}.$$

The boundary conditions imposed on the functions $w_k(X_1, y)$ are such that the discrepancy, due to the introduction of the outer approximation $u(x, y; \epsilon)$ in the boundary condition (1.2b) at the outflow boundary $x = a$, disappears and such that the functions $w_k(X_1, y)$ approach zero as $x \neq a$ and ϵ tends to zero. That is, we impose the boundary conditions

$$\begin{aligned} &w_0(0, y) = g_2(y) - u_0(a, y) \\ &w_k(0, y) = -u_k(a, y) \qquad \text{for } 1 \leq k \leq n, \\ &w_{n+1}(0, y) = 0 \end{aligned} \tag{3.11a}$$

and for $0 \leq k \leq n+1$,

$$w_k(X_1, y) \to 0 \qquad \text{as } X_1 \to \infty. \tag{3.11b}$$

It is easy to see that when $k = 0$ and $k = 1$ we obtain the solutions in the form

$$\begin{aligned} w_0(X_1, y) &= \left[g_2(y) - u_0(a, y)\right] \exp(-pX_1), \\ w_1(X_1, y) &= \left[-u_1(a, y) - \left(g_2(y) - u_0(a, y)\right)p^{-1}qX_1\right] \exp(-pX_1). \end{aligned}$$

In general, for $k \geq 2$, each of the functions $\pi_k(X_1, y)$ is the product of $\exp(-pX_1)$, the function of boundary layer type along the boundary $x = a$, and a polynomial of degree $k-1$ in X_1 with the coefficients depending on y. Hence the solutions w_k can be expressed as

$$w_k(X_1, y) = \ell_k(X_1, y) \exp(-pX_1),$$

where $\ell_k(X_1, y)$ is a polynomial of degree k in X_1 with the coefficients depending on y.

Remark 3.3. From ODE theory, if p is a positive number, the integral

$$\int_0^\infty f(s) \exp(-ps)\, ds$$

converges and

$$\exp(-px) \int_0^x f(s)\, ds$$

converges to 0 as x tends to ∞, then the solution of the ordinary differential equation

$$u''(x) + pu'(x) = f(x) \exp(-px)$$

defined over the unbounded interval $0 < x < \infty$ under the boundary conditions

$$u(0) = A \qquad \text{and} \qquad u(x) \to 0 \quad \text{as } x \to \infty,$$

has the form

$$\begin{aligned} u(x) = A \exp(-px) &- p^{-1} \exp(-px) \int_0^x f(s)\, ds \\ &- p^{-1} \int_x^\infty f(s) \exp(-ps)\, ds + p^{-1} \exp(-px) \int_0^\infty f(s) \exp(-ps)\, ds. \end{aligned} \tag{3.12}$$

THEOREM 3.4. *There exist two positive constants C and c independent of ϵ such that the inequalities*

$$\left| \frac{\partial^\ell}{\partial y^\ell} w_k(X_1, y) \right| \leq C \exp(-cX_1) \tag{3.13}$$

hold for $0 \leq k \leq n+1$ and $\ell = 0$ and 2.

Proof. The above inequalities are quite clear because the functions w_k, and $\partial^2 w_k/\partial y^2$ are products of $\exp(-pX_1)$ and a polynomial in X_1 with the coefficients depending on y.

Applying the differential operator L_ϵ to the series w gives

$$(3.14)\qquad \begin{aligned} L_\epsilon w &= \sum_{k=0}^{n+1} \epsilon^k L_\epsilon w_k \\ &= \sum_{k=0}^{n+1} \epsilon^k \Big(-\epsilon \frac{\partial^2 w_k}{\partial x^2} - \epsilon \frac{\partial^2 w_k}{\partial y^2} + p \frac{\partial w_k}{\partial x} + q w_k\Big). \end{aligned}$$

The equation (3.10) may be written as

$$(3.15)\qquad -\epsilon \frac{\partial^2 w_k}{\partial x^2} + p \frac{\partial w_k}{\partial x} = -\epsilon^{-1} \pi_k\Big(\frac{a-x}{\epsilon}, y\Big).$$

Substitution of (3.15) into (3.14) yields

$$(3.16)\qquad \begin{aligned} L_\epsilon w &= -\sum_{k=0}^{n+1} \epsilon^{k+1} \frac{\partial^2 w_k}{\partial y^2} + \sum_{k=0}^{n+1} \epsilon^k q w_k + \sum_{k=0}^{n+1} \epsilon^{k-1} \pi_k\Big(\frac{a-x}{\epsilon}, y\Big) \\ &= \epsilon^{n+1}\Big(-\epsilon \frac{\partial^2 w_{n+1}}{\partial y^2} - \frac{\partial^2 w_n}{\partial y^2} + q w_{n+1}\Big). \end{aligned}$$

It follows from (3.13) that the estimate

$$L_\epsilon w = O(\epsilon^{n+1})$$

is valid uniformly in the closure of Ω. Furthermore, let us examine this series on the boundary of Ω.

i) At $x = a$, the series

$$(3.17a)\qquad w = \sum_{k=0}^{n+1} \epsilon^k w_k(0, y) = g_2(y) - \sum_{k=0}^{n} \epsilon^k u_k(a, y)$$

is equal to $u^*(a, y; \epsilon)$.

ii) At $x = 0$, the series

$$(3.17b)\qquad w = \sum_{k=0}^{n+1} \epsilon^k w_k\Big(\frac{a}{\epsilon}, y\Big)$$

is asymptotically exponentially small with respect to ϵ in the interval $0 \le y \le b$.

iii) At $y = 0$ and $y = b$, the series

$$(3.17c)\qquad w = \sum_{k=0}^{n+1} \epsilon^k w_k\Big(\frac{a-x}{\epsilon}, y\Big)$$

is asymptotically exponentially small with respect to ϵ for each x in the interval $0 \le x < a$, but not in the closed interval $0 \le x \le a$. To circumvent this difficulty, we define two series, called the elliptic corner layers, at the outflow corners $(a, 0)$ and (a, b). The construction of these elliptic corner layers is given in Section 3.6.

3.3. Elliptic boundary layer along the characteristic boundary $y = 0$. Usually the given problem (1.1), (1.2a,b,c,d) may not have the necessary compatibility conditions for the solution $u_\epsilon(x, y)$ at the inflow corner $(0, 0)$ to insure the smoothness of the parabolic boundary layer function $z_0(x, Y)$, which satisfies the parabolic differential equation (3.26) in a semi-strip $0 < x < a$ and $0 < Y < \infty$. The second partial derivative of z_0 with respect to the time-like variable x is singular near the origin, and consequently the right-hand side of the differential equation for z_1 is singular in the vicinity of the origin. To remedy the presence of this "corner singularity," we introduce some functions $v_k(X, Y_1)$ with the stretched variables $X = x/\epsilon$ and $Y_1 = y/\epsilon$, which are defined by elliptic differential equations over the quarter plane $0 < X < \infty$ and $0 < Y_1 < \infty$ with the values zero as the boundary conditions along $X = 0$ and a suitable boundary condition along $Y_1 = 0$ such that the desired compatibility conditions for parabolic differential equations can be obtained to guarantee the boundedness of the second partial derivative of all functions $z_k(x, Y)$ with respect to x in the semi-strip domain. This enables us to carry out an iteration process related to the parabolic boundary layer.

The elliptic boundary layer along the characteristic boundary $y = 0$ is defined by the series

$$v(X, Y_1; \epsilon) = \sum_{k=0}^{n+1} \epsilon^k \, v_k(X, Y_1; \epsilon), \tag{3.18}$$

where the functions $v_k(X, Y_1; \epsilon)$ are defined iteratively by the elliptic differential equations

$$-\Big(\frac{\partial^2 v_k}{\partial X^2} + \frac{\partial^2 v_k}{\partial {Y_1}^2}\Big) + p\,\frac{\partial v_k}{\partial X} + \epsilon q v_k = 0 \tag{3.19}$$

over the quarter plane $0 < X < \infty$ and $0 < Y_1 < \infty$. We impose the boundary conditions for v_k in the following way:

$$v_k(0, Y_1; \epsilon) = 0, \quad 0 \le k \le n+1, \tag{3.20a}$$

and

$$v_k(X, 0; \epsilon) = \omega_k(X; \epsilon), \quad 0 \le k \le n+1, \tag{3.20b}$$

where the functions $\omega_k(X; \epsilon)$ have the expressions

$$\omega_0(X; \epsilon) = \sum_{i=1}^{N} \frac{X^i}{i!}\Big(g_3^{(i)}(0) - \frac{\partial^i}{\partial x^i}\, u_0(0, 0)\Big)\epsilon^i,$$

and for $1 \le k \le n$,

$$\omega_k(X;\epsilon) = -\sum_{i=1}^{N-2k} \frac{X^i}{i!}\frac{\partial^i}{\partial x^i} u_k(0,0)\,\epsilon^i,$$

and

$$\omega_{n+1}(X;\epsilon) = 0.$$

Note that $\omega_k(0;\epsilon) = 0$. In addition to the Dirichlet boundary conditions (3.20a,b), we impose the following conditions at infinity:

$$v_k(X,Y_1;\epsilon) \to 0 \qquad \text{as } X^2 + Y_1^2 \to \infty \text{ and } Y_1 > 0,\ 0 \le k \le n+1. \tag{3.20c}$$

We remark that an equivalent condition is

$$v_k(X,Y_1;\epsilon) \text{ does not grow exponentially as } X^2 + Y_1^2 \to \infty.$$

The elliptic equation (3.19) with the conditions (3.20a,b,c) has a unique solution, and the maximum principle is valid for this problem [12], [39]. The parameter ϵ appears in this problem as a regular perturbation parameter. Therefore, $v_k(X,Y_1;\epsilon)$ could itself be written as a finite series in ϵ plus a remainder term that is $O(\epsilon^{n+1})$. The particular form of the function $v_k(X,Y_1;\epsilon)$ was chosen to make subsequent computations more tractable.

THEOREM 3.5. *The solutions $v_k(X,Y_1;\epsilon)$ have the integral representations*

$$v_k(X,Y_1;\epsilon) = \frac{\tau Y_1}{\pi}\int_0^\infty \left[r_5^{-1}K_1(\tau r_5) - r_6^{-1}K_1(\tau r_6)\right]\omega_k(s;\epsilon)\exp\left[-\frac{p(s-X)}{2}\right]ds,$$

where

$$\tau = \left(\frac{p^2}{4} + \epsilon q\right)^{1/2},$$
$$r_5 = \left[(X-s)^2 + Y_1^2\right]^{1/2},$$
$$r_6 = \left[(X+s)^2 + Y_1^2\right]^{1/2},$$

and K_1 is the modified Bessel function of the second kind of the first order.

Proof. The transformation

$$v_k(X,Y_1;\epsilon) = v_k^*(X,Y_1;\epsilon)\exp\left(\frac{pX}{2}\right)$$

yields the differential equations for v_k^*

$$-\left(\frac{\partial^2 v_k^*}{\partial X^2} + \frac{\partial^2 v_k^*}{\partial {Y_1}^2}\right) + \left(\frac{p^2}{4} + \epsilon q\right)v_k^* = 0 \tag{3.21}$$

over the quarter plane $0 < X < \infty$ and $0 < Y_1 < \infty$ under the boundary conditions

$$v_k^*(X,0;\epsilon) = \omega_k(X;\epsilon)\exp\Big(-\frac{pX}{2}\Big),$$
$$v_k^*(0,Y_1;\epsilon) = 0,$$

and

$$v_k^*(X,Y_1;\epsilon) \to 0 \quad \text{as } X^2+Y_1^2 \to \infty.$$

Since the fundamental solution for this differential operator (3.21) is

$$\frac{1}{2\pi}K_0\Big(\tau\sqrt{X^2+Y_1^2}\Big),$$

the Green's function for this differential operator over the quarter plane is given by

$$G(X,Y_1;s,t) = \frac{1}{2\pi}\big[K_0(\tau r_1) - K_0(\tau r_2) + K_0(\tau r_3) - K_0(\tau r_4)\big]$$

where K_0 is the modified Bessel function of the second kind of the zeroth order, and

$$r_1 = \big[(X-s)^2+(Y_1-t)^2\big]^{1/2}, \qquad r_2 = \big[(X+s)^2+(Y_1-t)^2\big]^{1/2},$$
$$r_3 = \big[(X+s)^2+(Y_1+t)^2\big]^{1/2}, \qquad r_4 = \big[(X-s)^2+(Y_1+t)^2\big]^{1/2}.$$

It follows that the functions v_k^* have the expressions

$$v_k^*(X,Y_1;\epsilon) = \int_0^\infty \omega_k(s;\epsilon)\exp\Big(-\frac{ps}{2}\Big)\frac{\partial G}{\partial t}(X,Y_1;s,0)\,ds.$$

A computation gives, by using $K_0'(x) = -K_1(x)$,

$$\frac{\partial G}{\partial t}(X,Y_1;s,0) = \frac{\tau Y_1}{\pi}\big[r_5^{-1}K_1(\tau r_5) - r_6^{-1}K_1(\tau r_6)\big].$$

Therefore we have the desired integral representations for $v_k^*(X,Y_1;\epsilon)$. This completes the proof.

The value of the positive integer N should be at least $2n+1$ in order to guarantee the smoothness of the parabolic boundary layer function $(\partial^2/\partial x^2)z_n(x,Y)$ and the ordinary corner layer function $(\partial^2/\partial Y^2)W_{n+1}(X_1,Y)$, which will be discussed shortly.

In order to estimate the elliptic boundary layer, the following exponential estimates for $v_k(X, Y_1; \epsilon)$ are needed.

THEOREM 3.6. *There exist two positive constants C and c independent of ϵ such that the inequalities*

$$|v_k(X, Y_1; \epsilon)| \leq C \exp\left[-c\left(\sqrt{X^2 + Y_1^2} - X\right)\right] \tag{3.22}$$

hold for $0 \leq k \leq n + 1$.

Proof. The condition at infinity (3.20c) enables us to have the maximum principle for the boundary value problem over the unbounded domain. By the linearity, it suffices to prove that the solutions $v_k(X, Y_1; \epsilon)$ for the differential equation (3.19) over the quarter plane $0 < X < \infty$ and $0 < Y_1 < \infty$ under the boundary conditions (3.20a) and

$$v_k(X, 0; \epsilon) = X^i \epsilon^i,$$

and (3.20c) satisfy the estimate (3.22). Let the barrier function $U(X, Y_1; \epsilon)$ be defined by

$$U(X, Y_1; \epsilon) = \epsilon^i \sum_{j=0}^{i} c_j \, (X^2 + Y_1^2)^{j/2} \exp\left[-\tfrac{p}{2}\left(\sqrt{X^2 + Y_1^2} - X\right)\right],$$

where the values of positive constants c_j will be determined in the course of this proof. Note that the function U satisfies the restricted growth condition at infinity (3.20c). It is clear that

$$U(X, 0; \epsilon) = \epsilon^i \sum_{j=0}^{i} c_j \, X^j \geq v_k(X, 0; \epsilon)$$

for all k if $c_j = 1$ and $c_j \geq 0$ for $0 \leq j \leq i - 1$, and

$$U(0, Y_1; \epsilon) = \epsilon^i \sum_{j=0}^{i} c_j \, Y_1^j \exp\left(-\tfrac{p}{2} Y_1\right) > v_k(0, Y_1; \epsilon)$$

for all k. A computation yields

$$
\begin{aligned}
&-\left(\frac{\partial^2 U}{\partial X^2}+\frac{\partial^2 U}{\partial Y_1{}^2}\right)+p\frac{\partial U}{\partial X}+\epsilon q U\\
&=\epsilon^i\exp\left[-\frac{p}{2}\left(\sqrt{X^2+Y_1^2}-X\right)\right]\left(\sum_{j=0}^{i}c_j p(j+\tfrac12)(X^2+Y_1^2)^{(j-1)/2}\right.\\
&\qquad\left.-\sum_{j=0}^{i}c_j j^2(X^2+Y_1^2)^{(j-2)/2}+\sum_{j=0}^{i}\epsilon q c_j(X^2+Y_1^2)^{j/2}\right)\\
&=\epsilon^i\exp\left[-\frac{p}{2}\left(\sqrt{X^2+Y_1^2}-X\right)\right]\\
&\quad\cdot\left(\epsilon q c_i(X^2+Y_1^2)^{i/2}+\left[\epsilon q c_{i-1}+c_i p(i+\tfrac12)\right](X^2+Y_1^2)^{(i-1)/2}\right.\\
&\qquad+\sum_{j=1}^{i-1}\left[\epsilon q c_{j-1}+c_j p(j+\tfrac12)-c_{j+1}(j+1)^2\right]\\
&\qquad\quad\left.\cdot(X^2+Y_1^2)^{(j-1)/2}\left(c_0\frac{p}{2}-c_1\right)(X^2+Y_1^2)^{-1/2}\right)\\
&\ge\epsilon^i\exp\left[-\frac{p}{2}\left(\sqrt{X^2+Y_1^2}-X\right)\right]\left(c_i p(i+\tfrac12)(X^2+Y_1^2)^{(i-1)/2}\right.\\
&\qquad\left.+\sum_{j=0}^{i-1}\left[c_j p(j+\tfrac12)-c_{j+1}(j+1)^2\right](X^2+Y_1^2)^{(j-1)/2}\right),
\end{aligned}
$$

which is greater than or equal to zero if the numbers c_j are chosen so that the inequalities

$$c_j p(j+\tfrac12)\ge c_{j+1}(j+1)^2$$

hold for $j=i-1, i-2, \ldots, 2, 1, 0$, then it follows from the maximum principle that the inequalities

$$\left|v_k(X,Y_1;\epsilon)\right|\le\epsilon^i\sum_{j=0}^{i}c_j(X^2+Y_1^2)^{j/2}\exp\left[-\frac{p}{2}\left(\sqrt{X^2+Y_1^2}-X\right)\right]$$

hold for $0\le x<\infty$ and $0\le Y_1<\infty$. By the definition of the stretched variables X and Y_1, we obtain

$$\left|v_k\left(\frac{x}{\epsilon},\frac{y}{\epsilon};\epsilon\right)\right|\le\sum_{j=0}^{i}c_j(a^2+b^2)^{j/2}\exp\left[-\frac{p}{2\epsilon}\left(\sqrt{x^2+y^2}-x\right)\right].$$

Hence over the closure of Ω, we are led to the estimate

$$\left|v_k\left(\frac{x}{\epsilon},\frac{y}{\epsilon};\epsilon\right)\right|\le C\exp\left[-\frac{p}{2\epsilon}\left(\sqrt{x^2+y^2}-x\right)\right],$$

where C is some constant independent of ϵ. This completes the proof.

Now applying the differential operator L_ϵ to the series v yields

$$\begin{aligned} L_\epsilon v &= \sum_{k=0}^{n+1} \epsilon^k L_\epsilon v_k \\ &= \sum_{k=0}^{n+1} \epsilon^k \Big[-\epsilon \Big(\frac{\partial^2 v_k}{\partial x^2} + \frac{\partial^2 v_k}{\partial y^2} \Big) + p \frac{\partial v_k}{\partial x} + q v_k \Big]. \end{aligned}$$

Equation (3.19) can be written as

$$-\epsilon \Big(\frac{\partial^2 v_k}{\partial x^2} + \frac{\partial^2 v_k}{\partial y^2} \Big) + p \frac{\partial v_k}{\partial x} + q v_k = 0.$$

Hence it follows that

$$L_\epsilon v = 0 \tag{3.23}$$

is valid uniformly in the closure of Ω. Moreover, let us examine this series on the boundary of Ω.

i) At $y = 0$, the series becomes

$$v = \sum_{i=1}^{N} \frac{x^i}{i!} \Big[g_3^{(i)}(0) - \frac{\partial^i}{\partial x^i} u_0(0,0) \Big] - \sum_{k=1}^{n} \epsilon^k \Bigg[\sum_{i=1}^{N-2k} \frac{x^i}{i!} \frac{\partial^i}{\partial x^i} u_k(0,0) \Bigg]. \tag{3.24a}$$

ii) At $x = 0$,

$$v = 0. \tag{3.24b}$$

iii) At $y = b$, the series

$$v = \sum_{k=0}^{n} \epsilon^k v_k \Big(\frac{x}{\epsilon}, \frac{b}{\epsilon}; \epsilon \Big) \tag{3.24c}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \le x \le a$.

iv) At $x = a$, the series

$$v = \sum_{k=0}^{n} \epsilon^k v_k \Big(\frac{a}{\epsilon}, \frac{y}{\epsilon}; \epsilon \Big) \tag{3.24d}$$

is asymptotically exponentially small with respect to ϵ in the interval $0 < y \le b$, but not in the closed interval $0 \le y \le b$. To overcome this difficulty, we will define a series, called an elliptic corner layer (ECL), at the outflow corner $(a, 0)$. The ECL is constructed in Section 3.6.

3.4. Parabolic boundary layer along the characteristic boundary $y = 0$. The stretched variable Y along the characteristic boundary $y = 0$ is defined by $y = \sqrt{\epsilon}\,Y$. The parabolic boundary layer along the characteristic boundary $y = 0$ is defined by the series

$$z(x, Y; \epsilon) = \sum_{k=0}^{n} \epsilon^k z_k(x, Y), \tag{3.25}$$

where the functions $z_k(x, Y)$ are defined iteratively by the parabolic differential equations

$$-\frac{\partial^2 z_k}{\partial Y^2} + p\frac{\partial z_k}{\partial x} + q z_k = \mu_k(x, Y) \tag{3.26}$$

over the semi-strip $0 < x < a$ and $0 < Y < \infty$. The functions μ_k are given by

$$\mu_0(x, Y) = 0$$

and for $1 \leq k \leq n$,

$$\mu_k(x, Y) = \frac{\partial^2 z_{k-1}}{\partial x^2}.$$

We impose boundary conditions on the functions z_k to eliminate the discrepancy along the boundary $y = 0$ introduced by both the outer approximation $u(x, y; \epsilon)$ and the elliptic boundary layer $v(X, Y_1; \epsilon)$, and to arrange that the functions $z_k(x, Y)$ approach zero for $Y \neq 0$ and as $\epsilon \to 0$. For this, we impose the following initial-boundary conditions

$$z_k(0, Y) = 0, \tag{3.27a}$$

$$z_k(x, 0) = \gamma_k(x), \tag{3.27b}$$

and

$$z_k(x, Y) \to 0 \quad \text{as } Y \to \infty. \tag{3.27c}$$

The functions $\gamma_k(x)$ are defined by

$$\gamma_0(x) = g_3(x) - u_0(x, 0) - \sum_{i=1}^{N} \frac{x^i}{i!}\left[g_3^{(i)}(0) - \frac{\partial^i}{\partial x^i} u_0(0, 0)\right],$$

and for $1 \leq k \leq n$,

$$\gamma_k(x) = -\left[u_k(x, 0) - \sum_{i=1}^{N-2k} \frac{x^i}{i!}\frac{\partial^i}{\partial x^i} u_k(0, 0)\right].$$

Note that $\gamma_k(0) = 0$ for each k.

THEOREM 3.7. *The solutions $z_k(x,Y)$ for the initial-boundary value problems* (3.26), (3.27a,b,c) *have the integral representations*

$$z_0(x,Y) = \sqrt{\frac{2}{\pi}} \int_{Y/\sqrt{2x/p}}^{\infty} \exp\Big(-\frac{t^2}{2}\Big)\,\gamma_0\Big(x-\frac{pY^2}{2t^2}\Big) \exp\Big(-\frac{qY^2}{2t^2}\Big)\,dt, \tag{3.28}$$

and for $1 \le k \le n$,

$$\begin{aligned} z_k(x,Y) = {} & \sqrt{\frac{2}{\pi}} \int_{Y/\sqrt{2x/p}}^{\infty} \exp\Big(-\frac{t^2}{2}\Big)\,\gamma_k\Big(x-\frac{pY^2}{2t^2}\Big) \exp\Big(-\frac{qY^2}{2t^2}\Big)\,dt \\ & + \frac{1}{2\sqrt{\pi}} \int_0^{x/p}\int_0^{\infty} \frac{1}{\sqrt{s}}\Big\{\exp\Big[-\frac{(Y-t)^2}{4s}\Big] - \exp\Big[-\frac{(Y+t)^2}{4s}\Big]\Big\} \\ & \cdot \frac{\partial^2}{\partial x^2} z_{k-1}(x-ps,t)\exp(-qs)\,dt\,ds. \end{aligned} \tag{3.29}$$

Proof. The transformation

$$z_k(x,Y) = z_k^*\Big(\frac{x}{p},Y\Big)\exp\Big(-\frac{qx}{p}\Big)$$

yields the heat equation for $z_k^*(x,Y)$

$$-\frac{\partial^2 z_k^*}{\partial Y^2} + \frac{\partial z_k^*}{\partial x} = \mu_k(px,Y)\exp(qx) \tag{3.30}$$

over the semi-strip $0 < x < a/p$ and $0 < Y < \infty$ under the initial-boundary conditions

$$z_k^*(0,Y) = 0, \quad z_k^*(x,0) = \gamma_k(px)\exp(qx),$$

and

$$z_k^*(x,Y) \to 0 \quad \text{as } Y \to \infty.$$

The fundamental solution for the differential operator of (3.30) is

$$K(x,Y) = \frac{1}{\sqrt{4\pi x}}\exp\Big(-\frac{Y^2}{4x}\Big),$$

and hence Green's function for the differential operator over the semi-strip $0 < x < a/p$ and $0 < Y < \infty$ is given by

$$G(x,Y;s,t) = K(x-s,Y-t) - K(x-s,Y+t).$$

Therefore the solutions $z_k^*(x,Y)$ can be expressed as [5]

$$\begin{aligned} z_k^*(x,Y) = {} & \int_0^x \frac{\partial G}{\partial t}(x,Y;s,0)\,\gamma_k(ps)\exp(qs)\,ds \\ & + \int_0^x\int_0^{\infty} G(x,Y;s,t)\,\mu_k(ps,t)\exp(qs)\,dt\,ds. \end{aligned}$$

A computation shows that

$$\begin{aligned}\frac{\partial G}{\partial t}(x,Y;s,0) &= -2\frac{\partial}{\partial Y}K(x-s,Y)\\ &= \frac{Y}{2\sqrt{\pi}\,(x-s)^{3/2}}\exp\Big[-\frac{Y^2}{4(x-s)}\Big],\end{aligned}$$

and then we obtain

$$\begin{aligned}z_k^*(x,Y) &= \frac{Y}{2\sqrt{\pi}}\int_0^x \frac{1}{(x-s)^{3/2}}\exp\Big[-\frac{Y^2}{4(x-s)}\Big]\gamma_k(ps)\exp(qs)\,ds\\ &\quad+\int_0^x\int_0^\infty G(x,Y;s,t)\,\mu_k(ps,t)\exp(qs)\,dt\,ds.\end{aligned}$$

Making the changes of integrators as follows: let $Y/\sqrt{2(x-t)}=t$ in the first integral, and let $x-s=s'$ and then replace s' by s in the second integral of z_k^*, we get

$$\begin{aligned}z_k^*(x,Y) &= \sqrt{\frac{2}{\pi}}\int_{Y/\sqrt{2x}}^\infty \exp\Big(-\frac{t^2}{2}\Big)\gamma_k\Big(px-\frac{pY^2}{2t^2}\Big)\exp\Big(qx-\frac{qY^2}{2t^2}\Big)\,dt\\ &\quad+\frac{1}{2\sqrt{\pi}}\int_0^x\int_0^\infty \frac{1}{\sqrt{s}}\Big\{\exp\Big[-\frac{(Y-t)^2}{4s}\Big]-\exp\Big[-\frac{(Y+t)^2}{4s}\Big]\Big\}\\ &\qquad\cdot\mu_k(px-ps,t)\exp[q(x-s)]\,dt\,ds,\end{aligned}$$

which gives the desired expressions for $z_k(x,Y)$. This completes the proof.

THEOREM 3.8. *There exist two positive constants C and c independent of ϵ such that the inequalities*

$$\Big|\frac{\partial^i}{\partial x^i}z_k(x,Y)\Big| \le C\exp(-cY), \tag{3.31}$$

and

$$\Big|\frac{\partial^{2i}}{\partial Y^{2i}}z_k(x,Y)\Big| \le C\exp(-cY), \tag{3.32}$$

hold for $0\le i\le 2n+2-2k$. The inequalities (3.32) will be used in Section 3.5 to obtain estimates for the ordinary corner layer functions $W_k(X_1,Y)$.

Proof. First of all, let us prove the inequality (3.31) when $k=0$. Thanks to the construction of $\gamma_0(x)$, it is easy to see that

$$\gamma_0^{(i)}(0)=0,$$

for $i=0,1,2,\ldots,2n+1$ (the value of N should be at least $2n+1$), and hence we have

$$\frac{\partial^i}{\partial x^i}z_0(x,Y)=\sqrt{\frac{2}{\pi}}\int_{Y/\sqrt{2x/p}}^\infty \exp\Big(-\frac{t^2}{2}\Big)\gamma_0^{(i)}\Big(x-\frac{pY^2}{2t^2}\Big)\,dt,$$

for $0 \le i \le 2n+2$. Since

$$\left|\gamma_0^{(i)}\Big(x - \frac{pY^2}{2t^2}\Big)\right| \le C$$

and for arbitrary constant $c > 0$,

$$\exp\Big(-\frac{t^2}{2}\Big) \le \exp\Big(\frac{c^2}{2} - ct\Big) = \exp\Big(\frac{c^2}{2}\Big)\cdot\exp(-ct)$$
$$= C\exp(-ct),$$

we have

$$\left|\frac{\partial^i}{\partial x^i} z_0(x,Y)\right| \le C\int_{Y/\sqrt{2x/p}}^{\infty}\exp(-ct)\,dt = C\exp\Big(-c\frac{Y}{\sqrt{2x/p}}\Big)$$
$$\le C\exp\Big(-c\frac{Y}{\sqrt{2a/p}}\Big) = C\exp(-cY),$$

for $0 \le Y < \infty$. It follows from (3.26) that $\partial^{2i}z_0/\partial Y^{2i}$ is a linear combination of $\partial^j z_0/\partial x^j$ for $0 \le j \le i$. Consequently, the inequality (3.32) is obtained for $k = 0$. Next consider the case $k = 1$, we have from (3.29)

$$\frac{\partial^i}{\partial x^i} z_1(x,Y) = \sqrt{\frac{2}{\pi}}\int_{Y/\sqrt{2x/p}}^{\infty}\exp\Big(-\frac{t^2}{2}\Big)\gamma_1^{(i)}\Big(x - \frac{pY^2}{2t^2}\Big)\exp\Big(-\frac{qY^2}{2t^2}\Big)\,dt$$
$$+ \frac{1}{2\sqrt{\pi}}\int_0^{x/p}\int_0^{\infty}\frac{1}{\sqrt{s}}\Big[\exp\Big(-\frac{(Y-t)^2}{4s}\Big) - \exp\Big(-\frac{(Y+t)^2}{4s}\Big)\Big]$$
$$\cdot\frac{\partial^{2+i}}{\partial x^{2+i}} z_0(x-ps,t)\,dt\,ds$$

for $0 \le i \le 2n$, because of

$$\gamma_1^{(i)}(0) = 0$$

for $0 \le i \le 2n-1$ and

$$\frac{\partial^j}{\partial x^j} z_0(0,y) = 0$$

for $2 \le j \le 2n{+}1$. (Note that $(\partial^j/\partial x^j)z_0(0,Y)$ is a linear combination of $(\partial^{2i}/\partial Y^{2i})z_0$ $(0,Y)$ for $0 \le i \le j$.) As in the case $k = 0$, the first integral of $(\partial^i/\partial x^i)z_1(x,Y)$ has the desired estimate, and therefore we consider the second integral only. Now the estimates

$$\left|\frac{\partial^{2+i}}{\partial x^{2+i}} z_0(x-ps,t)\right| \le C\exp(-ct)$$

and

$$\left|\exp\Big(-\frac{(Y-t)^2}{4s}\Big) - \exp\Big(-\frac{(Y+t)^2}{4s}\Big)\right|$$
$$\le \exp\Big(-\frac{(Y-t)^2}{4s}\Big) \le C\exp\Big(-c\frac{|Y-t|}{2\sqrt{s}}\Big)$$
$$\le C\exp\Big(-c\frac{|Y-t|}{2\sqrt{x/p}}\Big) \le C\exp(-c|Y-t|)$$

imply

$$\begin{aligned}&\left|\text{second integral of } \frac{\partial^i}{\partial x^i} z_1(x,Y)\right|\\ &\quad\le C\int_0^{x/p}\int_0^\infty \frac{1}{\sqrt{s}}\left[\exp(-c|Y-t|)\exp(-ct)\right]dt\,ds\\ &\quad= C\int_0^{x/p}\frac{1}{\sqrt{s}}\,ds\int_0^\infty \exp\left[-c(|Y-t|+t)\right]dt\\ &\quad= C\exp(-cY).\end{aligned}$$

This completes the case $k = 1$ for inequality (3.31). From (3.26), one can see that $\partial^{2i}z_1/\partial Y^{2i}$ depends linearly on $\partial^j z_1/\partial x^j$ and $\partial^{\ell+1}z_0/\partial x^{\ell+1}$ for $0 \le i \le 2n$, $0 \le j \le i$ and $1 \le \ell \le i$.

Continuing in this manner, one can show the inequalities (3.31) and (3.32) for $k = 2, 3, \ldots, n$. This completes the proof.

Now applying the differential operator L_ϵ to the series z yields

$$\begin{aligned}L_\epsilon z &= \sum_{k=0}^{n} \epsilon^k L_\epsilon z_k\\ &= \sum_{k=0}^{n}\epsilon^k\left(-\epsilon\frac{\partial^2 z_k}{\partial x^2} - \epsilon\frac{\partial^2 z_k}{\partial y^2} + p\frac{\partial z_k}{\partial x} + qz_k\right).\end{aligned}\tag{3.33}$$

Equation (3.26) can be written as

$$-\epsilon\frac{\partial^2 z_k}{\partial y^2} + p\frac{\partial z_k}{\partial x} + qz_k = \mu_k\left(x, \frac{y}{\sqrt{\epsilon}}\right).\tag{3.34}$$

Substitution of (3.34) into (3.33) gives

$$\begin{aligned}L_\epsilon z &= -\sum_{k=0}^{n}\epsilon^{k+1}\frac{\partial^2 z_k}{\partial x^2} + \sum_{k=0}^{n}\epsilon^k\mu_k\left(x,\frac{y}{\sqrt{\epsilon}}\right)\\ &= -\epsilon\frac{\partial^2 z_n}{\partial x^2}.\end{aligned}\tag{3.35}$$

It follows from (3.31) that the estimate

$$L_\epsilon z = O\left(\epsilon^{n+1}\right)$$

is valid uniformly in the closure of Ω. Moreover, let us examine the parabolic boundary layer z along the boundary of Ω.

i) At $x = 0$,

$$z = 0.\tag{3.36a}$$

ii) At $y = 0$,

$$z = u^*(x,0;\epsilon) - v\left(\frac{x}{\epsilon},0;\epsilon\right).\tag{3.36b}$$

iii) At $y = b$, the series

$$z = \sum_{k=0}^{n} \epsilon^k z_k\Big(x, \frac{b}{\sqrt{\epsilon}}\Big) \tag{3.36c}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \leq x \leq a$.

iv) At $x = a$, the series

$$z = \sum_{k=0}^{n} \epsilon^k z_k\Big(a, \frac{y}{\sqrt{\epsilon}}\Big) \tag{3.36d}$$

is asymptotically exponentially small with respect to ϵ in the interval $0 < y \leq b$, but not $0 \leq y \leq b$. In the next section we define a series, called ordinary corner layer, at the outflow corner $(a, 0)$ to overcome this difficulty.

3.5. Ordinary corner layer at the outflow corner $(a, 0)$. As defined above, the stretched variables X_1 and Y are expressed by $x = a - \epsilon X_1$ and $y = \sqrt{\epsilon}Y$. The ordinary corner layer at the outflow corner $(a, 0)$ is defined by the series

$$W(X_1, Y; \epsilon) = \sum_{k=0}^{n+1} \epsilon^k W_k(X_1, Y), \tag{3.37}$$

where the functions $W_k(X_1, Y)$ are defined iteratively by the ordinary differential equations

$$\frac{\partial^2 W_k}{\partial {X_1}^2} + p\frac{\partial W_k}{\partial X_1} = \tau_k(X_1, Y) \tag{3.38}$$

over the unbounded interval $0 < X_1 < \infty$. In these equations, Y may be regarded as a parameter $0 \leq Y < \infty$ and the functions τ_k are defined as

$$\tau_0(X_1, Y) = 0,$$

and for $1 \leq k \leq n + 1$,

$$\tau_k(X_1, Y) = -\frac{\partial^2 W_{k-1}}{\partial Y^2} + qW_{k-1}.$$

The boundary conditions imposed on the functions W_k remove the discrepancy along the outflow boundary $x = a$ introduced by the parabolic boundary layer z and are such that the functions $W_k(X_1, Y)$ become boundary layer functions with respect to X_1. That is, we define

$$\begin{aligned} &W_k(0, Y) = -z_k(a, Y) \quad \text{for } 0 \leq k \leq n, \\ &W_{n+1}(0, Y) = 0, \end{aligned} \tag{3.39a}$$

and for $0 \le k \le n+1$,

$$W_k(X_1, Y) \to 0 \quad \text{as } X_1 \to \infty. \tag{3.39b}$$

It is easy to see that the function $W_0(X_1, Y)$ has the following representation:

$$W_0(X_1, Y) = -z_0(a, Y) \exp(-pX_1).$$

In general, for $k \ge 1$, the functions $\tau_k(X_1, Y)$ are the products of a boundary layer function $\exp(-pX_1)$ and a polynomial of degree $k-1$ in X_1 with the coefficients depending on the parabolic boundary layer functions $z_i(a, Y)$, $0 \le i \le k-1$, and their even-order partial derivatives with respect to Y. Therefore, it follows from (3.12) that the solutions $W_k(X_1, Y)$ can be expressed as

$$W_k(X_1, Y) = m_k(X_1, Y) \exp(-pX_1),$$

where the function m_k is a polynomial of degree k in X_1 with the coefficients depending on the functions $z_i(a, Y)$, $0 \le i \le k$ ($0 \le i \le n$ when $k = n+1$), and their even-order partial derivatives with respect to Y, $(\partial^{2j}/\partial Y^{2j}) z_i(a, Y)$, for $0 \le i \le k-1$, and $1 \le j \le k-i$.

Note that the functions $W_k(X_1, Y)$ and their second partial derivatives with respect to Y (the latter is required for the estimate of $L_\epsilon W$) determine how smooth the parabolic boundary layer functions $z_k(x, Y)$ should be in the domain $0 \le x \le a$ and $0 \le Y < \infty$.

THEOREM 3.9. *There exist two positive constants C and c independent of ϵ such that the inequalities*

$$\left| \frac{\partial^i}{\partial Y^i} W_k(X_1, Y) \right| \le C \exp\left[-c(X_1 + Y)\right] \tag{3.40}$$

hold for $i = 0$ and $i = 2$.

Proof. The functions W_k and $\partial^2 W_k / \partial Y^2$ are products of $\exp(-pX_1)$ and a polynomial in X_1 with the coefficients depending on the functions z_i and $\partial^{2j} z_i / \partial Y^{2j}$. It follows from (3.32) that we have the desired inequalities.

Now applying the differential operator L_ϵ to the ordinary corner layer W yields

$$\begin{aligned} L_\epsilon W &= \sum_{k=0}^{n+1} \epsilon^k L_\epsilon W_k \\ &= \sum_{k=0}^{N+1} \epsilon^k \left(-\epsilon \frac{\partial^2 W_k}{\partial x^2} - \epsilon \frac{\partial^2 W_k}{\partial y^2} + p \frac{\partial W_k}{\partial x} + q W_k \right). \end{aligned} \tag{3.41}$$

Equation (3.38) can be written as

$$-\epsilon \frac{\partial^2 W}{\partial x^2} + p \frac{\partial W_k}{\partial x} = -\epsilon^{-1} \tau_k \left(\frac{a-x}{\epsilon}, \frac{y}{\sqrt{\epsilon}} \right). \tag{3.42}$$

Substitution of (3.42) into (3.41) gives

$$L_\epsilon W = \epsilon^{n+1}\Big(-\frac{\partial^2 W_{n+1}}{\partial Y^2} + qW_{n+1}\Big). \tag{3.43}$$

It follows from (3.40) that the estimate

$$L_\epsilon W = O\big(\epsilon^{n+1}\big)$$

holds uniformly in the closure of Ω. Moreover, let us examine the ordinary corner layer W along the boundary of Ω.

i) At $x = a$,

$$W = -z(a, Y; \epsilon). \tag{3.44a}$$

ii) At $x = 0$, the series

$$W = \sum_{k=0}^{n+1} \epsilon^k W_k\Big(\frac{a}{\epsilon}, \frac{y}{\sqrt{\epsilon}}\Big) \tag{3.44b}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \leq y \leq b$.

iii) At $y = b$, the series

$$W = \sum_{k=0}^{n+1} \epsilon^k W_k\Big(\frac{a-x}{\epsilon}, \frac{b}{\sqrt{\epsilon}}\Big) \tag{3.44c}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \leq x \leq a$.

iv) At $y = 0$, the series

$$W = \sum_{k=0}^{n+1} \epsilon^k W_k\Big(\frac{a-x}{\epsilon}, 0\Big) \tag{3.44d}$$

is asymptotically exponentially small with respect to ϵ in the interval $0 \leq x < a$, but not the closed interval $0 \leq x \leq a$. We shall define a series, called the elliptic corner layer, at the outflow corner $(a, 0)$ to overcome this difficulty.

3.6. Elliptic corner layer at the outflow corner $(a,0)$. As defined above, the stretched variables X_1 and Y_1 are expressed by $x = a - \epsilon X_1$ and $y = \epsilon Y_1$. The elliptic corner layer at the outflow corner $(a, 0)$ is defined by the series

$$V(X_1, Y_1; \epsilon) = \sum_{k=0}^{n+1} \epsilon^k V_k(X_1, Y_1; \epsilon), \tag{3.45}$$

where the functions $V_k(X_1, Y_1; \epsilon)$ are defined by the elliptic differential equations

$$-\Big(\frac{\partial^2 V_k}{\partial {X_1}^2} + \frac{\partial^2 V_k}{\partial {Y_1}^2}\Big) - p\frac{\partial V_k}{\partial X_1} + \epsilon q V_k = 0 \tag{3.46}$$

over the quarter plane $0 < X_1 < \infty$ and $0 < Y_1 < \infty$. The boundary conditions for $V_k(X_1, Y_1; \epsilon)$ are specified so that

i) the discrepancy introduced by both the ordinary boundary layer $w(X_1, y; \epsilon)$ and the ordinary corner layer $W(X_1, Y; \epsilon)$ in the boundary condition along $y = 0$ is eliminated;

ii) the discrepancy introduced by the elliptic boundary layer $v(X, Y_1; \epsilon)$ in the boundary condition along $x = a$ is eliminated; and

iii) the functions $V_k(X_1, Y_1; \epsilon)$ are corner layer functions with respect to X_1 and Y_1.

Thus, we impose the following conditions

$$\text{(3.47a)} \qquad \begin{aligned} V_k(X_1, 0; \epsilon) &= \xi_k(X_1) \qquad\qquad 0 \le k \le n+1, \\ &\equiv -w_k(X_1, 0) - W_k(X_1, 0), \end{aligned}$$

$$\text{(3.47b)} \qquad \begin{aligned} V_k(0, Y_1; \epsilon) &= \varsigma_k(Y_1; \epsilon) \\ &\equiv \begin{cases} -v_k\Big(\dfrac{a}{\epsilon}, Y_1; \epsilon\Big), & 0 \le k \le n, \\ 0, & k = n+1, \end{cases} \end{aligned}$$

and

$$\text{(3.47c)} \qquad V_k(X_1, Y_1; \epsilon) \to 0 \qquad \text{as } X_1^2 + Y_1^2 \to \infty,\ 0 \le k \le n+1.$$

Note that $\xi_k(0) = \varsigma_k(0; \epsilon)$ for $0 \le k \le n+1$. The elliptic problem (3.46), (3.47a,b,c) has a unique solution, and the maximum principle is valid for this problem. Note that the parameter ϵ appears in the equation (3.46) as a regular perturbation problem. The particular form of the functions $V_k(X_1, Y_1; \epsilon)$ was chosen to make the computations more tractable.

THEOREM 3.10. *The solutions* $V_k(X_1, Y_1; \epsilon)$ *of the boundary value problem* (3.46), (3.47a,b,c) *have the integral representations*

$$\begin{aligned} V_k(X_1, Y_1; \epsilon) = {} & \frac{\tau Y_1}{\pi} \int_0^\infty \Big(\frac{K_1(\tau\rho_5)}{\rho_5} - \frac{K_1(\tau\rho_6)}{\rho_6} \Big) \xi_k(s) \exp\Big(\frac{p(s - X_1)}{2}\Big)\, ds \\ & + \frac{\tau X_1}{\pi} \int_0^\infty \Big(\frac{K_1(\tau\rho_7)}{\rho_7} - \frac{K_1(\tau\rho_8)}{\rho_8} \Big) \varsigma_k(t; \epsilon) \exp\Big(-\frac{pX_1}{2}\Big)\, dt, \end{aligned}$$

where

$$\rho_5 = \big[(X_1 - s)^2 + Y_1^2\big]^{1/2}, \qquad \rho_6 = \big[(X_1 + s)^2 + Y_1^2\big]^{1/2},$$
$$\rho_7 = \big[X_1^2 + (Y_1 - t)^2\big]^{1/2}, \qquad \rho_8 = \big[X_1^2 + (Y_1 + t)^2\big]^{1/2},$$

and

$$\tau = \Big[\frac{p^2}{4} + \epsilon q\Big]^{1/2}.$$

Proof. The transformation

$$V_k(X_1, Y_1; \epsilon) = V_k^*(X_1, Y_1; \epsilon) \exp\Big(-\frac{pX_1}{2}\Big)$$

yields the differential equation for V_k^*

$$-\Big(\frac{\partial^2 V_k^*}{\partial X_1{}^2} + \frac{\partial^2 V_k^*}{\partial Y_1{}^2}\Big) + \Big(\frac{p^2}{4} + \epsilon q\Big) V_k^* = 0$$

over the quarter plane $0 < X_1 < \infty$ and $0 < Y_1 < \infty$ under the boundary conditions

$$V_k^*(X_1, 0; \epsilon) = \xi_k(X_1) \exp\Big(\frac{pX_1}{2}\Big),$$

$$V_k^*(0, Y_1; \epsilon) = \varsigma_k(Y_1; \epsilon)$$

and

$$V_k^*(X_1, Y_1; \epsilon) \to 0 \qquad \text{as } X_1^2 + Y_1^2 \to \infty,\ Y_1 > 0.$$

As in the case of the elliptic boundary layer, the solution V_k^* has the expression of the form

$$V_k^*(X_1, Y_1; \epsilon) = \int_0^\infty \xi_k(s) \exp\Big(\frac{ps}{2}\Big) \frac{\partial G}{\partial t}(X_1, Y_1; s, 0)\, ds + \int_0^\infty \varsigma_k(t; \epsilon) \frac{\partial G}{\partial s}(X_1, Y_1; 0, t)\, dt,$$

where the Green's function for this problem is given by

$$G(X_1, Y_1; s, t) = \frac{1}{2\pi}\big[K_0(\tau\rho_1) - K_0(\tau\rho_2) + K_0(\tau\rho_3) - K_0(\tau\rho_4)\big],$$

with

$$\rho_1 = \big[(X_1 - s)^2 + (Y_1 - t)^2\big]^{1/2}, \qquad \rho_2 = \big[(X_1 + s)^2 + (Y_1 - t)^2\big]^{1/2},$$
$$\rho_3 = \big[(X_1 + s)^2 + (Y_1 + t)^2\big]^{1/2}, \qquad \rho_4 = \big[(X_1 - s)^2 + (Y_1 + t)^2\big]^{1/2}.$$

Computations give

$$\frac{\partial G}{\partial t}(X_1, Y_1; s, 0) = \frac{\tau Y_1}{\pi}\Big[\frac{K_1(\tau\rho_5)}{\rho_5} - \frac{K_1(\tau\rho_6)}{\rho_6}\Big],$$

and

$$\frac{\partial G}{\partial s}(X_1, Y_1; 0, t) = \frac{\tau X_1}{\pi}\Big[\frac{K_1(\tau\rho_7)}{\rho_7} - \frac{K_1(\tau\rho_8)}{\rho_8}\Big],$$

and hence we obtain the desired integral representations. This completes the proof.

THEOREM 3.11. *There exist two positive constants C and c independent of ϵ such that the inequalities*

$$|V_k(X_1, Y_1; \epsilon)| \le C \exp\Big[-c\Big(X_1 + \sqrt{\Big(\frac{a}{\epsilon}\Big)^2 + Y_1^2} - \frac{a}{\epsilon}\Big)\Big] \tag{3.48}$$

hold.

Proof. From (3.22), we have

$$|V_k(0, Y_1; \epsilon)| \le C \exp\Big[-\frac{p}{2}\Big(\sqrt{\Big(\frac{a}{\epsilon}\Big)^2 + Y_1^2} - \frac{a}{\epsilon}\Big)\Big].$$

The estimates (3.13) and (3.40) give

$$|V_k(X_1, 0; \epsilon)| \le C \exp(-cX_1)$$

for some constant c between $p/2$ and p. Let the barrier function $V^*(X_1, Y_1; \epsilon)$ be defined by

$$V^*(X_1, Y_1; \epsilon) = C \exp\Big[-\frac{p}{2}\Big(X_1 + \sqrt{\Big(\frac{a}{\epsilon}\Big)^2 + Y_1^2} - \frac{a}{\epsilon}\Big)\Big].$$

Then a computation yields

$$\begin{aligned} &-\Big(\frac{\partial^2 V^*}{\partial {X_1}^2} + \frac{\partial^2 V^*}{\partial {Y_1}^2}\Big) - p\frac{\partial V^*}{\partial X_1} + \epsilon q V^* \\ &= V^*\Big(\frac{p^2}{4}\frac{(a/\epsilon)^2}{(a/\epsilon)^2 + Y_1^2} + \frac{p}{2}\frac{(a/\epsilon)^2}{[(a/\epsilon)^2 + Y_1^2]^{3/2}} + \epsilon q\Big) > 0. \end{aligned}$$

Furthermore, we have

$$\begin{aligned} \big|V_k(0, Y_1; \epsilon)\big| &\le V^*(0, Y_1; \epsilon), \\ \big|V_k(X_1, 0; \epsilon)\big| &\le V^*(X_1, 0; \epsilon), \end{aligned}$$

for sufficiently large values of C in the definition of V^*. By the maximum principle, we are led to have the desired inequalities for $0 \le k \le n+1$, $0 \le X_1 < \infty$, and $0 \le Y_1 < \infty$. This completes the proof.

Now applying the differential operator L_ϵ to the elliptic corner layer V yields

$$L_\epsilon V = \sum_{k=0}^{n+1} \epsilon^k \Big(-\epsilon\frac{\partial^2 V_k}{\partial x^2} - \epsilon\frac{\partial^2 V_k}{\partial y^2} + p\frac{\partial V_k}{\partial x} + qV_k\Big). \tag{3.49}$$

Equation (3.46) can be written as

$$-\epsilon\Big(\frac{\partial^2 V_k}{\partial x^2} + \frac{\partial^2 V_k}{\partial y^2}\Big) + p\frac{\partial V_k}{\partial x} + qV_k = 0, \tag{3.50}$$

for each k. Therefore substitution of (3.50) into (3.49) gives

$$L_\epsilon V = 0, \tag{3.51}$$

uniformly in the closure of Ω. Moreover, let us examine this function V along the boundary of Ω.

i) At $y = 0$,

$$V = -w\Big(\frac{a-x}{\epsilon}, 0; \epsilon\Big) - W\Big(\frac{a-x}{\epsilon}, 0; \epsilon\Big). \tag{3.52a}$$

ii) At $x = a$,

$$V = -v\Big(\frac{a}{\epsilon}, \frac{y}{\epsilon}; \epsilon\Big). \tag{3.52b}$$

iii) At $y = b$, the series

$$V = \sum_{k=0}^{n+1} \epsilon^k V_k\Big(\frac{a-x}{\epsilon}, \frac{b}{\epsilon}; \epsilon\Big) \tag{3.52c}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \leq x \leq a$.

iv) At $x = 0$, the series

$$V = \sum_{k=0}^{n+1} \epsilon^k V_k\Big(\frac{a}{\epsilon}, \frac{y}{\epsilon}; \epsilon\Big) \tag{3.52d}$$

is asymptotically exponentially small with respect to ϵ in the closed interval $0 \leq y \leq b$.

By symmetry, the remaining functions v_k^T, z_k^T, W_k^T, and V_k^T can be constructed similarly along the upper characteristic boundary $y = b$ for the terms TEBL, TPBL, TOCL, and TECL, respectively.

3.7. Asymptotic representation of the solution. Consider the remainder term $R_{n+1}(x, y; \epsilon)$, which is defined by

$$\begin{aligned} R_{n+1} &= u_\epsilon(x,y) - u(x,y;\epsilon) - w(X_1, y;\epsilon) \\ &\quad - v(X, Y_1;\epsilon) - z(x, Y;\epsilon) - W(X_1, Y;\epsilon) - V(X_1, Y_1;\epsilon) \\ &\quad - v^T(X, Y_1^T;\epsilon) - z^T(x, Y^T;\epsilon) - W^T(X_1, Y^T;\epsilon) - V^T(X_1, Y_1^T;\epsilon), \end{aligned}$$

where u_ϵ is the solution of the boundary value problem (1.1), (1.2a,b,c,d); the functions u, w, v, z, W, V are defined in (3.4), (3.9), (3.18), (3.25), (3.37), (3.45), respectively; and analogously for the other terms. Here we have used the stretched variables: $X_1 = (a-x)/\epsilon$, $X = x/\epsilon$, $Y_1 = y/\epsilon$, $Y = y/\sqrt{\epsilon}$, $Y_1^T = (b-y)/\epsilon$, $Y^T = (b-y)/\sqrt{\epsilon}$. Then it follows from the analysis of the preceding sections that the function R_{n+1} satisfies the elliptic differential equation

$$\begin{aligned} L_\epsilon R_{n+1} = \epsilon^{n+1}\Big(&\Delta u_n + \epsilon \frac{\partial^2 w_{n+1}}{\partial y^2} + \frac{\partial^2 w_n}{\partial y^2} - q w_{n+1} \\ &+ \frac{\partial^2 z_n}{\partial x^2} + \frac{\partial^2 W_{n+1}}{\partial Y^2} - q W_{n+1} + \frac{\partial^2 z_n^T}{\partial x^2} + \frac{\partial^2 W_{n+1}^T}{\partial {Y^T}^2} - q W_{n+1}^T\Big), \end{aligned}$$

which is of order $O(\epsilon^{n+1})$ in the closure of Ω, and we have the following conditions for R_{n+1} along the boundary of Ω.

i) At $x = 0$, the remainder

$$\begin{aligned} R_{n+1} = &-w\Big(\frac{a}{\epsilon}, y;\epsilon\Big) - W\Big(\frac{a}{\epsilon}, \frac{y}{\sqrt{\epsilon}};\epsilon\Big) - V\Big(\frac{a}{\epsilon}, \frac{y}{\epsilon};\epsilon\Big) \\ &- W^T\Big(\frac{a}{\epsilon}, \frac{b-y}{\sqrt{\epsilon}};\epsilon\Big) - V^T\Big(\frac{a}{\epsilon}, \frac{b-y}{\epsilon};\epsilon\Big) \end{aligned}$$

is asymptotically exponentially small for $0 \leq y \leq b$.

ii) At $x = a$,

$$R_{n+1} = 0 \quad \text{for } 0 \leq y \leq b.$$

iii) At $y = 0$, the remainder

$$R_{n+1} = -v^T\Big(\frac{x}{\epsilon}, \frac{b}{\epsilon}; \epsilon\Big) - z^T\Big(x, \frac{b}{\sqrt{\epsilon}}; \epsilon\Big) - W^T\Big(\frac{a-x}{\epsilon}, \frac{b}{\sqrt{\epsilon}}; \epsilon\Big) - V^T\Big(\frac{a-x}{\epsilon}, \frac{b}{\epsilon}; \epsilon\Big)$$

is asymptotically exponentially small for $0 \leq x \leq a$.

iv) At $y = b$, the remainder

$$R_{n+1} = -v\Big(\frac{x}{\epsilon}, \frac{b}{\epsilon}; \epsilon\Big) - z\Big(x, \frac{b}{\sqrt{\epsilon}}; \epsilon\Big) - W\Big(\frac{a-x}{\epsilon}, \frac{b}{\sqrt{\epsilon}}; \epsilon\Big) - V\Big(\frac{a-x}{\epsilon}, \frac{b}{\epsilon}; \epsilon\Big)$$

is asymptotically exponentially small for $0 \leq x \leq a$.

By using Theorem 3.3., we have the following estimate:

$$R_{n+1} = O\big(\epsilon^{n+1}\big),$$

which holds uniformly in the closure of Ω. Furthermore, the functions w_{n+1}, v_{n+1}, W_{n+1}, and V_{n+1} are uniformly bounded in the closure of Ω. Consequently, we obtain finally the following:

THEOREM 3.12. *The solution $u_\epsilon(x, y)$ of the boundary value problem* (1.1), (1.2a,b,c,d) *has a uniform approximation $U(x, y; \epsilon)$ in the closure of the rectangular region Ω with error $O\big(\epsilon^{n+1}\big)$, where U is defined by the series.*

$$\begin{aligned} U(x, y; \epsilon) = \sum_{k=0}^{n} \epsilon^k \big[&u_k(x, y) + w_k(X_1, y) + v_k(X, Y_1) + z_k(x, Y) \\ &+ W_k(X_1, Y) + V_k(X_1, Y_1) + v_k^T(X, Y_1^T) + z_k^T(x, Y^T) \\ &+ W_k^T(X_1, Y^T) + V_k^T(X_1, Y_1^T)\big]. \end{aligned}$$

4. Magnetohydrodynamic flow in a rectangular duct with nonconducting walls. The design of magnetohydrodynamic generators, flow-meters, pumps and accelerators requires an understanding of the flows of conducting fluids in rectangular ducts under transverse magnetic fields. These flows have received much attention from theoreticians because the governing equations are linear but the phenomena are neither trivially simple nor physically attainable in the laboratory.

In 1937, Jul. Hartmann [19] solved the one-dimensional problem where the flow was between two parallel walls, the fluid being virtually infinite in directions perpendicular to the imposed transverse magnetic field. Coordinates are then dependent on the transverse coordinate only.

We are concerned with the flow of a steady, incompressible, electrically conducting fluid through a rectangular duct with a uniform, external magnetic field applied transverse to the flow and parallel to two of the walls. Various forms of this problem with different combinations of conducting and nonconducting bounding walls have been considered, see J. A. Shercliff [41], C. C. Chang and T. S. Lundgren [6], W. E. Williams [47], J. C. R. Hunt [21], J. C. R. Hunt and K. Stewartson [22], D. Chiang and T. Lundgren [7], J. C. R. Hunt and J. A. Shercliff [23], D. J. Temperley and L. Todd [42], and L. A. Kalyakin [25] - [28].

As is known from the work of Shercliff [41] and Chang and Lundgren [6], the problem of magnetohydrodynamic flow in a rectangular duct with an applied magnetic field transverse to the axis of the duct is described by two second-order elliptic partial differential equations for the fluid velocity V and the axial component of the magnetic field B, namely, the z-component of the momentum equation

$$\rho\nu\Delta V + \frac{B_0}{\mu_0}\frac{\partial B}{\partial x} = \frac{\partial p}{\partial z},$$

and the z-component of the curl of Ohm's law

$$\Delta B + \sigma\mu_0 B_0 \frac{\partial V}{\partial x} = 0,$$

where ρ is the mass density; ν is the kinematic viscosity; B_0 is the magnitude of the transverse magnetic field which is applied in the x-direction; μ_0 is the permeability in a vacuum; $\partial p/\partial z$ is the pressure gradient, which is a constant; and σ is the electrical conductivity of the incompressible fluid medium. Let the origin be the centerline of the duct and let $2a$ and $2b$ be the lengths of the sides of the duct (see Figure 4.1).

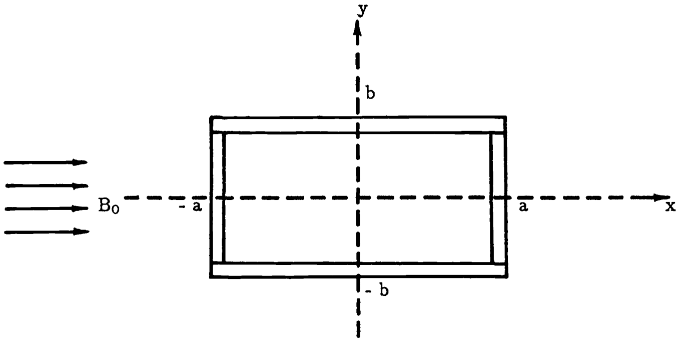


FIG. 4.1. *Rectangular duct.*

In the case of nonconducting walls, the boundary conditions are

$$V = B = 0 \quad \text{at } x = \pm a,$$
$$V = B = 0 \quad \text{at } y = \pm b.$$

We shall examine the flow for large Hartmann number M, which is defined by

$$M = B_0 a\sqrt{\frac{\sigma}{\rho\nu}},$$

or the small value of $\epsilon = 1/M$. With these dimensionless variables

$$\xi = \frac{x}{a}, \qquad \nu = \frac{y}{a}, \qquad V^* = 2\mu_0\sqrt{\rho\nu\sigma}\, B_0^{-1} V,$$
$$B^* = 2B_0^{-1}B, \qquad P = -2\mu_0 a B_0^{-2}\,\frac{\partial p}{\partial z},$$

the problem is to solve

$$\epsilon\Big(\frac{\partial^2 V^*}{\partial \xi^2} + \frac{\partial^2 V^*}{\partial \eta^2}\Big) + \frac{\partial B^*}{\partial \xi} = -P,$$
$$\epsilon\Big(\frac{\partial^2 B^*}{\partial \xi^2} + \frac{\partial^2 B^*}{\partial \eta^2}\Big) + \frac{\partial V^*}{\partial \xi} = 0,$$

with the boundary conditions

$$V^* = B^* = 0 \quad \text{at } \xi = \pm 1,$$
$$V^* = B^* = 0 \quad \text{at } \eta = \pm \ell, \ \ell = b/a.$$

The equations can be decoupled by the change of variables

$$u = P^{-1}(V^* - B^*),$$

and

$$v = P^{-1}(V^* - B^*).$$

The equations for u and v are

$$-\epsilon \Delta u + \frac{\partial u}{\partial \xi} = 1,$$

and

$$-\epsilon \Delta v - \frac{\partial v}{\partial \xi} = 1,$$

with the boundary conditions

$$u = v = 0 \quad \text{at } \xi = \pm 1,$$
$$u = v = 0 \quad \text{at } \eta = \pm \ell.$$

It is obvious that having determined the functions u, we can find v by the relationship

$$v(\xi, \eta) = u(-\xi, \eta).$$

Hence the function u alone needs to be investigated. The asymptotic approximation to u, being uniformly valid in the closure of the rectangular duct for large Hartmann number M, is still an open question in the literature, see Shercliff [41], Roberts [40, pp. 186–190], Cook, Ludford and Walker [9], and Temperley [43].

With an application of the preceding analysis, we conclude that the solution u can be written as

$$u(\xi, \eta) = (1 + \xi) - 2\exp(-\xi_2) + v_1(\xi_1, \eta_1; \epsilon) + V_1(\xi_2, \eta_1; \epsilon)$$
$$+ v_2(\xi_2, \eta_1; \epsilon) + V_2(\xi_2, \eta_2; \epsilon) + \text{AES},$$

where the stretched variables are

$$\xi_1 = \frac{\xi+1}{\epsilon}, \qquad \xi_2 = \frac{1-\xi}{\epsilon}, \qquad \eta_1 = \frac{\ell-\eta}{\epsilon} \quad \text{and} \quad \eta_2 = \frac{\eta+\ell}{\epsilon};$$

the function $1+\xi$ is known as a "core flow"; the function $-2\exp(-\xi_2)$ is called a "Hartmann boundary layer"; the functions $v_1(\xi_1,\eta_1;\epsilon)$ and $v_2(\xi_1,\eta_2;\epsilon)$ satisfy the elliptic partial differential equation of the form

$$-\left(\frac{\partial^2 v_i}{\partial {\xi_1}^2} + \frac{\partial^2 v_i}{\partial {\eta_i}^2}\right) + \frac{\partial v_i}{\partial \xi_1} = 0 \quad \text{for } i = 1,2,$$

over the quarter plane $0 < \xi_1 < \infty$ and $0 < \eta_i < \infty$ under the boundary conditions

$$v_i(\xi_1, 0; \epsilon) = -\epsilon\xi_1, \quad v_i(0,\eta_1;\epsilon) = 0,$$

and

$$v_i(\xi_1,\eta_i;\epsilon) \to 0 \quad \text{as } \xi_1^2 + \eta_i^2 \to \infty \text{ and } \eta_i > 0;$$

and the functions $V_1(\xi_2,\eta_1;\epsilon)$ and $V_2(\xi_2,\eta_2;\epsilon)$ satisfy the elliptic partial differential equation of the form

$$\left(\frac{\partial^2 V_i}{\partial {\xi_2}^2} + \frac{\partial^2 V_i}{\partial {\eta_i}^2}\right) + \frac{\partial V_i}{\partial \xi_2} = 0 \quad \text{for } i = 1,2,$$

over the quarter plane $0 < \xi_2 < \infty$ and $0 < \eta_i < \infty$ under the boundary conditions

$$V_i(\xi_2, 0; \epsilon) = 2\exp(-\xi_2),$$
$$V_i(0,\eta_i;\epsilon) = -v_i\left(\frac{2}{\epsilon},\eta_i;\epsilon\right),$$

and

$$V_i(\xi_2,\eta_i;\epsilon) \to 0 \quad \text{as } \xi_2^2 + \eta_i^2 \to \infty;$$

and AES denotes asymptotically exponentially small terms with respect to ϵ in the closure of the rectangular duct.

REFERENCES

[1] A. AZIZ AND T. Y. NA, *Perturbation Methods in Heat Transfer*, Hemisphere Publishing Corporation, New York, 1984.

[2] A. AZZAM, *On the first boundary value problem for elliptic equations in regions with corners*, Arabian J. Sci. Engrg., 4 (1979), pp. 129–135.

[3] A. BEJAN, *Convection Heat Transfer*, John Wiley, New York, 1984.

[4] V. F. BUTUZOV, *Asymptotic properties of solutions of singularly perturbed elliptic equations in rectangular regions*, Differentsial'nye Uravneniya, 11 (1975), pp. 1030-1041. Differential Equations, 11 (1975), pp. 780–787.

[5] J. R. CANNON, *The one-dimensional heat equation*, Encyclopedia of Mathematics and Its Application, Vol. 23, Addison-Wesley, Reading, MA, 1984.

[6] C. C. CHANG AND T. S. LUNDGREN, *Duct flow in magnetohydrodynamics*, Z. Angew. Math. Phys., 12 (1961), pp. 100–114.

[7] D. CHIANG AND T. LUNDGREN, *Magnetohydrodynamic flow in a rectangular duct with perfectly conducting electrodes*, Z. Angew. Math. Phys., 18 (1967), pp. 92–105.

[8] L. P. COOK AND G. S. S. LUDFORD, *The behavior as* $\epsilon \to 0+$ *of solutions to* $\epsilon\nabla^2 w = \partial w/\partial y$ *in* $|y| \leq 1$ *for discontinuous boundary data*, this Journal, 2 (1971), pp. 567–594.

[9] L. P. COOK, G. S. S. LUDFORD AND J. S. WALKER, *Corner regions in the asymptotic solution of* $\epsilon\nabla^2 u = \partial u/\partial y$ *with reference to* MHD *duct flow*, Proc. Camb. Phil. Soc., 72 (1972), pp. 117–122.

[10] L. P. COOK AND G. S. S. LUDFORD, *The behavior as* $\epsilon \to 0+$ *of solutions to* $\epsilon\nabla^2 w = \partial w/\partial y$ *on the rectangle* $0 \leq x \leq \ell$, $|y| \leq 1$, this Journal, 4 (1973), pp. 161–184.

[11] W. ECKHAUS AND E. M. DE JAGER, *Asymptotic solutions of singular perturbation problems for linear differential equations of elliptic type*, Arch. Rational Mech. Anal., 23 (1966), pp. 26–86.

[12] W. ECKHAUS, *Boundary layers in linear elliptic singular perturbations*, SIAM Rev., 14 (1972), pp. 225–270.

[13] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbations*, North-Holland Publishing Co., New York, 1979.

[14] K. O. FRIEDRICHS, *Theory of viscous fluids*, Fluid Dynamics, Chapter 4, Brown University, Providence, RI, 1941.

[15] J. GRASMAN, *On singular perturbations and parabolic boundary layers*, J. Engrg. Math., 2 (1968), pp. 163–172.

[16] J. GRASMAN, *On the Birth of Boundary Layers*, Doctoral thesis, Delft University of Technology, Math. Cent. Tract no. 36, Mathematisch Centrum, Amsterdam, 1971.

[17] J. GRASMAN, *The birth of a boundary layer in an elliptic singular perturbation problem*, Spectral Theory and Asymptotics of Differential Equations, E. M. De Jager, ed., North-Holland, Amsterdam, 1973, pp. 175-179.

[18] J. GRASMAN, *An elliptic singular perturbation problem with almost characteristic boundaries*, J. Math. Anal., 46 (1974), pp. 438–446.

[19] J. HARTMANN, *Theory of the laminar flow of an electrically conducting liquid in a homogeneous magnetic field*, Danske Vid. Selsk. Mat.-Fys. Medd., 15 (6) (1937).

[20] F. A. HOWES, *Perturbed boundary value problems whose reduced solutions are nonsmooth*, Indiana Univ. Math. J., 30 (1981), pp. 267–280.

[21] J. C. R. HUNT, *Magnetohydrodynamic flow in rectangular ducts*, J. Fluid Mech., 21 (1965), pp.577–590.

[22] J. C. R. HUNT AND K. STEWARTSON, *Magnetohydrodynamic flow in rectangular ducts.* II, J. Fluid Mech., 23 (1965), pp. 563–581.

[23] J. C. R. HUNT AND J. A. SHERCLIFF, *Magnetohydrodynamics at high Hartmann number*, Annual Review of Fluid Mechanics, 3 (1971), pp. 37–62.

[24] A. M. IL'IN AND E. F. LELIKOVA, *A method of joining asymptotic expansions for the equation* $\epsilon\Delta u - a(x,y)u_y = f(x,y)$ *in a rectangle*, Mat. Sb. (N.S.), 96(138) (1975), pp. 568-583. Math. USSR Sb., 25 (1975), pp. 533–548.

[25] L. A. KALYAKIN, *An asymptotic formula for the solution of a magnetohydrodynamic problem containing a small parameter*, I. *Rectilinear flow in a rectangular channel. A superconducting wall perpendicular to the magnetic field*, Differentsial'nye Uravneniya, 15 (1979a), pp. 668-680. Differential Equations, 15 (1979a), pp. 467–476.

[26] L. A. KALYAKIN, *Asymptotic expansion of the solution of a problem in magnetohydrodynamics involving a small parameter*, II. *Linear flow in a channel with a rectangular projection and a superconducting wall perpendicular to the magnetic field*, Differentsial'nye Uravneniya, 15 (1979b), pp. 1873-1887. Differential Equations, 15 (1979b), pp. 1336–1346.

[27] L. A. KALYAKIN, *Asymptotic expansion of a solution of a system of two linear* MHD *equations with a singular perturbation*, I. *A standard problem in an elliptic layer*, Differentsial'nye Uravneniya, 18 (1982), pp. 1724-1738. Differential Equations, 18 (1982), pp. 1238-1249.

[28] L. A. KALYAKIN, *Asymptotic expansion of a solution of a* MHD *system of two linear equations with a singular perturbation*, II. *A complete asymptotic expansion*, Differentsial'nye Uravneniya, 19 (1983), pp. 628-644. Differential Equations, 19 (1983), pp. 461–475.

[29] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[30] J. K. KNOWLES AND R. E. MESSICK, *On a class of singular perturbation problems*, J. Math. Anal. Appl., 9 (1964), pp. 42–58.

[31] G. E. LATTA, *Singular Perturbation Problems*, Doctoral dissertation, California Institute of Technology, Pasadena, CA, 1951.

[32] N. LEVINSON, *The first boundary value problem for $\epsilon\Delta u + Au_x + Bu_y + Cu = D$ for small ϵ*, Annals of Math., 51 (1950), pp. 428–455.

[33] P.-C. LIN AND F.-W. LIU, *The necessary and sufficient condition of uniformly convergent difference schemes for the elliptic-parabolic partial differential equation with a small parameter*, Appl. Math. Mech. (English Ed.), 5 (1984), pp. 1047–1055.

[34] P.-C. LU, *Introduction to the Mechanics of Viscous Fluids*, Holt, Rinehart and Winston, New York, 1973.

[35] J. MAUSS, *Étude de la solution asymptotique du problème de la couche limit parabolique*, C. R. Acad. Sc. Paris Série A, 265 (1967), pp. 838–840.

[36] J. MAUSS, *Problèmes de perturbations singulière*, Doctoral thesis, Dep. de Méchanique, Université de Paris, 1971.

[37] R. E. O'MALLEY, JR., *Topics in singular perturbations*, Adv. in Math., 2 (1968), pp. 365–470.

[38] D. W. PEACEMAN, *Fundamentals of Numerical Reservoir Simulation*, Elsevier Scientific Publishing Co., New York, 1977.

[39] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[40] P. H. ROBERTS, *An Introduction to Magnetohydrodynamics*, American Elsevier Publishing Company, Inc., New York, 1967.

[41] J. A. SHERCLIFF, *Steady motion of conducting fluids in pipes under transverse magnetic fields*, Proc. Camb. Phil. Soc., 49 (1953), pp. 136–144.

[42] D. J. TEMPERLEY AND L. TODD, *The effects of wall conducting in magnetohydrodynamic duct flow at high Hartmann numbers*, Proc. Camb. Phil. Soc., 69 (1971), pp. 337-351.

[43] D. J. TEMPERLEY, *Alternative approaches to the asymptotic solution of $\epsilon\nabla^2 u = \partial u/\partial y$, $0 < \epsilon \ll 1$, over a rectangle*, Z. Angew. Math. Mech., 56 (1976), pp. 461–468.

[44] M. VAN DYKE, *Perturbation Methods in Fluid Dynamics*, Academic Press, New York, 1964. (Annotated edition, Parabolic Press, Stanford, California, 1975.)

[45] M. I. VISHIK AND L. A. LYUSTERNIK, *Regular degeneration and boundary layer for linear differential equations with small parameter*, Uspekhi. Mat. Nauk., 12 (5) (1957), pp. 3–122. Amer. Math. Soc., Transl. (2), 20 (1962), pp. 239–364.

[46] W. WASOW, *Asymptotic solution of boundary value problems for the differential equation $\Delta U + \lambda(\partial/\partial x)U = \lambda f(x,y)$*, Duke Math. J., 11 (1944), pp. 405–415.

[47] W. E. WILLIAMS, *Magnetohydrodynamic flow in a rectangular tube at high Hartmann number*, J. Fluid Mech., 16 (1963), pp. 262–268.

A TRIPLE PRODUCT THEOREM FOR HYPERGEOMETRIC SERIES*

PETER HENRICI†

Abstract. The product of three hypergeometric series of type ${}_0F_1$ with arguments x, ωx, $\omega^2 x$ (ω = third root of unity) is expressed as a single hypergeometric series of type ${}_2F_7$. The proof uses differential equations; the basic idea (elimination of irreducible terms by means of the Cayley-Hamilton theorem) is more generally applicable.

Key words. hypergeometric series, product theorems for special functions

AMS(MOS) subject classifications. 33A30, 33A35, 34B30

1. THEOREM. *Let c be complex, $6c \neq 0, -1, -2, \cdots$, and let $\omega := \exp(2\pi i/3)$. In the usual notation for generalized hypergeometric series (see Bailey [1936]) there holds*

$$(1)\qquad \begin{aligned}&{}_0F_1[6c;\, x]\,{}_0F_1[6c;\, \omega x]\,{}_0F_1[6c;\, \omega^2 x]\\ &\quad = {}_2F_7\left[\begin{matrix}3c-\frac14,\, 3c+\frac14;\\ 6c,\, 2c,\, 2c+\frac13,\, 2c+\frac23,\, 4c-\frac13,\, 4c,\, 4c+\frac13;\end{matrix}\ (\tfrac49)^3x^3\right].\end{aligned}$$

The proof is given in §§ 2-4 below. The idea is to regard (1) as an identity between formal power series and to show that the series on the left and on the right formally satisfy the same hypergeometric differential equation. Section 5 features corollaries and remarks.

2. Derivatives. If

$$P = \sum_{n=0}^{\infty} a_n x^n$$

is any formal power series (over $\mathbb{C}$, or over any field of characteristic 0), we call

$$\theta P := \sum_{n=0}^{\infty} n a_n x^n$$

the *derivate* of P. (The relation to the ordinary derivative P' is obvious.) Derivatives satisfy many of the usual rules of calculus, such as the product rule,

$$\theta(PQ) = P\theta Q + Q\theta P,$$

and the further rules (such as the Leibniz rule for the higher derivatives of a product) that follow from it.

In terms of derivates, the (formal) generalized hypergeometric series

$$P = {}_pF_q\left[\begin{matrix}\alpha_1, \alpha_2, \cdots, \alpha_p;\\ \beta_1, \beta_2, \cdots, \beta_q;\end{matrix}\ x\right],$$

where p and q are arbitrary nonnegative integers, satisfies the differential equation

$$(2)\qquad \{\theta(\theta+\beta_1-1)\cdots(\theta+\beta_q-1) - x(\theta+\alpha_1)\cdots(\theta+\alpha_p)\}P = 0;$$

see Bailey [1936]. Similarly if $a \in \mathbb{C}$ and s is a positive integer,

$$Q = {}_pF_q\left[\begin{matrix}\alpha_1, \alpha_2, \cdots, \alpha_p;\\ \beta_1, \beta_2, \cdots, \beta_q;\end{matrix}\ ax^s\right]$$

* Received by the editors December 16, 1985; accepted for publication December 1, 1986.

† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514. It is with great regret that SIAM learned of the untimely death of Peter Henrici in March 1987. The editorial staff of the SIAM Journal on Mathematical Analysis feel especially his absence, as Peter was an active member of the Editorial Board for many years.—J.B.

satisfies

$$\{\theta(\theta+s\beta_1-s)\cdots(\theta+s\beta_q-s)-as^{n+1-m}x^s(\theta+s\alpha_1)\cdots(\theta+s\alpha_p)\}Q=0. \tag{3}$$

Conversely, if the β_i are such that none of the numbers $s\beta_i-s$ is zero or a negative integer, the series Q is a formal solution of (3), and it is the only such solution that has zeroth coefficient 1.

3. The triple product $\Pi_0F_1[6c,\omega^k x]$ and its derivatives. To prove (1), let

$$U:={}_0F_1[6c;x],\quad V:={}_0F_1[6c;\omega x],\quad W:={}_0F_1[6c;\omega^2x].$$

If $a:=6c-1$, the differential equation satisfied by U is

$$\theta(\theta+a)-xU=0.$$

We thus have

$$\theta^2U=-a\theta U+xU \tag{4a}$$

and similarly

$$\theta^2V=-a\theta V+\omega xV, \tag{4b}$$

$$\theta^2W=-a\theta W+\omega^2xW. \tag{4c}$$

To discover a differential equation satisfied by

$$Z:=UVW,$$

we form the derivatives θZ, $\theta^2Z,\cdots$ and eliminate derivates of U, V, W of order $\geqq2$ by means of (4). Some of the remaining terms $\sum\theta^iU\theta^jV\theta^kW$ may be expressed in terms of Z or

$$\theta Z=UV\theta W+VW\theta U+WU\theta V,$$

and thus are "reducible," but many terms cannot be so expressed. It is convenient to define a basis for the space of irreducible terms as follows:

$$\begin{aligned}
F&:=U\theta V\theta W+V\theta W\theta U+W\theta U\theta V,\\
P&:=\theta U\theta V\theta W,\\
A&:=VW\theta U+\omega WU\theta V+\omega^2UV\theta W,\\
B&:=U\theta V\theta W+\omega V\theta W\theta U+\omega^2W\theta U\theta V,\\
C&:=VW\theta U+\omega^2WU\theta V+\omega^4UV\theta W,\\
D&:=U\theta V\theta W+\omega^2V\theta W\theta U+\omega^4W\theta U\theta V.
\end{aligned} \tag{5}$$

We require the derivates of these irreducible forms. We have, for instance,

$$\theta F=3\theta U\theta V\theta W+U\theta^2V\theta W+U\theta V\theta^2W+V\theta^2W\theta U+V\theta W\theta^2U+W\theta^2U\theta V+W\theta U\theta^2V$$

and thus, using (4),

$$\begin{aligned}
\theta F&=3P+U(\omega xV-a\theta V)\theta W+U\theta V(\omega^2xW-a\theta W)+V(\omega^2xW-a\theta W)\theta U\\
&\quad+V\theta W(xU-a\theta U)+W(xU-a\theta U)\theta V+W\theta U(\omega xV-a\theta V)\\
&=3P+x[(1+\omega)UV\theta W+(\omega+\omega^2)VW\theta U+(\omega^2+1)WU\theta V]\\
&\quad-2a[U\theta V\theta W+V\theta W\theta U+W\theta U\theta V]
\end{aligned}$$

or, using $1+\omega+\omega^2=0$,

$$\theta F=-2aF+3P-xA. \tag{6a}$$

In a similar manner we obtain

$$\begin{aligned} \theta P &= -3aP+xB, \\ \theta A &= -aA-B, \\ \theta B &= -2aB+2xC, \\ \theta C &= -aC-D-3xZ, \\ \theta D &= -2aD-x\theta Z. \end{aligned} \tag{6b}$$

Defining the vector of irreducible terms

$$\mathbf{v}=\begin{pmatrix} F \\ P \\ xA \\ xB \\ x^2C \\ x^2D \end{pmatrix},$$

the matrix

$$\mathbf{M}=\begin{pmatrix} -2a & 3 & -1 & 0 & 0 & 0 \\ 0 & -3a & 0 & 1 & 0 & 0 \\ 0 & 0 & 1-a & -1 & 0 & 0 \\ 0 & 0 & 0 & 1-2a & 2 & 0 \\ 0 & 0 & 0 & 0 & 2-a & -1 \\ 0 & 0 & 0 & 0 & 0 & 2-2a \end{pmatrix},$$

and the vector of operators

$$\mathbf{s}(\theta)=\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ -\theta \end{pmatrix},$$

the relations (6) are expressed as

$$\theta\mathbf{v}=\mathbf{M}\mathbf{v}+x^3\mathbf{s}(\theta)Z. \tag{7}$$

(Here we have used, for instance, that $\theta(x^2C)=2xC+x^2\theta C$.)

4. Elimination of the irreducible terms. Using (4), we have

$$\begin{aligned} \theta^2 Z &= UV\theta^2 W+VW\theta^2 U+WU\theta^2 V+2(U\theta V\theta W+V\theta W\theta U+W\theta U\theta V) \\ &= UV(\omega^2 xW-a\theta W)+VW(xU-a\theta U)+WU(\omega xV-a\theta V)+2(\cdots) \\ &= -a\theta Z+2F. \end{aligned}$$

Introducing $\mathbf{c}:=(2,0,0,0,0,0)^T$, this is written:

$$\theta(\theta+a)Z=\mathbf{c}^T\mathbf{v}. \tag{8}$$

By successive applications of θ, always using (7), there results

$$\theta^2(\theta+a)Z=\mathbf{c}^T(\mathbf{M}\mathbf{v}+x^3\mathbf{s}(\theta)Z),$$
$$\theta^3(\theta+a)Z=\mathbf{c}^T\{\mathbf{M}^2\mathbf{v}+x^3[\mathbf{M}+(\theta+3)\mathbf{I}]\mathbf{s}(\theta)Z\},$$

and generally, as may be verified by induction,

(9) $$\theta^{k+1}(\theta+a)Z=\mathbf{c}^T\{\mathbf{M}^k\mathbf{v}+x^3[\mathbf{M}^{k-1}+(\theta+3)\mathbf{M}^{k-2}+\cdots+(\theta+3)^{k-1}\mathbf{I}]\mathbf{s}(\theta)Z\},$$

$k=0,1,\cdots$. A hypergeometric differential equation for Z (of the type (3) where $s=3$) will result if $\mathbf{v}$ can be eliminated by forming a suitable linear combination of sufficiently many of the relations (9). To effect the elimination, the Cayley-Hamilton theorem offers itself. The characteristic polynomial of the matrix $\mathbf{M}$ is

$$p(\lambda)=(\lambda+2a)(\lambda+3a)\cdots$$
$$=\sum_{k=0}^{6}\gamma_k\lambda^k,$$

say. For later reference we note that

(10) $$\gamma_6=1,\quad \gamma_5=11a-6,\quad \gamma_4=49a^2-57a+13.$$

We multiply the kth relation (9) by γ_k ($k=0,1,\cdots,6$) and add. On the left we obtain

$$\sum_{k=0}^{6}\gamma_k\theta^{k+1}(\theta+a)Z=\theta(\theta+a)p(\theta)Z.$$

On the right, the factor of $\mathbf{v}$ is

$$\mathbf{c}^T\sum_{k=0}^{6}\gamma_k\mathbf{M}^k=\mathbf{c}^Tp(\mathbf{M}),$$

which vanishes by the Cayley-Hamilton theorem. The vector $\mathbf{v}$ thus is eliminated. There remains

$$x^3\mathbf{c}^T\left\{\sum_{k=1}^{6}\gamma_k[\mathbf{M}^{k-1}+(\theta+3)\mathbf{M}^{k-2}+\cdots+(\theta+3)^{k-1}\mathbf{I}]\mathbf{s}(\theta)Z\right\}$$
$$=x^3\mathbf{c}^T\left\{\sum_{k=1}^{6}\gamma_k(\theta+3)^{k-1}\mathbf{I}+\sum_{k=2}^{6}\gamma_k(\theta+3)^{k-2}\mathbf{M}\right.$$
$$\left.+\sum_{k=3}^{6}\gamma_k(\theta+3)^{k-3}\mathbf{M}^2+\cdots+\gamma_6\mathbf{M}^5\right\}\mathbf{s}(\theta)Z.$$

We evaluate this expression by first forming the iterates $\mathbf{c}^T\mathbf{M}^j$, $j=0,1,2,\cdots,5$. Since only the last two components of $\mathbf{s}(\theta)$ are $\neq 0$, only the last two components of these iterates are relevant for the result. These components are different from zero only for $j\geqq 3$, and they have the following values (see Table 1). With the values of the γ_k given

TABLE 1

k	5	6
$(\mathbf{c}^T\mathbf{M}^3)_k$	16	0
$(\mathbf{c}^T\mathbf{M}^4)_k$	$-120a+52$	-16
$(\mathbf{c}^T\mathbf{M}^5)_k$	$584a^2-452a+128$	$152a-84$

in (10), we thus find

$$\begin{aligned}\mathbf{c}^T&\left\{\sum_{j=0}^{5}\sum_{k=j+1}^{6}\gamma_k(\theta+3)^{k-1}\mathbf{M}^j\right\}\mathbf{s}(\theta)\\ &=\mathbf{c}^T\{[\gamma_4+\gamma_5(\theta+3)+\gamma_6(\theta+3)^2]\mathbf{M}^3\mathbf{s}(\theta)\\ &\qquad+[\gamma_5+\gamma_6(\theta+3)]\mathbf{M}^4\mathbf{s}(\theta)+\gamma_6\mathbf{M}^5\mathbf{s}(\theta)\}\\ &=64\theta^2+(192a+192)\theta+114a^2+288a+108\\ &=64\left(\theta+\frac{3a}{2}+\frac{3}{4}\right)\left(\theta+\frac{3a}{2}+\frac{9}{4}\right).\end{aligned}$$

It follows that $Z=UVW$ is a solution of the generalized hypergeometic differential equation

$$\left\{\theta(\theta+a)p(\theta)-64x^3\left(\theta+\frac{3a}{2}+\frac{3}{4}\right)\left(\theta+\frac{3a}{2}+\frac{9}{4}\right)\right\}Z=0.$$

By (3), if expressed in terms of $c=a/6+1/6$, this solution must be identical with the series on the right of (1).

5. Remarks and corollaries. (i) By using obvious identities for Pochhammer symbols, the series on the right of (1) may also be expressed as

$$(11)\qquad {}_2F_7=\sum_{k=0}^{\infty}\frac{(12c-1+3k)_k}{k!(6c)_k(6c)_{2k}(6c)_{3k}}x^{3k}.$$

(ii) Any product theorem for hypergeometric series, i.e., any theorem expressing a product of hypergeometric series as a single such series, on comparing coefficients furnishes a formula expressing a sum of products of Pochhammer symbols (or of binomial coefficients) by a single such product. In this manner we obtain from (1), using the representation (11) and letting $b:=6c$,

$$(12)\qquad \sum_{k+l+m=n}\frac{\omega^{l-m}}{k!l!m!(b)_k(b)_l(b)_m}=\begin{cases}0, & n\not\equiv 0 \bmod 3,\\ \dfrac{(2b+3q-1)_q}{q!(b)_q(b)_{2q}(b)_{3q}}, & n=3q.\end{cases}$$

For instance, for $b=1$ this yields

$$(13)\qquad \sum_{0\leqq k\leqq m\leqq n}\binom{n}{m}^2\binom{m}{k}^2\omega^{k+m}=\begin{cases}0, & n\not\equiv 0 \bmod 3,\\ \dbinom{4q}{2q}\dbinom{2q}{q}, & n=3q.\end{cases}$$

Formula (12) is of a similar type as the formulas

$$(14)\qquad \sum_{k+l+m=n}\frac{\omega^{l-m}}{k!l!m!}=\begin{cases}1, & n=0,\\ 0, & n>0,\end{cases}$$

and

$$(15)\qquad \sum_{k+l+m=n}\frac{(a)_k(a)_l(a)_m}{k!l!m!}\omega^{l-m}=\begin{cases}0, & n\not\equiv 0 \bmod 3,\\ (a)_q/q!, & n=3q,\end{cases}$$

which follows from the trivial product theorems[1]

$$ {}_0F_0[x]\,{}_0F_0[\omega x]\,{}_0F_0[\omega^2 x] = 1 $$

(i.e., $\exp((1+\omega+\omega^2)x)=1$) and

$$ {}_1F_0[a;\, x]\,{}_1F_0[a;\, \omega x]\,{}_1F_0[a;\, \omega^2 x] = {}_1F_0[a;\, x^3] $$

(i.e., $(1-x)^{-a}(1-\omega x)^{-a}(1-\omega^2 x)^{-a} = (1-x^3)^{-a}$).

(iii) The method used in this work has the advantage of not requiring the result to be known in advance. Although the algebraic manipulations, if done by hand, soon become unmanageable, the method can be used in principle whenever the irreducible terms in the derivatives of a product satisfy a recurrence relation of the general form (7).

REFERENCE

W. N. Bailey [1936], *Generalized Hypergeometric Series*, University Press, Cambridge.

[1] According to an observation by a referee, the cases $n \equiv 0 \bmod 3$ of formulas (12), (13), and (15) are special cases of the following easily derived summation formula:

If N is an integer greater than one, $\omega_N = \exp(2\pi i/N)$, and if $a_0, a_1, a_2, \cdots$ is a sequence of complex numbers, then

$$ \sum_{\substack{n_1+\cdots+n_N=n\\ n_1,\cdots,n_N \geqq 0}} a_{n_1}a_{n_2}\cdots a_{n_N}\omega_N^{n_1+2n_2+\cdots+(N-1)n_{N-1}} = 0,\ n \not\equiv N \bmod N. $$

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

CANONICAL EQUATIONS AND SYMMETRY TECHNIQUES FOR q-SERIES*

A. K. AGARWAL†, E. G. KALNINS‡ AND WILLARD MILLER, JR.§

Abstract. The authors introduce symmetry techniques for the classification and derivation of generating functions for families of basic hypergeometric functions, in analogy with the Lie theory techniques for ordinary hypergeometric functions. To each family of basic hypergeometric functions there is associated a canonical system of partial q-difference equations and the symmetries of these equations are used to derive q-series identities and orthogonality relations for the special functions.

Key words. q-series, q-difference equations, basic hypergeometric functions

AMS(MOS) subject classifications. 33A30, 33A65, 39A10, 20N99

1. Introduction. In [11], [13], [14] a Lie algebraic method was developed which associated with each family of multivariable hypergeometric functions a canonical system of partial differential equations constructed from the differential recurrence relations obeyed by the family. (The basic idea behind this method followed from the work of Weisner [16].) The hypergeometric functions arise by partial separation of variables in the canonical systems and any analytic solution of these equations can be considered as a generating function for this family. Furthermore the generating functions can be characterized in terms of symmetry operators for the canonical systems.

In this paper we present the foundations of an analogous theory for families of many-variable basic hypergeometric functions. To each family we associate a canonical system of partial q-difference equations constructed from the q-difference recurrence relations obeyed by the family. The basic hypergeometric functions arise by partial separation of variables in the canonical systems and any analytic solution of these equations is a generating function for the family. Symmetry operators for the canonical system can be used to characterize the generating functions. Thus a direct link is established between symmetries of the canonical system and identities obeyed by q-series.

In § 2 we show how to derive the canonical system of q-difference equations associated with a given family of q-series, using as examples the one-variable hypergeometric functions ${}_r\varphi_s$ and the two-variable function f_2, a q-analogue of the Appell function F_2. In § 3 we describe how to relate two different families of basic hypergeometric functions, that is, the procedures of *embedding* and *augmentation*. In the procedure of embedding the canonical system for one basic hypergeometric family restricts through a specialization of variables to the canonical system for a second hypergeometric family, so that the restricted family can be considered as a generating function for the second family. Augmentation is a process inverse to this. By augmentation we can write the defining equations for a generating function as the restriction of a canonical system of higher dimension.

In § 4 we apply our techniques to derive and characterize in terms of symmetries, a variety of generating functions for the families ${}_r\varphi_s$. In § 5 we treat the family ${}_2\varphi_1$ in somewhat more detail. (This family needs special treatment because its canonical

* Received by the editors February 19, 1986; accepted for publication (in revised form) November 5, 1986.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

‡ Mathematics Department, University of Waikato, Hamilton, New Zealand.

§ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by National Science Foundation grant MCS 82-19847.

system admits symmetries not shared by the systems for general ${}_r\varphi_s$.) Furthermore we show how orthogonality relations for q-series follow from symmetry ideas.

Our theory provides a simple uniform procedure for derivation and symmetry classification of a wide variety of q-identities, in analogy with the Lie theoretic procedures for ordinary hypergeometric series. The full power of the theory becomes evident in the study of many-variable q-series and in the study of Askey-Wilson polynomials, as we will show in future papers. However, in distinction to the case of differential equations we do not have the tools of local Lie transformation group theory or the relationship between Lie symmetries and separation of variables to help us obtain the generating functions in the most compact form. Our procedures enable us to classify and characterize generating functions in terms of symmetry operators; unaided, they do not enable us to write the generating functions in simplest form, i.e., factorized or in terms of a new choice of variables. It will be very interesting to see if (as is the case for differential equations) factorization and coordinates have symmetry operator interpretations.

The symmetry techniques presented here apply to formal power series and are essentially independent of convergence criteria. Hence, we shall ordinarily not specify the domains of validity for the identities derived in this paper. In most cases they can be determined easily for one-variable hypergeometric functions by the ratio test. For multivariable hypergeometric functions the full domain of convergence may be very difficult to determine (or even unknown). In those cases one can specialize some of the parameters in the functions (so that infinite series truncate to finite series for example) to guarantee convergence.

Finally we note that the symbolic method of Burchnall and Chaundy for ordinary hypergeometric series [4], [5], and some works of Hahn on q-series [7], [8] contain points of similarity with our method, although these authors did not use symmetry techniques.

2. The "basic" idea.

We begin our study of canonical equations for q-series by deriving the canonical form associated with the q-hypergeometric functions ${}_r\varphi_s$:

$$(2.1)\qquad {}_r\varphi_s\left(\begin{matrix}a_1,\cdots,a_r\\ b_1,\cdots,b_s\end{matrix};x\right)=\sum_{n=0}^{\infty}\frac{(a_1;q)_n\cdots(a_r;q)_n x^n}{(b_1;q)_n\cdots(b_s;q)_n(q;q)_n}$$

where $a_i=q^{\alpha_i}$, $b_j=q^{\beta_j}$ and

$$(2.2)\qquad (a;q)_n=\frac{(a;q)_\infty}{(aq^n;q)_\infty},\qquad (a;q)_\infty=\prod_{m=0}^{\infty}(1-q^m a).$$

Here α_i, β_j, x are complex variables, $(\beta_j\neq 0,-1,-2,\cdots)$ and we normally require that $0<q<1$. Note that for n a nonnegative integer we have

$$(2.3)\qquad (a;q)_n=(1-a)(1-qa)\cdots(1-q^{n-1}a).$$

As is well known [3], [5] ${}_r\varphi_s$ is a q-analogue of the hypergeometric series

$$(2.4)\qquad {}_rF_s\left(\begin{matrix}\alpha_1,\cdots,\alpha_r\\ \beta_1,\cdots,\beta_s\end{matrix};x\right)=\sum_{n=0}^{\infty}\frac{(\alpha_1)_n\cdots(\alpha_r)_n}{(\beta_1)_n\cdots(\beta_s)_n}\frac{x^n}{n!}$$

where

$$(2.5)\qquad (\alpha)_n=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$$

and $\Gamma(\,\cdot\,)$ is the gamma function. For n a nonnegative integer

$$(2.6)\qquad (\alpha)_n=\alpha(\alpha+1)\cdots(\alpha+n-1).$$

Here ${}_r\varphi_s$ and ${}_rF_s$ are related by

$$(2.7)\qquad {}_rF_s\left(\begin{matrix}\alpha_i\\ \beta_j\end{matrix};\,x\right)=\lim_{q\to 1}{}_r\varphi_s\left(\begin{matrix}a^i\\ b_j\end{matrix};\,\frac{x}{(1-q)^{r-s-1}}\right).$$

Let T_u be the q-dilation operator corresponding to the variable u, i.e., T_u maps a function f of the variables $u, v, w, \cdots$ to the function

$$(2.8)\qquad T_uf(u,v,w,\cdots)=f(qu,v,w,\cdots).$$

From the q-series (2.1) one can easily verify the recurrence relations

$$(1-a_kT_x)\,{}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,x\right)=(1-a_k)\,{}_r\varphi_s\left(\begin{matrix}e^ka_i\\ b_j\end{matrix};\,x\right),\qquad 1\leqq k\leqq r,$$

$$(2.9)\qquad (1-b_lq^{-1}T_x)\,{}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,x\right)=(1-b_lq^{-1})\,{}_r\varphi_s\left(\begin{matrix}a_i\\ e_lb_j\end{matrix};\,x\right),\qquad 1\leqq l\leqq s,$$

$$x^{-1}(1-T_x)\,{}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,x\right)=\frac{(1-a_1)\cdots(1-a_r)}{(1-b_1)\cdots(1-b_s)}\,{}_r\varphi_s\left(\begin{matrix}qa_i\\ qb_j\end{matrix};\,x\right)$$

where

$$(2.10)\qquad \begin{aligned} e^ka_i&=\begin{cases}a_i & \text{if } i\neq k,\\ qa_k & \text{if } i=k,\end{cases}\\ e_lb_j&=\begin{cases}b_j & \text{if } j\neq l,\\ q^{-1}b_l & \text{if } j=l.\end{cases}\end{aligned}$$

Note that relations (2.9) imply the fundamental q-difference equation satisfied by the ${}_r\varphi_s$:

(2.11)

$$\{x(1-a_1T_x)\cdots(1-a_rT_x)-(1-T_x)(1-b_1q^{-1}T_x)\cdots(1-b_sq^{-1}T_x)\}\,{}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,x\right)=0.$$

Indeed, for $\beta_j\neq 0, -1, -2, \cdots$ the only solution of this equation which is analytic in x at $x=0$ is ${}_r\varphi_s(\begin{smallmatrix}a_i\\ b_j\end{smallmatrix};\,x)$.

Now we define the function ${}_r\Phi_s$ of $2(r+s)+1$ variables by

$$(2.12)\qquad {}_r\Phi_s(a_i,b_j;u_p)={}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,\frac{u_{r+1}\cdots u_{r+s+1}}{u_1\cdots u_r}\right)u_1^{-\alpha_1}\cdots u_r^{-\alpha_r}u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1}.$$

Let $\Delta_p^{\pm}$ be the q-difference operators

$$(2.13)\qquad \begin{aligned}\Delta_p^{+}&=u_p^{-1}(1-T_{u_p}),\\ \Delta_p^{-}&=u_p^{-1}(1-T_{u_p}^{-1}),\qquad 1\leqq p\leqq r+s+1.\end{aligned}$$

In terms of these operators, relations (2.9) take the simple form

$$
\begin{aligned}
&\Delta_k^-\,{}_r\Phi_s=(1-a_k)\,{}_r\Phi_s\begin{pmatrix}e^k a_i\\ b_j\end{pmatrix}, && 1\leqq k\leqq r,\\
(2.14)\qquad &\Delta_{r+l}^+\,{}_r\Phi_s=(1-b_lq^{-1})\,{}_r\Phi_s\begin{pmatrix}a_i\\ e_lb_j\end{pmatrix}, && 1\leqq l\leqq s,\\
&\Delta_{r+s+1}^+\,{}_r\Phi_s=\frac{(1-a_1)\cdots(1-a_r)}{(1-b_1)\cdots(1-b_s)}\,{}_r\Phi_s\begin{pmatrix}qa_i\\ qb_j\end{pmatrix}
\end{aligned}
$$

and (2.11) becomes the (canonical) partial q-difference equation

$$
(2.15)\qquad \left(\prod_{k=1}^{r}\Delta_p^- - \prod_{p=r+1}^{r+s+1}\Delta_p^+\right){}_r\Phi_s=0.
$$

Furthermore ${}_r\Phi_s$ satisfies the eigenvalue equations

$$
(2.16)\qquad
\begin{aligned}
&T_{r+s+1}^{-1}T_k^{-1}\,{}_r\Phi_s=q^{\alpha_k}\,{}_r\Phi_s, && 1\leqq k\leqq r,\\
&T_{r+s+1}^{-1}T_{r+l}\,{}_r\Phi_s=q^{\beta_l-1}\,{}_r\Phi_s, && 1\leqq l\leqq s.
\end{aligned}
$$

Indeed, ${}_r\Phi_s$ is characterized by (2.15), (2.16): It is (to within a constant multiple) the only solution of these equations analytic in the u_p at $u_{r+s+1}=0$.

We can regard an analytic solution $\Psi(u_p)$ of the canonical equation

$$
(2.17)\qquad (\Delta_1^-\cdots\Delta_r^- - \Delta_{r+1}^+\cdots\Delta_{r+s+1}^+)\Psi=0
$$

as a generating function for basic hypergeometric functions. Indeed, expanding Ψ as a power series

$$
(2.18)\qquad \Psi(u_p)=\sum_{\alpha_i,\beta_j} f_{\alpha_i\beta_j}(x)u_1^{-\alpha_1}\cdots u_r^{-\alpha_r}u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1}
$$

where $x=u_{r+1}\cdots u_{r+s+1}/u_1\cdots u_r$, we see that if Ψ is analytic at $x=0$ and if no nonzero term occurs with some $\beta_j=0,-1,-2,\cdots$ then always

$$
(2.19)\qquad f_{\alpha_i\beta_j}(x)=c_{\alpha_i\beta_j}\,{}_r\varphi_s\left(\begin{matrix}a_i\\ b_j\end{matrix};\,x\right)
$$

for some constants $c_{\alpha_i\beta_j}$. We shall typically compute such a generating function Ψ by characterizing it as a simultaneous eigenfunction of a set of $r+s$ commuting symmetry operators for (2.17). By a symmetry operator for the canonical equation we mean a linear operator L which maps any local analytic solution Ψ for (2.17) into another local analytic solution $L\Psi$. Clearly the dilation operators $T_{r+s+1}^{-1}T_k^{-1}$ $(1\leqq k\leqq r)$, $T_{r+s+1}^{-1}T_{r+l}$ $(1\leqq l\leqq s)$ are commuting symmetries, and the eigenvalue equations (2.16) characterize the basis solutions ${}_r\Phi_s$ in terms of these symmetries. Furthermore the q-difference operators Δ_i^- $(1\leqq i\leqq r)$ and Δ_{r+h}^+ $(1\leqq h\leqq s+1)$ are commuting symmetries. Note also that any permutation of the variables $\{u_i\colon 1\leqq i\leqq r\}$ is a symmetry of (2.17) as is any permutation of the variables $\{u_{r+h}\colon 1\leqq h\leqq s+1\}$. (For example, the transposition symmetry (u_{r+1}, u_{r+s+1}) implies that

$$
{}_r\varphi_s\left(\begin{matrix}a_ib_1^{-1}q\\ q^2b_1^{-1},\ b_jb_1^{-1}q\end{matrix};\,x\right)x^{1-\beta_1}
$$

is another solution of (2.11).)

The canonical equation (2.17) for q-hypergeometric functions is a clear analogue of the canonical equation for the hypergeometric functions ${}_rF_s$ [13]. Indeed the basis functions

$$(2.20)\qquad {}_r\mathscr{F}_a\left(\begin{matrix}\alpha_i\\ \beta_j\end{matrix}; u_p\right) = {}_rF_s\left(\begin{matrix}\alpha_i\\ \beta_j\end{matrix}; \frac{u_{r+1}\cdots u_{r+s+1}}{u_1\cdots u_r}\right) u_1^{-\alpha_1}\cdots u_r^{-\alpha_r} u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1}$$

satisfy the canonical equation

$$(2.21)\qquad (\partial_{u_1}\cdots\partial_{u_r} - (-1)^r\partial_{u_{r+1}}\cdots\partial_{u_{r+s+1}})\,{}_r\mathscr{F}_s = 0$$

and the eigenvalue equations

$$(2.22)\qquad \begin{aligned} &(-D_{r+s+1}-D_k)\,{}_r\mathscr{F}_s = \alpha_k\,{}_r\mathscr{F}_s, && 1\leqq k\leqq r,\\ &(-D_{r+s+1}+D_{r+l})\,{}_r\mathscr{F}_s = (\beta_l-1)\,{}_r\mathscr{F}_s, && 1\leqq l\leqq s,\end{aligned}$$

where $D_p = u_p\partial_{u_p}$. Furthermore ${}_r\mathscr{F}_s$ is the only solution of (2.21), (2.22) that is analytic in the u_p at $u_{r+s+1}=0$.

Our procedure applies to a family of q-analogues for the ${}_rF_s$. Let δ be a function with domain $\{1, 2, \cdots, r+s+1\}$ and range contained in the set $\{+, -\}$. The canonical equation

$$(2.23)\qquad (\Delta_1^{\delta(1)}\cdots\Delta_r^{\delta(r)} - \Delta_{r+1}^{\delta(r+1)}\cdots\Delta_{r+s+1}^{\delta(r+s+1)})\Psi = 0$$

and eigenvalue equations

$$(2.24)\qquad \begin{aligned} &T_{r+s+1}^{-1}T_k^{-1}\Psi = q^{\alpha_k}\Psi, && 1\leqq k\leqq r,\\ &T_{r+s+1}^{-1}T_{r+l}\Psi = q^{\beta_l-1}\Psi, && 1\leqq l\leqq s,\end{aligned}$$

have the unique (to within a constant multiple) solution

$$(2.25)\qquad {}_r\Phi_s^{\delta}\left(\begin{matrix}a_i\\ b_i\end{matrix}; u_p\right) = {}_r\varphi_s^{\delta}\left(\begin{matrix}a_i\\ b_j\end{matrix}; \frac{u_{r+1}\cdots u_{r+s+1}}{u_1\cdots u_r}\right) u_1^{-\alpha_1}\cdots u_r^{-\alpha_r} u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1}$$

where

$$(2.26)\qquad \begin{aligned} {}_r\varphi_s^{\delta}\left(\begin{matrix}a_i\\ b_j\end{matrix}; x\right) = \sum_{n=0}^{\infty} &\frac{(q^{\delta'(1)\alpha_1}; q^{\delta'(1)1})_n}{(q^{\delta(r+1)\beta_1}; q^{\delta(r+1)1})_n}\cdots\frac{(q^{\delta'(r)\alpha_r}; q^{\delta'(r)1})_n}{(q^{\delta(r+s)\beta_s}; q^{\delta(r+s)1})_n}\\ &\cdot\frac{x^n}{(q^{\delta(r+s+1)}; q^{\delta(r+s+1)1})_n}\end{aligned}$$

and

$$(2.27)\qquad \delta'(p) = \begin{cases} + & \text{if } \delta(p) = -,\\ - & \text{if } \delta(p) = +.\end{cases}$$

Each of these q-analogues of ${}_rF_s$ can be further treated by the methods presented in this paper.

Canonical equations for many-variable hypergeometric q-series can be derived almost as easily as for the one-variable case. Consider for example the Appell function F_2:

$$(2.28)\qquad F_2\left(\begin{matrix}\alpha, \beta, \beta'\\ \gamma, \gamma'\end{matrix}; x, y\right) = \sum_{n,m=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_m(\beta')_n x^m y^n}{(\gamma)_m(\gamma')_n m!n!}.$$

As shown in [11] the canonical differential equations are

$$(2.29)\qquad (\partial_{u_1}\partial_{u_2} - \partial_{u_3}\partial_{u_4})\mathscr{F}_2 = 0, \qquad (\partial_{u_1}\partial_{u_5} - \partial_{u_6}\partial_{u_7})\mathscr{F}_2 = 0$$

with eigenvalue equations

$$\begin{gathered} D_1+D_3+D_6\sim -\alpha, \qquad D_2+D_3\sim -\beta, \\ D_4-D_3\sim \gamma-1, \quad D_5+D_6\sim -\beta', \quad D_7-D_6\sim \gamma'-1. \end{gathered} \tag{2.30}$$

Here $A\sim\alpha$ stands for $A\mathscr{F}_2=\alpha\mathscr{F}_2$, and $D_i=u_i\partial_{u_i}$. Furthermore,

$$\mathscr{F}_2=F_2\begin{pmatrix} \alpha,\beta,\beta' & ; & \dfrac{u_3u_4}{u_1u_2}, \dfrac{u_6u_7}{u_1u_5} \\ \gamma,\gamma' & & \end{pmatrix} u_1^{-\alpha}u_2^{-\beta}u_5^{-\beta'}u_4^{\gamma-1}u_7^{\gamma'-1}.$$

Now consider the q-analogue

$$f_2\begin{pmatrix} a,b,b' & ;x,y \\ c,c' & \end{pmatrix}=\sum_{n,m=0}^{\infty}\frac{(a;q)_{m+n}(b;q)_m(b';q)_nx^my^n}{(c;q)_m(c';q)_n(q;q)_m(q;q)_n}$$

where $a=q^\alpha$, $b=q^\beta$, etc. The function

$$f_2=f_2\begin{pmatrix} a,b,b' & ; & \dfrac{u_3u_4}{u_1u_2}, \dfrac{u_6u_7}{u_1u_5} \\ c,c' & & \end{pmatrix} u_1^{-\alpha}u_2^{-\beta}u_5^{-\beta'}u_4^{\gamma-1}u_7^{\gamma'-1} \tag{2.31}$$

satisfies the recurrence relations

$$\begin{gathered} \Delta_1^-f_2=(1-a)f_2(aq), \qquad \Delta_2^-f_2=(1-b)f_2(bq), \\ \Delta_3^+f_2=\frac{(1-a)(1-b)}{(1-c)}f_2\begin{pmatrix} aq,bq \\ cq \end{pmatrix}, \\ \Delta_4^+f_2=(1-cq^{-1})f_2(cq^{-1}), \qquad \Delta_5^-f_2=(1-b')f_2(b'q), \\ \Delta_6^+f_2=\frac{(1-a)(1-b')}{(1-c')}f_2\begin{pmatrix} aq,b'q \\ c'q \end{pmatrix}, \\ \Delta_7^+f_2=(1-c'q^{-1})f_2(c'q^{-1}), \end{gathered} \tag{2.32}$$

hence the canonical equations

$$\Delta_1^-\Delta_2^--\Delta_3^+\Delta_4^+\sim 0, \qquad \Delta_1^-\Delta_5^--\Delta_6^+\Delta_7^+\sim 0. \tag{2.33}$$

(Here again $A\sim\chi$ signifies that f_2 is an eigenfunction of the operator A with eigenvalue χ.) Furthermore f_2 satisfies the dilation eigenvalue equations

$$\begin{gathered} T_1T_3T_6\sim a^{-1}, \quad T_2T_3\sim b^{-1}, \quad T_4T_3^{-1}\sim cq^{-1}, \\ T_5T_6\sim b'^{-1}, \quad T_7T_6^{-1}\sim c'q^{-1}, \end{gathered} \tag{2.34}$$

and for $\gamma, \gamma'\neq 0, -1, -2, \cdots$ the only solution of equations (2.33), (2.34) analytic in the u_i at $u_3=u_6=0$ is (2.31). The standard pair of q-difference equations for the function f_2

$$\begin{gathered} [(1-aT_xT_y)(1-bT_x)-x^{-1}(1-T_x)(1-cq^{-1}T_x)]f_2=0, \\ [(1-aT_xT_y)(1-b'T_y)-y^{-1}(1-T_y)(1-c'q^{-1}T_y)]f_2=0, \end{gathered}$$

is obtained directly from the canonical equations by setting $x=u_3u_4/u_1u_2$, $y=u_6u_7/u_1u_5$ and factoring out the remaining "ignorable" variables. Note the perfect correspondence between the differential equations (2.29), (2.30) for the Appell function and the q-difference equations (2.33), (2.34) for the q-analogue.

Just as in the single-variable case we can study a family of q-analogues for F_2, one for each function δ with domain $\{1, 2, \cdots, 7\}$ and range contained in $\{+, -\}$. The canonical equations are

$$\begin{aligned}&\Delta_1^{\delta(1)}\Delta_2^{\delta(2)} - \Delta_3^{\delta(3)}\Delta_4^{\delta(4)} \sim 0,\\ &\Delta_1^{\delta(1)}\Delta_5^{\delta(5)} - \Delta_6^{\delta(6)}\Delta_7^{\delta(7)} \sim 0\end{aligned} \tag{2.35}$$

and the corresponding q-series is

$$\begin{aligned}f_2^\delta\begin{pmatrix} a, b, b' \\ c, c' \end{pmatrix}; x, y\Big) &= \sum_{m,n=0}^{\infty} \frac{(q^{\delta'(1)\alpha}; q^{\delta'(1)1})_{m+n}(q^{\delta'(2)\beta}; q^{\delta'(2)1})_m}{(q^{\delta(r)\gamma}; q^{\delta(r)1})_m(q^{\delta(7)\gamma'}; q^{\delta(7)1})_n}\\ &\quad\cdot \frac{(q^{\delta'(5)\beta'}; q^{\delta'(5)1})_n x^m y^n}{(q^{\delta(3)}; q^{\delta(3)1})_m(q^{\delta(6)}; {}^{\delta(6)1})_n}.\end{aligned} \tag{2.36}$$

Similar computations can be performed for any two-variable (or many-variable) q-hypergeometric series g. Corresponding to each parameter $a = q^\alpha$ such that the symbol $(a; q)_{Am+Bn}$ appears in the numerator of the expansion

$$g = \sum_{m,n} \frac{(a; q)_{Am+Bn} \cdots}{(c; q)_{Cm+Dn} \cdots} \frac{x^m y^m}{(q; q)_m(q; q)_n}$$

there is a "raising operator" $E^\alpha = \Delta^-$ associated with the recurrence relation

$$(1 - aT_x^A T_y^B)g = (1 - a)g(aq).$$

Similarly, for each denominator parameter c we can construct a "lowering operator" $E_\gamma = \Delta^+$ associated with

$$(1 - cq^{-1}T_x^C T_y^D)g = (1 - cq^{-1})g(cq^{-1}).$$

Application of $x^{-1}(1 - T_x)$, corresponding to Δ^+, takes each numerator parameter a to aq^A and each denominator parameter c to cq^C, whereas application of $y^{-1}(1 - T_y)$, corresponding to Δ^+, takes each numerator parameter a to aq^B and each denominator parameter c to cq^D.

Hrabowski [9] has discussed the general procedures for associating a system of canonical differential equations and eigenvalue equations with a given hypergeometric series and, conversely, for associating one or more hypergeometric series with a given system of canonical differential equations and eigenvalue equations. His analysis applies, with minor modifications, to q-hypergeometric series as we will discuss in a subsequent paper. The principal distinction is that there are two types of q-difference operators $\Delta_u^\pm$ and only one type of differential operator ∂_u.

3. Embeddings and augmentations. If the canonical equations of a q-hypergeometric series can be identified with a subset of the canonical equations of a second q-hypergeometric series, then by a suitable restriction of coordinates we can regard the second series to be a generating function for the first series. As an illustrative example we consider the canonical equation

$$\Delta_1^-\Delta_2^- - \Delta_3^+\Delta_4^+ \sim 0 \tag{3.1}$$

for the basic hypergeometric function

$$\begin{aligned}&{}_2\Phi_1 = {}_2\varphi_1\begin{pmatrix} a, b \\ c \end{pmatrix}; \frac{u_3u_4}{u_1u_2}\Big)u_1^{-\alpha}u_2^{-\beta}u_3^{\gamma-1},\\ &T_4^{-1}T_1^{-1} \sim q^\alpha, \qquad T_4^{-1}T_2^{-1} \sim q^\beta, \qquad T_4^{-1}T_3 \sim q^{\gamma-1},\end{aligned} \tag{3.2}$$

and the canonical equations (2.33) and eigenvalue equations (2.34) for the q-Appell function f_2. Identifying the variables $u_1, \cdots, u_4$ for ${}_2\Phi_1$ with the variables $u_1, \cdots, u_4$ for f_2 we see from (2.33) and (3.1) that for any choice of u_5, u_6, u_7 the function $f_2(u_p)$ can be regarded as a solution of the canonical equation (3.1) for ${}_2\Phi_1$. For uniqueness we require $u_5 = u_6 = u_7 = 1$ and obtain the solution

$$f_2 = f_2\begin{pmatrix} a, b, b' \\ c, c' \end{pmatrix}; \frac{u_3u_4}{u_1u_2}, \frac{1}{u_1}\Big) u_1^{-\alpha}u_2^{-\beta}u_4^{\gamma-1} \tag{3.3}$$

of (3.1). Our approach is to characterize the generating function (3.3) in terms of symmetry operators for (3.1). However the remaining canonical equation for f_2 and the eigenvalue equations (2.34) involve the variables u_5, u_6, u_7. We need to evaluate the operators $\Delta_5^-, \Delta_6^+, \Delta_7^+$ applied to f_2 for $u_5 = u_6 = u_7 = 1$, in terms of operators acting only on functions of the variables $u_1, \cdots, u_4$. From (2.34) we find

$$T_6 \sim a^{-1}T_1^{-1}T_3^{-1}, \qquad T_5^{-1} \sim b'a^{-1}T_1^{-1}T_3^{-1}, \qquad T_7 \sim c'q^{-1}a^{-1}T_1^{-1}T_3^{-1}.$$

Thus the solution (3.3) is characterized by the equations

$$\begin{gathered} T_2T_3 \sim b^{-1}, \qquad T_4T_3^{-1} \sim cq^{-1}, \\ \Delta_1^-\left(1 - \frac{b'}{a}T_1^{-1}T_3^{-1}\right) - \left(1 - \frac{1}{a}T_1^{-1}T_3^{-1}\right)\left(1 - \frac{c'}{qa}T_1^{-1}T_3^{-1}\right) \sim 0. \end{gathered} \tag{3.4}$$

Note that the operators T_2T_3, $T_4T_3^{-1}$, Δ_1^- and $T_1^{-1}T_3^{-1}$ are all symmetries of $\Delta_1^-\Delta_2^- - \Delta_3^+\Delta_4^+$, so that we have characterized the solution (3.3) of (3.1) in terms of a set of (mixed) eigenvalue equations for symmetries of (3.1). (It is not always the case that the generating function of the restricted canonical system obtained through this process is characterized in terms of symmetries of the restricted system. An example is the restriction of the canonical system for the Appell function F_1 to the wave equation [11]. However, in this case and in all other such examples known to the authors appropriate functional linear combinations of the mixed eigenvalue equations can be expressed in terms of symmetry operators and the resulting system still uniquely characterizes the generating function.) This is the process of *embedding.*

The process inverse to embedding is augmentation. Here we are given a canonical system of q-difference equations and a characterization of a generating function for this system by a set of mixed eigenvalue equations expressed in terms of symmetry operators for the canonical system. Our aim is to establish simple rules for determination of the generating function as an explicit q-hypergeometric series by recasting the defining equations as a canonical system with dilation eigenvalue equations in a greater number of variables than the original problem.

To see how this procedure works, consider the generating function, characterized as the function analytic in $u_1, \cdots, u_4$ at $u_3 = 0$ and satisfying the canonical equation (3.1) and the mixed eigenvalue equations (3.4). Since the last equation in (3.4) is neither a canonical equation or a dilation eigenvalue equation for (3.1), we cannot determine the power series expression for the generating function by inspection. However, we can replace the expressions

$$1 - b'a^{-1}T_1^{-1}T_3^{-1}, \qquad 1 - a^{-1}T_1^{-1}T_3^{-1}, \qquad 1 - c'q^{-1}a^{-1}T_1^{-1}T_3^{-1}$$

by Δ_5^-, Δ_6^+, Δ_7^+, respectively, for $u_5 = u_6 = u_7 = 1$ where $T_5^{-1} \sim b'a^{-1}T_1^{-1}T_3^{-1}$, $T_6 \sim a^{-1}T_1^{-1}T_3^{-1}$ and $T_7 = c'q^{-1}a^{-1}T_1^{-1}T_3^{-1}$. Then for general u_p the defining equations of the generating function take the canonical form (2.33), (2.34) with the unique solution (2.31), analytic at $u_3 = u_6 = 0$. Setting $u_5 = u_6 = u_7 = 1$ we obtain the generating function.

(Note that the choice Δ_5^-, Δ_6^+, Δ_7^+ is unique. If we had taken Δ_5^+ for example, we would have obtained the condition $T_5T_1T_3 \sim b'a^{-1}$, but $T_5T_1T_3$ is *not* a symmetry of the canonical equations (2.33).) We also note that q-analogues of the Appell function F_3 and the Horn function $\mathcal{H}_2$ correspond to these same canonical equations but have different analyticity properties [11].

The following sections contain several more examples of augmentation.

4. Generating functions for ${}_r\varphi_s$. Here we will present several examples showing how generating functions for the ${}_r\varphi_s$, (2.1) are associated with the canonical equation

$$(4.1) \qquad (\Delta_1^- \cdots \Delta_r^- - \Delta_{r+1}^+ \cdots \Delta_{r+s+1}^+)\Psi = 0.$$

Recall that

$${}_r\Phi_s = {}_r\varphi_s\left(\begin{matrix} a_i \\ b_j \end{matrix}; \frac{u_{r+1}\cdots u_{r+s+1}}{u_1\cdots u_r}\right) u_1^{-\alpha_1}\cdots u_r^{-\alpha_r} u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1}$$

is a solution of (4.1) and that Δ_i^-, $(1 \leqq i \leqq r)$, Δ_{r+k}^+, $(1 \leqq k \leqq s)$, $T_{r+s+1}T_i$ and $T_{r+s+1}^{-1}T_k$ are symmetries of this equation.

There appears to be no convenient general q-analogue of the local Lie theory which permits us to compute Lie group symmetries of differential equations from Lie algebra symmetries through the process of exponentiation. However, in particular cases the analogy is successful. Consider the q-exponential

$$(4.2) \qquad e_q(x) = \sum_{n=0}^{\infty} \frac{x^n}{(q;q)_n} = \frac{1}{(x;q)_\infty}, \qquad |x| < 1$$

satisfying $\Delta_x^+ e_q = e_q$ [15, p. 92]. In a formal sense at least, the operator $e_q(\lambda\Delta_1^-)$, $\lambda \in \mathbb{C}$, is a symmetry of (4.1). Applying this operator to a basis solution ${}_r\Phi_s$ and making use of (2.14), (4.2) we obtain

$$(4.3) \qquad e_q(\lambda\Delta_1^-)\,{}_r\Phi_s(a_1) = \sum_{n=0}^{\infty} \frac{\lambda^n (a_1;q)_n}{(q;q)_n}\,{}_r\varphi_s(a_1q^n).$$

To compute the left-hand side of (4.3) we utilize Heines's (q-binomial) theorem [15, p. 92], [2],

$$\sum_{m=0}^{\infty} \frac{(a;q)_m}{(q;q)_m} t^m = \frac{(at;q)_\infty}{(t;q)_\infty}$$

to derive

$$(4.4) \qquad e_q(\lambda\Delta_x^-)x^n = x^n \frac{(\lambda q^{-n}/x;q)_\infty}{(\lambda/x;q)_\infty} = \frac{x^n}{(\lambda/x,q)_{-n}}.$$

From (4.4), (2.1) and (2.12) we find

$$(4.5) \qquad \begin{aligned} e_q(\lambda\Delta_1^-)\,{}_r\Phi_s(a_1) &= \frac{(\lambda/a_1u_1;q)_\infty}{(\lambda/u_1;q)_\infty}\,{}_r\varphi_{s+1}\left(\begin{matrix} a_i \\ b_j \end{matrix}; \lambda/a_1u_1; \frac{u_{r+1}\cdots u_{r+s+1}}{u_1\cdots u_r}\right) \\ &\quad \cdot u_1^{-\alpha_1}\cdots u_r^{-\alpha_r} u_{r+1}^{\beta_1-1}\cdots u_{r+s}^{\beta_s-1} \end{aligned}$$

so that

$$(4.6) \qquad \frac{(x/a_1;q)_\infty}{(x;q)_\infty}\,{}_r\varphi_{s+1}\left(\begin{matrix} a_i \\ b_j \end{matrix}; x/a_1; y\right) = \sum_{n=0}^{\infty} \frac{(a_1^{-1};q)_n}{(q;q)_n} x^n\,{}_r\varphi_s\left(\begin{matrix} a_1q^n, a_2, \cdots, a_r \\ b_j \end{matrix}; y\right).$$

Another way to understand this result is to note that the right-hand side of (4.3) is an eigenfunction of the operator $T_{r+s+1}T_1-\lambda q^{-1}\Delta_1^- T_{r+s+1}T_1$ with eigenvalue a_1^{-1}. The method of augmentation can then be employed to derive (4.5). Still another point of view is that since ${}_r\Phi_s(a_1)$ is an eigenfunction of the symmetry operator $T_{r+s+1}T_1$ with eigenvalue a_1^{-1} then $e_q(\lambda\Delta_1^-){}_r\Phi_s(a_1)$ must be an eigenfunction of the formal symmetry operator

$$e_q(\lambda\Delta_1^-)T_{r+s+1}T_1E_q(-\lambda\Delta_1^-) \tag{4.7}$$

with the same eigenvalue. Here [15, p. 93]

$$\begin{aligned} E_q(x)&=\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{(q;q)_n}x^n=(-x;q)_\infty,\\ \Delta_x^-E_q&=-q^{-1}E_q,\qquad e_q(x)E_q(-x)=1.\end{aligned} \tag{4.8}$$

Let X, Y be linear operators.

LEMMA.

$$f(\lambda)\equiv e_q(\lambda X)YE_q(-\lambda X)=\sum_{n=0}^{\infty}\frac{\lambda^n}{(q;q)_n}[X,Y]_n \tag{4.9}$$

where

$$\begin{aligned} &[X,Y]_0=Y,\qquad [X,Y]_1=XY-YX,\\ &[X,Y]_{n+1}=X[X,Y]_n-q^n[X,Y]_nX,\qquad n=1,2,\cdots.\end{aligned} \tag{4.10}$$

Proof. This result is equivalent to the identity $\Delta_\lambda^+ f(\lambda)=Xf(\lambda)-T_\lambda f(\lambda)X$ and the identity can be verified by formal power series expansion in the variable λ. (The authors learned of this result from Mourad Ismail and Dennis Stanton.)

For $X=\Delta_1^-$, $Y=T_{r+s+1}T_1$ it is easily verified that $[X,Y]_1=(1-q^{-1})\Delta_1^-T_{r+s+1}T_1$, $[X,Y]_2=0$ so, by the lemma:

$$e_q(\lambda\Delta_1^-)T_{r+s+1}T_1E_q(-\lambda\Delta_1^-)=T_{r+s+1}T_1-q^{-1}\lambda\Delta_1^-T_{r+s+1}T_1.$$

Using the q-binomial theorem we can verify that

$$E_q(-\lambda\Delta_x^+)x^n=x^n\frac{(\lambda/x;q)_\infty}{(\lambda q^n/x;q)_\infty}. \tag{4.11}$$

From (2.14) we have

$$E_q(-\lambda\Delta_{r+1}^+){}_r\Phi_s(b_1)=\sum_{n=0}^{\infty}\frac{(b\lambda/q)^n(b^{-1}q;q)_n}{(q;q)_n}{}_r\Phi_s(b_1q^{-n}),$$

$$\begin{aligned} E_q(-\lambda\Delta_{r+s+1}^+){}_r\Phi_s\begin{pmatrix}a_i\\ b_j\end{pmatrix}=\sum_{n=0}^{\infty}&\frac{(-\lambda)^nq^{n(n-1)/2}}{(q;q)_n}\\ &\cdot\frac{(a_1;q)_n\cdots(a_r;q)_n}{(b_1;q)_n\cdots(b_s;q)_n}{}_r\Phi_s\begin{pmatrix}a_iq^n\\ b_jq^n\end{pmatrix}.\end{aligned}$$

Applying (4.11) we obtain the generating functions

$$\begin{aligned} \frac{(qx/b;q)_\infty}{(x;q)_\infty}{}_{r+1}\varphi_s\begin{pmatrix}a_i,x\\ b_j\end{pmatrix};y\Bigg)&=\sum_{n=0}^{\infty}\frac{(q/b_1;q)_n}{(q;q)_n}{}_r\varphi_s\begin{pmatrix}a_i\\ b_1q^{-n},b_2\cdots b_s\end{pmatrix};y\Bigg)x^n,\\ {}_{r+1}\varphi_s\begin{pmatrix}a_i,x\\ b_j\end{pmatrix};y\Bigg)&=\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{(q;q)_n}\frac{(a_1;q)_n\cdots(a_r;q)_n}{(b_1;q)_n\cdots(b_s;q)_n}{}_r\varphi_s\begin{pmatrix}a_iq^n\\ b_jq^n\end{pmatrix};y\Bigg)(-xy)^n.\end{aligned} \tag{4.12}$$

Now we consider some examples where the generating functions are directly characterized in terms of symmetry operators for the canonical equation (4.1). Let A, B, C be nonnegative integers with $B \geqq 1$, $A+B=r$, $C+B=s$. We look for an eigenfunction of (4.1) characterized by the following conditions:

$$\begin{array}{ll}
\multicolumn{2}{l}{\Delta_1^- \Delta_2^- \cdots \Delta_A^- - \Delta_{r+1}^+ \Delta_{r+2}^+ \cdots \Delta_{r+C}^+ \Delta_{r+s+1}^+ \sim 0,} \\
\multicolumn{2}{l}{\Delta_{A+1}^- \Delta_{A+2}^- \cdots \Delta_r^- - \Delta_{r+C+1}^+ \Delta_{r+C+2}^+ \cdots \Delta_{r+s}^+ \sim 0,} \\
T_1^{-1} T_{r+s+1}^{-1} \sim a_1, & T_{r+1} T_{r+s+1}^{-1} \sim c_1 q^{-1}, \\
\vdots & \vdots \\
T_A^{-1} T_{r+s+1}^{-1} \sim a_A, & T_{r+C} T_{r+s+1}^{-1} \sim c_C q^{-1}, \\
T_{A+1} T_r^{-1} \sim b_1 q^{-1}, & T_{r+C+1}^{-1} T_r^{-1} \sim d_1, \\
\vdots & \vdots \\
T_{A+B-1} T_r^{-1} \sim b_{B-1} q^{-1}, & T_{r+C+B}^{-1} T_r^{-1} \sim d_B.
\end{array} \tag{4.13}$$

Choosing u_r and u_{r+s+1} as the distinguished variables such that the generating function is analytic at $u_r = u_{r+s+1} = 0$, we see that equations (4.13) are in canonical form with solution

$$\begin{aligned}
&{}_A\varphi_C\left(\begin{matrix} a_1, \cdots, a_A \\ c_1, \cdots, c_C \end{matrix}; \frac{u_{r+1} \cdots u_{r+C}}{u_1 \cdots u_A} u_{r+s+1}\right) u_1^{-\alpha_1} \cdots u_A^{-\alpha_A} u_{r+1}^{\gamma_1 - 1} \cdots u_{r+C}^{\gamma_C - 1} \\
&\quad \cdot {}_B\varphi_{B-1}\left(\begin{matrix} d_1, \cdots, d_B \\ b_1, \cdots, b_{B-1} \end{matrix}; \frac{q b_1 \cdots b_{B-1} u_{A+1} \cdots u_{A+B-1}}{d_1 \cdots d_B u_{r+C+1} \cdots u_{r+s}} u_r\right) \\
&\quad \cdot u_{A+1}^{\beta_1 - 1} \cdots u_{A+B-1}^{\beta_{B-1}} u_{r+C+1}^{-\delta_1} \cdots u_{r+s}^{-\delta_B} \\
&= \sum_{n=0}^{\infty} X_n \, {}_{A+B}\varphi_{C+B}\left(\begin{matrix} a_1, \cdots, a_A; q^{-n}, q^{1-n}/b_1, \cdots, q^{1-n}/b_{B-1} \\ c_1, \cdots, c_C; q^{1-n}/d_1, \cdots, q^{1-n}/d_B \end{matrix}; \frac{u_{r+1} \cdots u_{r+s+1}}{u_1 \cdots u_r}\right) \\
&\quad \cdot u_1^{-\alpha_1} \cdots u_A^{-\alpha_A} u_{A+1}^{\beta_1 + n - 1} \cdots u_{A+B-1}^{\beta_{B-1} + n - 1} u_r^n u_{r+1}^{\gamma_1 - 1} \cdots u_{r+C}^{\gamma_C - 1} u_{r+C+1}^{-\delta_1 - n} \cdots u_{r+s}^{-\delta_B - n}.
\end{aligned}$$

Setting $u_{r+s+1} = 0$, we find that

$$X_n = \left(\frac{q b_1 \cdots b_{B-1}}{d_1 \cdots d_B}\right)^n \frac{(d_1; q)_n \cdots (d_B; q)_n}{(b_1; q)_n \cdots (b_{B-1}; q)_n (q; q)_n}$$

and the generating function simplifies to

$$\begin{aligned}
&{}_A\varphi_C\left(\begin{matrix} a_1, \cdots, a_A \\ c_1, \cdots, c_C \end{matrix}; \frac{d_1 \cdots d_B}{b_1 \cdots b_{B-1}}; tz\right) {}_B\varphi_{B-1}\left(\begin{matrix} d_1, \cdots, d_B \\ b_1, \cdots, b_{B-1} \end{matrix}; t\right) \\
&= \sum_{n=0}^{\infty} \frac{(d_1; q)_n \cdots (d_B; q)_n}{(b_1; q)_n \cdots (b_{B-1}; q)_n} \\
&\quad \cdot {}_{A+B}\varphi_{C+B}\left(\begin{matrix} a_1, \cdots, a_A; q^{-n}, q^{1-n}/b_1, \cdots, q^{1-n}/b_{B-1} \\ c_1, \cdots, c_C; q^{1-n}/d_1, \cdots, q^{1-n}/d_B \end{matrix}; qz\right) \frac{t^n}{(q; q)_n}.
\end{aligned} \tag{4.14}$$

Finally we derive a generating function for the q-series

$$
\begin{aligned}
&{}_{p+1}\hat{\varphi}_{p+r}\left(\begin{matrix} a_1, \cdots, a_{p+1} \\ b_1, \cdots, b_{p+r} \end{matrix}; q^{-1}, x\right) \\
&= \sum_{n=0}^{\infty} \frac{(a_1; q^{-1})_n \cdots (a_{p+1}; q^{-1})_n}{(b_1; q^{-1})_n \cdots (b_{p+r}; q^{-1})_n} \frac{x^n}{(q^{-1}; q^{-1})_n} \\
&= \sum_{n=0}^{\infty} \frac{(a_1^{-1}; q)_n \cdots (a_{p+1}^{-1}; q)_n (-1)^{rn} q^{rn(n-1)/2}}{(b_1^{-1}; q)_n \cdots (b_{p+r}^{-1}; q)_n (q; q)_n} \left(\frac{x a_1 \cdots a_{p+1}}{b_1 \cdots b_{p+r} q}\right)^n.
\end{aligned}
\tag{4.15}
$$

Consider the equations

$$\Delta_1^+ \cdots \Delta_{p+1}^+ - \Delta_{p+2}^- \cdots \Delta_{2p+r+2}^- \sim 0, \tag{4.16}$$

$$
\begin{aligned}
&T_1^{-1} T_{2p+r+2}^{-1} \sim q^{\alpha_i} = a_i^{-1}, \qquad 1 \leqq i \leqq p+1, \\
&T_j T_{2p+r+2}^{-1} \sim q^{\beta_j - 1} = b_j^{-1} q^{-1}, \qquad p+2 \leqq j \leqq 2p+r+1,
\end{aligned}
\tag{4.17}
$$

in canonical form. The solution of these equations, analytic at $u_{2p+r+2} = 0$ is:

$$
\begin{aligned}
&{}_{p+1}\hat{\varphi}_{p+r}\left(\begin{matrix} a_1, \cdots, a_{p+1} \\ b_1, \cdots, b_{p+r} \end{matrix}; q^{-1}, \frac{u_{p+2} \cdots u_{2p+r+2}}{u_1 \cdots u_{p+1}}\right) \\
&\cdot u_1^{-\alpha_1} \cdots u_{p+1}^{-\alpha_{p+1}} u_{p+2}^{\beta_1 - 1} \cdots u_{2p+r+1}^{\beta_{p+r} - 1}.
\end{aligned}
\tag{4.18}
$$

We search for a generating function satisfying (4.16) and the following conditions:

$$
\begin{aligned}
&\text{(a)} \quad \Delta_1^+ + c T_1 T_{2p+r+2} \sim 1, \\
&\text{(b)} \quad T_1^{-1} T_{2p+r+2}^{-1} \sim a_i^{-1}, \qquad 2 \leqq i \leqq p+1, \\
&\qquad\quad T_j T_{2p+r+2}^{-1} \sim b_j^{-1} q^{-1}, \qquad p+2 \leqq j \leqq 2p+r+1.
\end{aligned}
\tag{4.19}
$$

Introducing a new variable u_{2p+r+3} and conditions

$$T_1 T_{2p+r+2} T_{2p+r+3} \sim c^{-1}, \qquad \Delta_1^+ - \Delta_{2p+r+3}^- \sim 0 \tag{4.20}$$

we see that (4.20) reduces to (4.19) (a) when $u_{2p+r+3} = 1$ and conditions (4.16), (4.19) (b), (4.20) are in canonical form. It is straightforward to write down the generating function analytic at $u_1 = u_{2p+r+3} = 0$ and, with simplification, to obtain the identity

$$
\begin{aligned}
&\frac{(ct; q)_\infty}{(t; q)_\infty} {}_{p+1}\Phi_{p+r+1}\left[\begin{matrix} c, a_2^{-1}, \cdots, a_{p+1}^{-1} \\ ct, b_1^{-1}, \cdots, b_{p+r}^{-1} \end{matrix}; q, xt\right] \\
&= \sum_{k=0}^{\infty} \frac{(c; q)_k}{(q; q)_k} {}_{p+1}\Phi_{p+r}\left[\begin{matrix} q^{-n}, a_2^{-1}, \cdots, a_{p+1}^{-1} \\ b_1^{-1}, \cdots, b_{p+r}^{-1} \end{matrix}; q, xq^n\right] t^k
\end{aligned}
\tag{4.21}
$$

where

$$
\begin{aligned}
&{}_{p+1}\Phi_{p+r}\left[\begin{matrix} a_1, \cdots, a_{p+1} \\ b_1, \cdots, b_{p+r} \end{matrix}; q, x\right] \\
&= \sum_{n=0}^{\infty} \frac{(a_1; q)_n \cdots (a_{p+1}; q)_n (-1)^r q^{rn(n-1)/2}}{(b_1; q)_n \cdots (b_{p+r}; q)_n} \frac{x^n}{(q; q)_n}.
\end{aligned}
\tag{4.22}
$$

The generating functions derived above are not "deep." Indeed each can be proven by equating coefficients of powers of appropriate variables and using the q-binomial theorem. Furthermore, more general generating functions hold when some of the q-shifted factorials are replaced by arbitrary sequences; see [6]. Our point is that generating functions in their totality can be classified and derived using symmetry methods. In the following section we consider cases where the q-series have a richer symmetry structure and the generating functions are more interesting.

5. Generating functions for ${}_2\varphi_1$. The canonical equations for the q-hypergeometric functions ${}_2\varphi_1$ and ${}_1\varphi_1$ admit certain simple symmetry operators which do not extend to symmetries of the equations for general ${}_r\varphi_s$. This is closely related to the fact that ${}_2\varphi_1$ and ${}_1\varphi_1$ obey q-difference recurrence relations not shared by general ${}_r\varphi_s$. We shall examine the case ${}_2\varphi_1$ in some detail. Here the canonical equation is

$$Q \equiv \Delta_1^- \Delta_2^- - \Delta_3^+ \Delta_4^+ \sim 0 \tag{5.1}$$

and the eigenvalue equations for the basis

$${}_2\Phi_1 = {}_2\varphi_1\begin{pmatrix} a, b & ; & \dfrac{u_3 u_4}{u_1 u_2} \\ c & & \end{pmatrix} u_1^{-\alpha} u_2^{-\beta} u_3^{\gamma-1}$$

are

$$T_4^{-1} T_1^{-1} \sim a, \qquad T_4^{-1} T_2^{-1} \sim b, \qquad T_4^{-1} T_3 \sim cq^{-1}. \tag{5.2}$$

In addition to the dilation operators $T_4 T_1$, $T_4 T_2$, $T_4 T_3^{-1}$, their products and inverses, we have as symmetries of (5.1) the operators

$$\begin{gathered}
E^\alpha = \Delta_1^-, \quad E^\beta = \Delta_2^-, \quad E_\gamma = \Delta_3^+, \quad E^{\alpha\beta\gamma} = \Delta_4^+, \\
E_\alpha = -q^{-1} u_3 u_4 T_1^{-2} T_4^{-2} \Delta_2^- + u_1 u_3 T_1^{-1} T_4^{-1} \Delta_3^+ - q u_1 T_3 T_4^{-1} + u_1 T_1^{-1} T_3 T_4^{-2}, \\
E_\beta = -u_3 u_4 q^{-1} T_2^{-2} T_4^{-2} \Delta_1^- + u_2 u_3 T_2^{-1} T_4^{-1} \Delta_3^+ - q u_2 T_3 T_4^{-1} + u_2 T_2^{-1} T_3 T_4^{-2}, \\
E^\gamma = u_3\left[q T_3 T_4^{-2} - T_1^{-1} T_2^{-1} T_4^{-2} + q^2 \frac{u_1 u_2}{u_3} T_3 \Delta_4^- T_4^{-1} \right], \\
E_{\alpha\beta\gamma} = -u_1 u_2 T_4^{-1} \Delta_3^+ + q^{-1} u_4 T_4^{-1} - q^{-2} u_4 T_1^{-1} T_2^{-1} T_4^{-2}, \\
E^{\alpha\gamma} = u_3 T_4^{-1} \Delta_1^- + q u_2 \Delta_4^-, \qquad E^{\beta\gamma} = u_3 T_4^{-1} \Delta_2^- + q u_1 \Delta_4^-, \\
E_{\alpha\gamma} = -u_1 \Delta_3^+ + q^{-1} u_4 T_1^{-1} T_4^{-1} \Delta_2^-, \qquad E_{\beta\gamma} = -u_2 \Delta_3^+ + q^{-1} u_4 T_2^{-1} T_4^{-1} \Delta_1^-.
\end{gathered} \tag{5.3}$$

These symmetry operators correspond to recurrence relations for the ${}_2\varphi_1$ since, when applied to the standard basis ${}_2\Phi_1\binom{a,b}{c}$, they yield

$$\begin{gathered}
E^\alpha\, {}_2\Phi_1 = (1-a)\,{}_2\Phi_1(aq), \qquad E_\alpha\, {}_2\Phi_1 = (a-c)\,{}_2\Phi_1(aq^{-1}), \\
E^\beta\, {}_2\Phi_1 = (1-b)\,{}_2\Phi_1(bq), \qquad E_\beta\, {}_2\Phi_1 = (b-c)\,{}_2\Phi_1(bq^{-1}), \\
E^\gamma\, {}_2\Phi_1 = \frac{(c-a)(c-b)}{c-1}\,{}_2\Phi_1(cq), \qquad E_\gamma\, {}_2\Phi_1 = \left(1-\frac{c}{q}\right){}_2\Phi_1(cq^{-1}), \\
E^{\alpha\beta\gamma}\, {}_2\Phi_1 = \frac{(1-a)(1-b)}{1-c}\,{}_2\Phi_1\begin{pmatrix} aq, bq \\ cq \end{pmatrix}, \qquad E_{\alpha\beta\gamma}\, {}_2\Phi_1 = \left(\frac{c}{q}-1\right){}_2\Phi_1\begin{pmatrix} aq^{-1}, bq^{-1} \\ cq^{-1} \end{pmatrix}, \\
E^{\alpha\gamma}\, {}_2\Phi_1 = \frac{(b-c)(1-a)}{1-c}\,{}_2\Phi_1\begin{pmatrix} aq \\ cq \end{pmatrix}, \qquad E_{\alpha\gamma}\, {}_2\Phi_1 = \left(1-\frac{c}{q}\right){}_2\Phi_1\begin{pmatrix} aq^{-1} \\ cq^{-1} \end{pmatrix}, \\
E^{\beta\gamma}\, {}_2\Phi_1 = \frac{(a-c)(1-b)}{1-c}\,{}_2\Phi_1\begin{pmatrix} bq \\ cq \end{pmatrix}, \qquad E_{\beta\gamma}\, {}_2\Phi_1 = \left(1-\frac{c}{q}\right){}_2\Phi_1\begin{pmatrix} bq^{-1} \\ cq^{-1} \end{pmatrix}.
\end{gathered} \tag{5.4}$$

There appear to be no simple standardized expressions for these symmetries as q-difference operators applied to the null space of Q. Indeed one can always multiply each such symmetry by a dilation symmetry and obtain a recurrence relation (5.4) equivalent to the original relation. Second, given any q-difference operator D we can add DQ to any symmetry operator, since DQ acts as the zero operator on the null space of Q. One can use the above modifications to simplify considerably some of the expressions for the operators (5.3), but at the expense of complicating relations (5.4).

From the raising and lowering operators E we can form the following equations, each equivalent to the canonical equation (5.1):

$$\begin{gathered}
E^{\alpha}E_{\alpha}+T_4^{-1}(qT_3-T_1^{-1})(1-q^{-1}T_4^{-1}T_1^{-1})\sim 0,\\
E^{\beta}E_{\beta}+T_4^{-1}(qT_3-T_2^{-1})(1-q^{-1}T_4^{-1}T_2^{-1})\sim 0,\\
E^{\gamma}E_{\gamma}+T_4^{-2}(T_3-T_1^{-1})(T_3-T_2^{-1})\sim 0,\\
E^{\alpha\beta\gamma}E_{\alpha\beta\gamma}+(1-q^{-1}T_4^{-1}T_1^{-1})(1-q^{-1}T_4^{-1}T_2^{-1})\sim 0,\\
E^{\alpha\gamma}E_{\alpha\gamma}-T_4^{-1}(T_2^{-1}-T_3^{-1})(1-q^{-1}T_4^{-1}T_1^{-1})\sim 0,\\
E^{\beta\gamma}E_{\beta\gamma}-T_4^{-1}(T_1^{-1}-T_3^{-1})(1-q^{-1}T_4^{-1}T_2^{-1})\sim 0.
\end{gathered} \tag{5.5}$$

The operators E and the dilation symmetries T_4T_1, T_4T_2, $T_4T_3^{-1}$ form a q-analogue of the 15-dimensional conformal symmetry algebra of the wave equation in four-dimensional space-time. Although these q-symmetry operators do not generate a finite-dimensional Lie algebra under operator commutation they still permit us to construct the invariants (5.5).

The method of augmentation can be used to obtain explicit expressions for many generating functions characterized by E symmetry operators. For example, while the conditions

$$E_{\alpha}\sim -c,\quad T_2T_4\sim b^{-1},\quad T_3T_4^{-1}\sim cq^{-1} \tag{5.6}$$

would be difficult to solve directly, due to the complicated expression for E_{α}, we note that the first of these conditions implies $E^{\alpha}E_{\alpha}\sim -cE^{\alpha}$ so from the first expression (5.5) for $E^{\alpha}E_{\alpha}$

$$\Delta_1^- -(1-c^{-1}T_4^{-1}T_1^{-1})(1-q^{-1}T_4^{-1}T_1^{-1})\sim 0.$$

Setting $T_5\sim c^{-1}T_4^{-1}T_1^{-1}$, $T_6\sim q^{-1}T_4^{-1}T_1^{-1}$, we see that the desired generating functions are the restrictions to $u_5=u_6=1$ of certain solutions of the canonical system

$$\begin{gathered}
\Delta_1^-\Delta_2^- -\Delta_3^+\Delta_4^+\sim 0,\qquad \Delta_1^- -\Delta_5^+\Delta_6^+\sim 0,\\
T_2T_4\sim b^{-1},\qquad T_3T_4^{-1}\sim cq^{-1},\\
T_1T_4T_5\sim c^{-1},\qquad T_1T_4T_6\sim q^{-1}.
\end{gathered} \tag{5.7}$$

We can immediately write down a series solution:

$$\Psi=f_2\left(\begin{matrix}c,b,0\\ c,c\end{matrix};\frac{u_3u_4}{u_1u_2},\frac{u_5u_6}{u_1}\right)u_1^{-\gamma}u_2^{-\beta}u_3^{\gamma-1}u_6^{\gamma-1} \tag{5.8}$$

where f_2 is defined by (2.31). Setting $u_5=u_6=1$ we find the (not very interesting) generating function

$$f_2\left(\begin{matrix}c,b,0\\ c,c\end{matrix};z,t\right)=\sum_{n=0}^{\infty}\frac{1}{(q;q)_n}\,{}_2\varphi_1\left(\begin{matrix}cq^n,b\\ c\end{matrix};z\right)t^n. \tag{5.9}$$

We shall return to this example after introducing the q-Kummer transformation symmetry.

The q-Kummer transformation

$$(5.10)\qquad {}_2\varphi_1\left(\begin{matrix}a,b\\ c\end{matrix};x\right)=\frac{(ax;q)_\infty}{(x;q)_\infty}\sum_{n=0}^{\infty}\frac{(a;q)_n(c/b;q)_n q^{n(n-1)/2}}{(c;q)_n(q;q)_n}\frac{(-bn)^n}{(ax;q)_n}$$

[1] and the q-Euler transform

$$(5.11)\qquad {}_2\varphi_1\left(\begin{matrix}a,b\\ c\end{matrix};x\right)=\frac{(abx/c;q)_\infty}{(x;q)_\infty}\,{}_2\varphi_1\left(\begin{matrix}c/a,c/b\\ c\end{matrix};\frac{abx}{c}\right)$$

[15, p. 97], [2] can be related to symmetries of (5.1). For this we consider the restriction of the operator Q, (5.1), to the space of convergent Laurent series in the monomials

$$(5.12)\qquad f_{k,\alpha,\beta,\gamma}=\left(\frac{u_3u_4}{u_1u_2}\right)^k u_1^{-\alpha}u_2^{-\beta}u_3^{\gamma-1}$$

where k is a nonnegative integer and α, β, γ are complex numbers such that $\gamma\neq 0, -1, -2, \cdots$. (That is, we do not consider the complication of logarithmic solutions.) We define the operator R_1 on this space as the unique linear operator such that

$$(5.13)\qquad \begin{aligned}R_1(f_{k,\alpha,\beta,\gamma})&=\frac{(azq^k;q)_\infty}{(z;q)_\infty}q^{k(k-1)/2}\left(\frac{-cz}{b}\right)^k u_1^{-\alpha}u_2^{\beta-\gamma}u_3^{\gamma-1}\\ &=\left(\frac{-c}{b}\right)^k q^{k(k-1)/2}\sum_{n=0}^{\infty}\frac{(aq^k;q)_n}{(q;q)_n}z^{n+k}u_1^{-\alpha}u_2^{\beta-\gamma}u_3^{\gamma-1},\end{aligned}$$

$$z=\frac{u_3u_4}{u_1u_2},\quad a=q^\alpha,\quad b=q^\beta,\quad c=q^\gamma.$$

(This is a q-analogy of the inversion in a cone conformal symmetry of $\partial_{12}-\partial_{34}\sim 0$.) Similarly we define R_2 by

$$(5.14)\qquad R_2(f_{k,\alpha,\beta,\gamma})=\left(\frac{-cz}{a}\right)^k q^{k(k-1)/2}\frac{(bzq^k;q)_\infty}{(z;q)_\infty}u_1^{\alpha-\gamma}u_2^{-\beta}u_3^{\gamma-1}$$

and linearity, and S by linearity and

$$(5.15)\qquad S(f_{k,\alpha,\beta,\gamma})=\frac{(cz/ab;q)_\infty}{(z;q)_\infty}\left(\frac{cz}{ab}\right)^k u_1^{\beta-\gamma}u_2^{\alpha-\gamma}u_3^{\gamma-1}.$$

It is not difficult to show that (5.10), (5.11) are equivalent to the assertion that R_1, R_2 and S are symmetries of (5.1). (The direct proofs of (5.10), (5.11) involve nothing more complicated than the q-Vandermonde theorem [3].)

Note that a basis for the solution space of $Q\sim 0$ and the eigenvalue equations $T_1^{-1}T_4^{-1}\sim a$, $T_2^{-1}T_4^{-1}\sim b$, $T_3T_4^{-1}\sim cq^{-1}$ consists of the functions

$${}_2\Phi_1\left(\begin{matrix}a,b\\ c\end{matrix}\right)={}_2\varphi_1\left(\begin{matrix}a,b\\ c\end{matrix};z\right)u_1^{-\alpha}u_2^{-\beta}u_3^{\gamma-1}\quad\text{and}$$

$${}_2\Phi_1'\left(\begin{matrix}a,b\\ c\end{matrix}\right)={}_2\varphi_1\left(\begin{matrix}qa/c,qb/c\\ q^2/c\end{matrix};z\right)z^{1-\gamma}u_1^{-\alpha}u_2^{-\beta}u_3^{\gamma-1}.$$

Furthermore the operators R_1, R_2, S satisfy

$$
\begin{aligned}
&R_1\,{}_2\Phi_1\begin{pmatrix} a,\, b \\ c \end{pmatrix} = {}_2\Phi_1\begin{pmatrix} a,\, c/b \\ c \end{pmatrix},\\
&R_1\,{}_2\Phi_1'\begin{pmatrix} a,\, b \\ c \end{pmatrix} = (-c^{1/2}/b)^{1-\gamma}{}_2\Phi_1'\begin{pmatrix} a,\, c/b \\ c \end{pmatrix},\\
&R_2\,{}_2\Phi_1\begin{pmatrix} a,\, b \\ c \end{pmatrix} = {}_2\Phi_1\begin{pmatrix} c/a,\, b \\ c \end{pmatrix},\\
&R_2\,{}_2\Phi_1'\begin{pmatrix} a,\, b \\ c \end{pmatrix} = (-c^{1/2}/a)^{1-\gamma}{}_2\Phi_1'\begin{pmatrix} c/a,\, b \\ c \end{pmatrix},\\
&R_1^2 = R_2^2 = S^2 = I, \qquad R_1R_2 = R_2R_1 = S,
\end{aligned}
\tag{5.16}
$$

where I is the identity operator. Other easily derived properties of R_1 are as follows:

$$
\begin{aligned}
&R_1E^\alpha R_1^{-1} = E^\alpha, \qquad R_1E^\beta R_1^{-1} = E_\beta T_2T_4,\\
&R_1E_\alpha R_1^{-1} = E_\alpha, \qquad R_1E_{\alpha\gamma}R_1^{-1} = -E_{\alpha\beta\gamma},\\
&R_1E_\beta R_1^{-1} = qE^\beta T_3T_2, \qquad R_1E^\gamma R_1^{-1} = -qE^{\beta\gamma}T_3T_2,\\
&R_1E_\gamma R_1^{-1} = E_{\beta\gamma}, \qquad R_1E^{\alpha\beta\gamma}R_1^{-1} = E^{\alpha\gamma}T_2T_4,\\
&R_1T_1^{-1}T_4^{-1}R_1^{-1} = T_1^{-1}T_4^{-1}, \qquad R_1T_2^{-1}T_4^{-1}R_1^{-1} = qT_2T_3,\\
&R_1T_3T_4^{-1}R_1^{-1} = T_3T_4^{-1}.
\end{aligned}
\tag{5.17}
$$

Similar results for R_2 follow from (5.17) and the interchanges $1 \leftrightarrow 2$, $\alpha \leftrightarrow \beta$, and the corresponding results for S follow from $S = R_1R_2 = R_2R_1$.

As an example of the use of these symmetries we consider a generating function Ψ characterized by

$$
E^\alpha E_\beta \sim \lambda, \qquad T_4^{-1}T_3 \sim \lambda, \qquad T_4^2T_1T_2 \sim 1/\lambda\mu q. \tag{5.18}
$$

Due to the occurrence of the operator E_β, it is not easy to find a simple form for the generating function by direct computation from Ψ. However, we can transform this problem into a simpler one. Indeed, $\Psi' = R_1\psi$ satisfies

$$
qE^\alpha E^\beta T_3T_2 \sim \lambda, \qquad T_4^{-1}T_3 \sim \lambda, \qquad T_1^{-1}T_2 \sim \mu. \tag{5.19}
$$

Although these equations are not in canonical form they are easy to solve by substituting a formal power series for Ψ'. The solution analytic at $u_1 = u_4 = 0$ is

$$
\Psi' = \sum_{k=0}^{\infty} \frac{q^{k(k-1)/2 - lk - l(l+1)/2}u_1^k u_2^{k+\mu}u_3^{l+\lambda}(q^{\mu+1}u_4)^l}{(q^{\mu+1};q)_k(q^{\lambda+1};q)_l(q;q)_k(q;q)_l}.
$$

From (5.13) we then find easily that

$$
\Psi = R_1\Psi' = \sum_{k=0}^{\infty}\frac{(z^{-1};q)_k(-zt/q)^k}{(q^{\mu+1};q)_k(q;q)_k}\sum_{l=0}^{\infty}\frac{(-q^{\lambda-1}zt)^l q^{l^2}}{(q^{\lambda+1};q)_l(q;q)_l}u_2^{-\mu-\lambda-1}u_3^\lambda \tag{5.20}
$$

where $z = u_3u_4/u_1u_2$, $t = u_1/u_2$. (The factorization in this expression is not surprising, based on variable separation for the corresponding differential equation problem and the fact that the operators E^α, E_β commute.) The generating relation is

$$
\Psi = \sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{(q;q)_n(q^{\mu+1};q)_n}\,{}_2\varphi_1\begin{pmatrix} q^{-n},\, q^{\mu+\lambda+n+1} \\ q^{\lambda+1} \end{pmatrix};\, z\Big)t^n u_2^{-\mu-\lambda-1}u_3^\lambda \tag{5.21}
$$

where the coefficient of ${}_2\varphi_1$ has been determined by setting $z=0$ in (5.20). This is the generating function for little q-Jacobi polynomials [10]; set $z=qx$.

For our final example we return to (5.6) and note that the function $\Psi' = R_2\Psi$ satisfies the conditions

$$E^{\alpha}T_1T_4 \sim -1, \quad T_2T_4 \sim b^{-1}, \quad T_3T_4^{-1} \sim cq^{-1}$$

which are easily solved by power series substitution to yield

$$\Psi' = \sum_{k,l=0}^{\infty} \frac{q^{-k(k+1)/2-lk}(q^{\beta}; q)_k(-z)^k u_1^{l+k} u_2^{-\beta} u_3^{\gamma-1}}{(q^{\gamma}; q)_k(q; q)_k(q; q)_l}.$$

Then from (5.14) we find

$$\Psi = R_2\Psi' = \frac{(q^{\beta}z; q)_{\infty}}{(z; q)_{\infty}(t; q)_{\infty}} \sum_{k=0}^{\infty} \frac{q^{k^2}(q^{\gamma}zt)^k t^{\gamma} u_2^{-\beta} u_3^{\gamma-1}}{(q^{\beta}z; q)_k(q^{\gamma}; q)_k(q; q)_k}$$

$$= \sum_{n=0}^{\infty} \frac{1}{(q; q)_n} {}_2\varphi_1\left(\begin{matrix} cq^n, b \\ c \end{matrix}; z\right) t^{n+\gamma} u_2^{-\beta} u_3^{\gamma-1}$$

where $z = u_3u_4/u_1u_2$, $t = u_1^{-1}$.

Our symmetry approach has profound relationships with the theory of orthogonal polynomials. We shall illustrate these relationships by presenting a new derivation of the orthogonality for little q-Jacobi polynomials which we normalize in the form

$$(5.22) \qquad \Phi_n^{(a,b)}(x) = {}_2\varphi_1\left(\begin{matrix} q^{-n}, q^{n+1}ab \\ aq \end{matrix}; qx\right), \qquad n = 0, 1, 2, \cdots$$

with $-1 < q < 1$, $0 < aq < 1$, $bq < 1$ [10]. The symmetries $E^{\alpha\beta\gamma}$ and $E_{\alpha\beta\gamma}$, (5.3), induce recurrences for these polynomials:

$$(5.23) \qquad \begin{aligned} \tau^{(a,b)}\Phi_n^{(a,b)}(x) &= \frac{q(1-q^{-n})(1-q^{n+1}ab)}{(1-aq)} \Phi_{n-1}^{(aq,bq)}(x), \\ \tau^{*(aq,bq)}\Phi_{n-1}^{(aq,bq)}(x) &= -(1-q)\Phi_n^{(a,b)}(x) \end{aligned}$$

where

$$\tau^{(a,b)} = \Delta_x^+, \qquad \tau^{*(aq,bq)} = (x-1)T_x^{-1} + (aq - abq^2x).$$

The existence of this pair of "raising" and "lowering" operators suggests that there might exist a Hilbert space structure with respect to which τ^* and τ are mutually adjoint, so that $\tau^*\tau$ would be selfadjoint.

To be more explicit, let $w_{a,b}(x)$ be a (complex-valued) weight function and $S_{a,b}$ the indefinite inner-product space of polynomials $f(x)$ with respect to the inner product

$$(5.24) \qquad (f_1, f_2)_{a,b} = \frac{1}{2\pi i} \int_C f_1(x) f_2(x) w_{a,b}(x) \frac{dx}{x}$$

where the contour C is a deformation of the circle $|x| = 1 + \varepsilon$, $\varepsilon > 0$ in the complex x-plane. Consider $\tau^{(a,b)}$ and $\tau^{*(aq,bq)}$ as mappings:

$$\tau^{(a,b)}: S_{a,b} \to S_{aq,bq}, \qquad \tau^{*(aq,bq)}: S_{aq,bq} \to S_{a,b}$$

and determine $w_{a,b}(x)$ so that

$$(5.25) \qquad (\tau f, g)_{aq,bq} = (f, \tau^* g)_{a,b}$$

for all $f \in S_{a,b}$, $g \in S_{aq,bq}$. A straightforward "integration by parts" yields the following conditions:

$$\frac{-q}{x}\frac{w_{aq,bq}(x/q)}{w_{a,b}(x)} = x-1, \qquad \frac{1}{x}\frac{w_{aq,bq}(x)}{w_{a,b}(x)} = aq(1-bqx)$$

with the solution

$$(5.26) \qquad \begin{aligned} w_{a,b}(x) &= \frac{(x/a;\,q)_\infty(qa/x;\,q)_\infty}{(xbq;\,q)_\infty(1/x;\,q)_\infty s(a,\,q)}, \\ s(a,\,q) &= (aq;\,q)_\infty(1/a;\,q)_\infty(-aq;\,q)_\infty(-1/a;\,q)_\infty. \end{aligned}$$

It follows immediately that $\tau^*\tau$ is selfadjoint on $S_{a,b}$ and from the recurrence relations (5.23) we have

$$(5.27) \qquad \tau^*\tau\Phi_n^{(a,b)} = \lambda_n\Phi_n^{(a,b)}, \qquad \lambda_n = -q(1-q^{-n})(1-q^{n+1}ab).$$

Clearly $\lambda_n \neq \lambda_m$ if $n \neq m$ and since eigenfunctions corresponding to distinct eigenvalues are orthogonal we have

$$(5.28) \qquad (\Phi_n^{(a,b)}, \Phi_m^{(a,b)})_{a,b} = 0 \quad \text{for } n \neq m.$$

From (5.23) and (5.25) with $f = \Phi_n^{(a,b)}$, $g = \Phi_{n-1}^{(aq,bq)}$ we obtain the following recurrence:

$$(5.29) \qquad \|\Phi_n^{(a,b)}\|_{a,b}^2 = \frac{-q(1-q^{-n})(1-q^{n+1}ab)}{(1-aq)^2}\|\Phi_{n-1}^{(aq,bq)}\|_{aq,bq}^2.$$

From (5.29) we can compute $\|\Phi_n^{(a,b)}\|_{a,b}^2$ once we know $\|1\|_{a,b}^2 = (\Phi_0^{(a,b)}, \Phi_0^{(a,b)})_{a,b}$ for all admissible a, b.

We now turn to the task of computing $\|1\|_{a,b}^2$. We know that $(\Phi_1^{(a,b)}, \Phi_0^{(a,b)})_{a,b} = 0$ and, substituting the explicit expression (5.22) for the orthogonal polynomial $\Phi_1^{(a,b)}(x)$, we can write this relation in the form

$$(5.30) \qquad \|1\|_{a,bq}^2 = \frac{(1-bq)}{(1-abq^2)}\|1\|_{a,b}^2.$$

(Here we have used the evident fact that $(1-xbq, 1)_{a,b} = \|1\|_{a,bq}^2$.) To obtain an additional condition on the norm we consider the symmetries E^γ, E_γ in the form:

$$(5.31) \qquad \begin{aligned} \mu^{(a,b)}\Phi_n^{(a,b)} &= (1-a)\Phi_n^{(aq^{-1},bq)}, \\ \mu^{*(aq^{-1},bq)}\Phi_n^{(aq^{-1},bq)} &= \frac{q^{-n}(1-aq^n)(1-bq^{n+1})}{a(1-a)}\Phi_n^{(a,b)} \end{aligned}$$

where

$$\mu^{(a,b)} = 1 - aT_x^1, \qquad \mu^{(a,b)}: S_{a,b} \to S_{aq^{-1},bq},$$

$$\mu^{*(aq^{-1},bq)} = \frac{x-1}{ax}T_x^{-1} + \frac{1-bqx}{ax}, \qquad \mu^{*(aq^{-1},bq)}: S_{aq^{-1},bq} \to S_{a,b}.$$

It is easily verified that

$$(5.32) \qquad (\mu f, g)_{aq^{-1},bq} = (f, \mu^* g)_{a,b}$$

for all $f \in S_{a,b}$, $g \in S_{aq^{-1},bq}$ so that μ and μ^* are mutually adjoint. Setting $f = \Phi_0^{(a,b)}$, $g = \Phi_0^{(aq^{-1},bq)}$ in this relation, we see immediately that

$$(5.33) \qquad \|1\|_{aq^{-1},bq}^2 = \frac{(1-bq)}{a(1-a)}\|1\|_{a,b}^2.$$

The recurrences (5.30), (5.33) have the solution

$$\|1\|^2_{a,b}=\frac{(abq^2;\,q)_\infty K(q)}{(bq;\,q)_\infty(aq;\,q)_\infty(-1/a;\,q)_\infty(-aq;\,q)_\infty}.$$

Thus

$$\frac{1}{2\pi i}\int_C \frac{(x/a;\,q)_\infty(qa/x;\,q)_\infty dx}{(1/a;\,q)_\infty(1/x;\,q)_\infty(xbq;\,q)_\infty x}=\frac{(abq^2;\,q)_\infty K(q)}{(bq;\,q)_\infty(aq;\,q)_\infty}. \tag{5.34}$$

(Here we are assuming $a>1+\varepsilon>1>qa$.) With the choice $bq=1/a$ the integral becomes trivial to evaluate and we find that $K(q)=1/(q;\,q)_\infty$. The complex orthogonality relations just determined can be recast as real discrete orthogonality through evaluation of the contour integral by residues at the poles $x=q^k$, $k=0,1,2,\cdots$. The final result is

$$\begin{aligned}&\sum_{k=0}^{\infty}\frac{(aq)^k(q^{k+1};\,q)_\infty}{(bq^{k+1};\,q)_\infty}\Phi_n^{(a,b)}(q^k)\Phi_m^{(a,b)}(q^k)\\ &\qquad=\frac{(aq)^n(abq^{n+1};\,q)_\infty(q;\,q)_n\delta_{m,n}}{(bq^{n+1};\,q)_\infty(aq;\,q)_\infty(aq;\,q)_n(1-abq^{2n+1})}.\end{aligned} \tag{5.35}$$

Note that the proof of this result follows entirely from the symmetries; no special function identities are needed.

The ideas behind this derivation of orthogonality relations can be generalized substantially. In particular in [12] it is shown how to derive the orthogonality relations for the Askey-Wilson polynomials (the most general extension of the classical orthogonal polynomials known) using this symmetry method. A simple corollary of the derivation is an identity for ${}_4\varphi_3$ polynomials (Sears' transformation) that includes the q-Kummer and q-Euler transforms as special cases.

The fundamental symmetry concepts introduced in this paper extend to the very important q-series of the form

$${}_{r+1}\varphi_r\begin{pmatrix}a_i,x\\ b_j\end{pmatrix};q\Big) \quad\text{and}\quad {}_{r+1}\varphi_r\begin{pmatrix}a_i,ax,a/x\\ b_j\end{pmatrix};q\Big).$$

They will be the subject of future papers by the authors.

REFERENCES

[1] G. E. Andrews, *On the q-analogue of Kummer's theorem and applications*, Duke Math. J., 40 (1973), pp. 525-528.

[2] ———, *The theory of partitions*, in Encyclopedia of Mathematics and Its Applications, Vol. 2, Addison-Wesley, Reading, MA, 1976.

[3] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge University Press, 1935; Reprinted by Stechert-Hafner, New York, 1964.

[4] J. L. Burchnall and T. W. Chaundy, *Expansions of Appell's double hypergeometric functions* I, Quart. J. Math. Oxford Ser., 12 (1940), pp. 249-270.

[5] T. W. Chaundy, *Expansions of hypergeometric functions*, Quart. J. Math. Oxford Ser., 13 (1940), pp. 159-171.

[6] J. Fields and M. Ismail, *Polynomial expansions*, Math. Comp., 29 (1976), pp. 894-902.

[7] W. Hahn, *Uber Orthogonal polynome, die q-Differenzengleichen genugen*, Math. Nachr., 2 (1949), pp. 4-34.

[8] ———, *Beitrage zur theorie der heischen reihen*, Math. Nachr., 2 (1949), pp. 340-379.

[9] J. Hrabowski, *Multiple hypergeometric functions and simple Lie groups* SL *and* Sp, this Journal, 16 (1985), pp. 876-886.

[10] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and* $_4\Phi_3$ *polynomials*, J. Approx. Theory, 36 (1982), pp. 43-54.

[11] E. G. KALNINS, H. L. MANOCHA AND W. MILLER, *The Lie theory of two-variable hypergeometric functions*, Stud. Appl. Math., 62 (1980), pp. 143-173.

[12] E. G. KALNINS AND W. MILLER, *Symmetry techniques for q-series: the Askey-Wilson polynomials*, Rocky Mountain J. Math., to appear.

[13] W. MILLER, *Lie theory and generalized hypergeometric functions*, this Journal, 3 (1972), pp. 31-44.

[14] ———, *Lie theory and generalizations of the hypergeometric functions*, SIAM J. Appl. Math., 25 (1973), pp. 226-235.

[15] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.

[16] L. WEISNER, *Group-theoretic origin of certain generating functions*, Pacific J. Math., 5 (1955), pp. 1033-1039.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

003

A PROBABILISTIC PROOF OF RAMANUJAN'S ${}_1\psi_1$ SUM*

KEVIN W. J. KADELL†

Abstract. The q-binomial theorem and Ramanujan's ${}_1\psi_1$ sum provide q-analogues of the beta and F distributions, respectively. We obtain these discrete distributions using order statistics, thus deriving the summation formulas probabilistically. We obtain a q-analogue of the χ^2 distribution and interpret it probabilistically. We show that an appropriate ratio follows our q-analogue of the F distribution.

Key words. Ramanujan's ${}_1\psi_1$ sum, q-binomial theorem, q-gamma function, Gaussian polynomial, order statistics, beta distribution, F distribution, and χ^2 distribution

AMS(MOS) subject classifications. Primary 33A15; secondary 05A15, 60C05, 62E15

1. Introduction and summary. Fix q with $|q|<1$ and set

$$(1.1)\qquad \begin{aligned} &(\alpha)_0=1,\\ &(\alpha)_n=\prod_{i=0}^{n-1}(1-\alpha q^i), \qquad n\geqq 1,\\ &(\alpha)_\infty=\lim_{n\to\infty}(\alpha)_n=\prod_{i=0}^{\infty}(1-\alpha q^i). \end{aligned}$$

The q-binomial theorem is (see [3, (2.2.1)])

$$(1.2)\qquad \sum_{n=0}^{\infty}\frac{(a)_n}{(q)_n}t^n=\frac{(at)_\infty}{(t)_\infty}, \qquad |t|<1.$$

We can interpret $(\alpha)_n$ for all integers n by

$$(1.3)\qquad (\alpha)_n=\frac{(\alpha)_\infty}{(\alpha q^n)_\infty}.$$

Ramanujan's ${}_1\psi_1$ sum is

$$(1.4)\qquad \sum_{n=-\infty}^{\infty}\frac{(a)_n}{(b)_n}t^n=\frac{(at)_\infty(qa^{-1}t^{-1})_\infty(q)_\infty(ba^{-1})_\infty}{(t)_\infty(ba^{-1}t^{-1})_\infty(b)_\infty(qa^{-1})_\infty}, \qquad |ba^{-1}|<|t|<1.$$

The sum on the left side of (1.4) is usually denoted ${}_1\psi_1\left[{a;q,t \atop b}\right]$. By (1.3), we have

$$(1.5)\qquad \frac{1}{(q)_n}=0, \qquad n<0.$$

Thus Ramanujan's sum (1.4) reduces to the q-binomial theorem (1.2) when $b=q$. Ismail [12] gave an elegant proof of (1.4) by observing that (1.4) reduces to (1.2) when $b=q^n$ for all integers n and that both sides of (1.4) are analytic at $b=0$. Ramanujan's sum (1.4) then follows using the following well-known theorem.

THEOREM 1. *If f and g are analytic at z_0 and agree at infinitely many points which include z_0 as an accumulation point, then $f=g$.*

* Received by the editors June 4, 1986; accepted for publication July 30, 1986.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287. Most of this work was performed on a postdoctoral appointment at the Univesity of Wisconsin, Madison, Wisconsin 53706, supported by a grant from the National Science Foundation. The work of this author was supported in part by a faculty grant-in-aid from Arizona State University, Tempe, Arizona 85287.

See Andrews [1], [2], Andrews and Askey [4], Askey [6], Hahn [11] and M. Jackson [14] for some other proofs of (1.4).

There are a number of other opportunites to use Ismail's argument to prove summation formulas for basic hypergeometric series. Askey and Ismail [7] used it to prove the summation formula for a very well poised ${}_6\psi_6$ using the limiting form of Jackson's theorem. Gauss [10] introduced the q-binomial coefficient

$$\begin{bmatrix} m \\ k \end{bmatrix} = \frac{(1-q^m)(1-q^{m-1})\cdots(1-q^{m-k+1})}{(1-q)(1-q^2)\cdots(1-q^k)}, \qquad 0 \leqq k \leqq m, \tag{1.6}$$

which clearly tends to $\binom{m}{k}$ as q tends to 1. Setting $a=q^{-m}$, $t=xq^{N+m}$, where $m \geqq 0$ in (1.2), we obtain

$$\sum_{k=0}^{m} q^{Nk+\binom{k}{2}} \begin{bmatrix} m \\ k \end{bmatrix} (-x)^k = (xq^N)_m. \tag{1.7}$$

Equating coefficients in (1.7), we have the representation

$$\sum_{N \leqq n_1 < n_2 < \cdots < n_k \leqq N+m-1} q^{[\sum_{i=1}^{k} n_i]} = q^{Nk+\binom{k}{2}} \begin{bmatrix} m \\ k \end{bmatrix}, \qquad 0 \leqq k \leqq m, \tag{1.8}$$

from which it immediately follows that $\left[\begin{smallmatrix} m \\ k \end{smallmatrix}\right]$ is a polynomial. Equivalent formulations of (1.8) occur in a number of related topics. Kendall and Stuart [15, pp. 496-512] evaluate the characteristic function of the Wilcoxon rank sum (Mann-Whitney U) statistic. Polya [16] gives the probability generating function of an equivalent statistic on lattice paths. Andrews [3, Chaps. 3 and 13] gives an extensive account of the role of $\left[\begin{smallmatrix} m \\ k \end{smallmatrix}\right]$ in the theory of partitions and finite vector spaces.

The basic result (1.8) implies (1.7) directly. Thus (1.2) holds for $a=q^{-m}$, $t=xq^{N+m}$. Set $z=at$. Equation (1.2) now follows using Theorem 1 twice since both sides are analytic at $z=0$ and at $t=0$.

Askey [5] has observed that (1.2) and (1.4) provide q-analogues of the beta integral

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^1 t^{(x-1)}(1-t)^{(y-1)}\,dt \tag{1.9}$$

$$= c^x \int_0^\infty s^{(x-1)} \frac{ds}{(1+cs)^{(x+y)}}, \tag{1.10}$$

where $c>0$, Re $(x)>0$, and Re $(y)>0$, on the intervals $(0,1)$ and $(0,\infty)$, respectively. When these are normalized to integrate to 1, the resulting probability mass functions provide q-analogues of the beta and F distributions, respectively.

In § 2, we use (1.8) to show that the order statistics of a certain discrete random sample follow our q-beta distribution. This proves the q-binomial theorem (1.2) using Theorem 1. In § 3, we use a simple transformation to show that the F distribution also arises from order statistics. We give the q-F distribution with both degrees of freedom equal to 2. In § 4, we give an extension of (1.8). We use the order statistics of a discrete random sample to obtain the q-F distribution. This establishes Ramanujan's sum (1.4) using Theorem 1. In § 5, we set the denominator degrees of freedom in our q-F distribution equal to ∞. This gives a q-analogue of the χ^2 distribution with $2x$ degrees of freedom, which we interpret probabilistically. We obtain a different q-analogue of the χ^2 distribution from our q-beta distribution. We show that an appropriate ratio follows our q-F distribution.

2. The q-beta distribution. Let x and y be positive integers and let u_i, $1 \leqq i \leqq x+y-1$, be i.i.d. uniform random variables on $(0, 1)$. We denote the order statistics by

$$u_{(1)} < u_{(2)} < \cdots < u_{(x+y-1)}. \tag{2.1}$$

The xth order statistic $u_{(x)}$ has a beta distribution with parameters x and y. The density is

$$be_{x,y}(t) = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} t^{(x-1)}(1-t)^{(y-1)}, \qquad 0 < t < 1, \tag{2.2}$$

which integrates to 1 by (1.9). Feller [8, pp. 21-24] establishes this fact by a direct combinatorial argument. We require the case $q = 1$ of (1.8). It is

$$\sum_{N \leqq n_1 < n_2 < \cdots < n_k \leqq N+m-1} 1 = \binom{m}{k} = \frac{m!}{k!(m-k)!}. \tag{2.3}$$

Let $u_{(x)} = u_n$ and set

$$S = \{s_1, \cdots, s_{x-1}\} = \{s \mid u_s < u_{(x)}\}, \qquad L = \{l_1, \cdots, l_{y-1}\} = \{l \mid u_l > u_{(x)}\}. \tag{2.4}$$

There are

$$(x+y-1)\binom{x+y-2}{x-1} = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}$$

ways of choosing n, S and L. We have

$$\begin{aligned} &\frac{1}{\Delta t} \text{Prob}\left[t - \frac{\Delta t}{2} \leqq u_{(x)} < t + \frac{\Delta t}{2}\right] \\ &= \sum_{n,S,L} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \text{Prob}\,[u_s < u_n \text{ for all } s \in S \text{ and } u_l > u_n \text{ for all } l \in L]\,du_n \\ &= \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} u^{(x-1)}(1-u)^{(y-1)}\,du. \end{aligned} \tag{2.5}$$

The density $be_{x,y}(t)$ of $u_{(x)}$ follows by letting Δt tend to 0.

Fix q with $0 < q < 1$. F. H. Jackson [13] introduced the q-gamma function

$$\Gamma_q(a) = \frac{(q)_\infty}{(q^a)_\infty}(1-q)^{(1-a)}, \tag{2.6}$$

which clearly tends to $\Gamma(a)$ when a is a positive integer. Askey [5] treats the general case. F. H. Jackson [13] also gave the q-integral

$$\int_0^1 f(t)\,d_qt = (1-q) \sum_{n=0}^{\infty} q^n f(q^n). \tag{2.7}$$

This is the approximation which results when the unit interval is partitioned by the powers of q. The measure du of a uniform random variable u is replaced by

$$d_q(q^n) = q^n(1-q) = \text{Prob}\,[q^{n+1} \leqq u < q^n], \qquad n \geqq 0, \tag{2.8}$$

which is the probability mass function of a geometric distribution with parameter $\theta = 1 - q$. While n follows a generalization of the exponential distribution, q^n follows a natural discrete version of the uniform.

A random sample from a continuous distribution will be distinct with probability 1. Choose $x+y-1$ distinct observations from the distribution (2.8). Our sample contains

$$q^{n_{(x+y-1)}} < \cdots < q^{n_{(2)}} < q^{n_{(1)}} \leqq 1, \tag{2.9}$$

where the exponents are

$$0 \leqq n_{(1)} < n_{(2)} < \cdots < n_{(x+y-1)}. \tag{2.10}$$

We require the limiting case of (1.8) as m tends to ∞. It is

$$\sum_{N \leqq n_1 < n_2 < \cdots < n_k} q^{[\sum_{i=1}^{k} n_i]} = \frac{q^{Nk+\binom{k}{2}}}{(q)_k}. \tag{2.11}$$

The joint probability mass function of the order statistics is given by

$$q^{-\binom{x+y-1}{2}} (q)_{(x+y-1)} q^{[\sum_{i=1}^{x+y-1} n_{(i)}]}, \tag{2.12}$$

where (2.10) holds and 0 otherwise. This is clearly correct up to a constant which is supplied by setting $N=0$, $k=x+y-1$, in (2.11). The xth order statistic from our sample (2.9) is $q^{n_{(y)}}$. We have the probability mass function

$$\begin{aligned} {}_q be_{x,y}(q^t) &= \operatorname{Prob}[n_{(y)} = y-1+t] \\ &= q^{-\binom{x+y-1}{2}} (q)_{(x+y-1)} \left(\sum_{0 \leqq n_{(1)} < \cdots < n_{(y-1)} \leqq y-2+t} q^{[\sum_{i=1}^{y-1} n_{(i)}]} \right) \\ &\quad \cdot q^{(y-1+t)} \left(\sum_{y+t \leqq n_{(y+1)} < \cdots < n_{(x+y-1)}} q^{[\sum_{i=y+1}^{x+y-1} n_{(i)}]} \right), \end{aligned} \tag{2.13}$$

where $t \geqq 0$. We may evaluate the two sums on the right side of (2.13) by setting $N=0$, $m=y-1+t$ and $k=y-1$ in (1.8) and $N=y+t$, $k=x-1$ in (2.11). This yields

$$\begin{aligned} {}_q be_{x,y}(q^t) &= q^{-\binom{x+y-1}{2}} (q)_{(x+y-1)} q^{\binom{y-1}{2}} \begin{bmatrix} y-1+t \\ y-1 \end{bmatrix} \frac{q^{(y+t)(x-1)+\binom{x-1}{2}}}{(q)_{(x-1)}} \\ &= q^{tx} \frac{(q)_{(x+y-1)}}{(q)_{(x-1)}(q)_{(y-1)}} \frac{(q)_{(y-1+t)}}{(q)_t} \\ &= \frac{\Gamma_q(x+y)}{\Gamma_q(x)\Gamma_q(y)} (q^t)^{(x-1)} (q^{t+1})_{(y-1)} [q^t(1-q)]. \end{aligned} \tag{2.14}$$

Since ${}_q be_{x,y}(q^t)$ is a probability, we have

$$\begin{aligned} 1 &= \sum_{t=0}^{\infty} {}_q be_{x,y}(q^t) \\ &= \sum_{t=0}^{\infty} q^{tx} \frac{(q)_{(x+y-1)}}{(q)_{(x-1)}(q)_{(y-1)}} \frac{(q)_{(y-1+t)}}{(q)_t} \\ &= \frac{(q^x)_\infty}{(q^{x+y})_\infty} \sum_{t=0}^{\infty} \frac{(q^y)_t}{(q)_t} (q^x)^t. \end{aligned} \tag{2.15}$$

Solving for the sum on the right side of (2.15), we find that (1.2) holds for $a=q^y$, $t=q^x$, where x and y are positive integers. Since both sides are analytic at $a=0$ and at $t=0$, the result (1.2) follows from two applications of Theorem 1.

We may also rewrite (2.15) as the q-analogue

$$\int_0^1 t^{(x-1)} \frac{(qt)_\infty}{(q^y t)_\infty} d_q t = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)} \tag{2.16}$$

of the beta integral (1.9). Two applications of Theorem 1 establish (2.16) for Re $(x)>0$, Re $(y)>0$. See Askey [5].

3. A transformation. Let $c>0$ be a fixed scale parameter. Then

$$v_i=c^{-1}\frac{u_i}{(1-u_i)},\qquad 1\leqq i\leqq x+y-1, \tag{3.1}$$

are i.i.d. random variables with the density function

$$\frac{c}{(1+ct)^2},\qquad t>0. \tag{3.2}$$

Since our transformation is monotonic, it preserves order. Thus the xth order statistic

$$v_{(x)}=c^{-1}\frac{u_{(x)}}{(1-u_{(x)})} \tag{3.3}$$

has the density function

$$c^x\frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}\frac{s^{(x-1)}}{(1+cs)^{x+y}},\qquad s>0. \tag{3.4}$$

This integrates to 1 by (1.10). For $c=xy^{-1}$, we observe that $v_{(x)}$ has an F distribution with numerator and denominator degrees of freedom $2x$ and $2y$, respectively. This also follows from Fisher's principle [9]. Let e_i, $1\leqq i\leqq x+y$, be i.i.d. exponential random variables (with any scale parameter) and set $s=e_1+e_2+\cdots+e_{x+y}$. Then the random variables

$$\left(\frac{e_1}{s},\frac{e_1+e_2}{s},\cdots,\frac{e_1+e_2+\cdots+e_{x+y-1}}{s}\right) \tag{3.5}$$

have the same joint distribution as the order statistics

$$(u_{(1)},u_{(2)},\cdots,u_{(x+y-1)}). \tag{3.6}$$

For a given value of the sum s, the joint density of the random variables (3.5) is constant. Integrating over s, the joint density of (3.5) is constant, as is that of the order statistics (3.6). Since they have the same support, they both have the joint density

$$f(t_1,t_2,\cdots,t_{x+y-1})=\begin{cases}(x+y-1)! & \text{if } 0<t_1<\cdots<t_{x+y-1}<1,\\ 0 & \text{otherwise.}\end{cases} \tag{3.7}$$

The density of $u_{(x)}$ may be obtained by integrating over all of the variables except t_x. This is, of course, what was done in § 2 with the q-analogue (2.12) of (3.7).

Choosing the common scale parameter $E(e_1)=2$, we have

$$e_i\sim\chi^2(2),\qquad 1\leqq i\leqq x+y. \tag{3.8}$$

We have the independent random variables

$$\omega_x=e_1+e_2+\cdots+e_x\sim\chi^2(2x),\qquad \omega_y=e_{x+1}+\cdots+e_{x+y}\sim\chi^2(2y). \tag{3.9}$$

Observe for a Poisson process with parameter 1, that ω_x is the waiting time for the xth event and ω_y is the waiting time between the xth and the $(x+y)$th events. Clearly $s=\omega_x+\omega_y\sim\chi^2(2x+2y)$. We have

$$u_{(x)}\sim\frac{e_1+e_2+\cdots+e_x}{s}=\frac{\omega_x}{\omega_x+\omega_y} \tag{3.10}$$

and

$$v_{(x)}=c^{-1}\frac{u_{(x)}}{(1-u_{(x)})}\sim c^{-1}\frac{\omega_x}{\omega_y}\sim c^{-1}\frac{\chi^2(2x)}{\chi^2(2y)}, \tag{3.11}$$

which has an F distribution when $c=xy^{-1}$.

We require a q-analogue of the density function (3.2). Let u be uniformly distributed on $(0,1)$ and set $s=q^n$, where n is the unique integer satisfying

$$q^{n+1}\leqq c^{-1}\frac{u}{(1-u)}<q^n. \tag{3.12}$$

We have the probability mass function

$$\begin{aligned}\text{Prob}\left[q^{n+1}\leqq c^{-1}\frac{u}{(1-u)}<q^n\right]&=\text{Prob}\left[\frac{cq^{n+1}}{(1+cq^{n+1})}\leqq u<\frac{cq^n}{(1+cq^n)}\right]\\&=\frac{c(1-q)q^n}{(1+cq^n)(1+cq^{n+1})}=\frac{c(1-q)q^n}{(-cq^n)_2}.\end{aligned} \tag{3.13}$$

This provides a q-analogue of the F distribution with both degrees of freedom equal to 2.

4. The *q-F* distribution. We want to find the distribution of the order statistics of an appropriate random sample from the distribution (3.13). This requires a new representation of the Gaussian polynomial $[{m \atop k}]$. Replace each n_i on the right side of (2.3) by n_i+i-1. Then (2.3) becomes

$$\sum_{\substack{N\leqq n_1,n_k\leqq N+m+k-2\\ n_i+2\leqq n_{i+1},1\leqq i\leqq k-1,}} 1=\binom{m}{k}. \tag{4.1}$$

We have the following theorem.

THEOREM 2.

$$\sum_{\substack{N\leqq n_1,n_k\leqq N+m+k-2\\ n_i+2\leqq n_{i+1},1\leqq i\leqq k-1,}}\prod_{i=1}^{k}\frac{q^{n_i}}{(-cq^{n_i})_2}=\frac{q^{Nk+2\binom{k}{2}}}{(-cq^N)_k(-cq^{N+m})_k}\begin{bmatrix}m\\k\end{bmatrix},\qquad 0\leqq k\leqq m. \tag{4.2}$$

Proof. We proceed by induction on k. For $k=1$, we have

$$\begin{aligned}\sum_{N\leqq n_1\leqq N+m-1}\frac{q^{n_1}}{(-cq^{n_1})_2}&=\frac{1}{c(1-q)}\sum_{n=N}^{N+m-1}\text{Prob}\left[\frac{cq^{n+1}}{(1+cq^{n+1})}\leqq u<\frac{cq^n}{(1+cq^n)}\right]\\&=\frac{1}{c(1-q)}\text{Prob}\left[\frac{cq^{N+m}}{(1+cq^{N+m})}\leqq u<\frac{cq^N}{(1+cq^N)}\right]\\&=\frac{q^N}{(1+cq^N)(1+cq^{N+m})}\frac{(1-q^m)}{(1-q)},\end{aligned} \tag{4.3}$$

as required. For each $k\geqq 1$, we proceed by induction on m. For $m=k$, the sum on the left side of (4.2) contains only the term with

$$n_i=N+2i-2,\qquad 1\leqq i\leqq k, \tag{4.4}$$

and we easily verify (4.2). Let $F(N,m,k)$ denote the sum on the left side of (4.2). Classifying the terms of the sum according to whether $n_1>N$ or $n_1=N$, we have

$$F(N,m,k)=F(N+1,m-1,k)+\frac{q^N}{(-cq^N)_2}F(N+2,m-1,k-1). \tag{4.5}$$

Substitute the right side of (4.2) into (4.5) and divide by

$$\frac{q^{Nk+2\binom{k}{2}}}{(-cq^N)_k(-cq^{N+m})_k}\begin{bmatrix} m-1 \\ k-1 \end{bmatrix}.$$

This gives

$$\frac{(1-q^m)}{(1-q^k)}=q^k\frac{(1-q^{m-k})}{(1-q^k)}\frac{(1+cq^N)}{(1+cq^{N+k})}+\frac{(1+cq^{N+m})}{(1+cq^{N+k})}, \tag{4.6}$$

which is easily verified. Equation (4.5) completes our induction on k and m. □

Observe that the condition of minimal difference 2 is precisely what is required for each term of the sum on the left side of (4.2) to have only simple poles as a function of c. We have the limiting cases $m=\infty$

$$\sum_{\substack{N\leqq n_1 \\ n_i+2\leqq n_{i+1}, 1\leqq i\leqq k-1,}} \prod_{i=1}^{k}\frac{q^{n_i}}{(-cq^{n_i})_2}=\frac{q^{Nk+2\binom{k}{2}}}{(-cq^N)_k}\frac{1}{(q)_k} \tag{4.7}$$

and $N=-\infty$, $N+m=M$,

$$\sum_{\substack{n_k\leqq M+k-2 \\ n_i+2\leqq n_{i+1}, 1\leqq i\leqq k-1,}} \prod_{i=1}^{k}\frac{q^{n_i}}{(-cq^{n_i})_2}=\frac{c^{-k}q^{\binom{k}{2}}}{(-cq^M)_k}\frac{1}{(q)_k}. \tag{4.8}$$

We also require the case $N=-\infty$ of (4.7) (or $M=\infty$ of (4.8)). It is

$$\sum_{n_i+2\leqq n_{i+1}, 1\leqq i\leqq k-1,} \prod_{i=1}^{k}\frac{q^{n_i}}{(-cq^{n_i})_2}=\frac{c^{-k}q^{\binom{k}{2}}}{(q)_k}. \tag{4.9}$$

Choose a sample q^{n_i}, $1\leqq i\leqq x+y-1$, of $x+y-1$ random variables from the distribution (3.13) subject to the condition

$$n_{(i)}+2\leqq n_{(i+1)}, \qquad 1\leqq i\leqq x+y-2. \tag{4.10}$$

The joint probability mass function of the order statistics is given by

$$q^{-\binom{x+y-1}{2}}(q)_{(x+y-1)}\prod_{i=1}^{x+y-1}\frac{cq^{n_{(i)}}}{(-cq^{n_{(i)}})_2}, \tag{4.11}$$

where (4.10) holds and 0 otherwise. The constant is obtained by setting $k=x+y-1$ in (4.9). The xth order statistic is $q^{n_{(y)}}$. It has probability mass function

$$\begin{aligned} {}_q^cF_{2x,2y}(q^s) &= \text{Prob}\,[n_{(y)}=y-1+s] \\ &= q^{-\binom{x+y-1}{2}}(q)_{(x+y-1)}\Bigg(\sum_{\substack{n_{(y-1)}\leqq y-3+s \\ n_{(i)}+2\leqq n_{(i+1)}, 1\leqq i\leqq y-2,}}\prod_{i=1}^{y-1}\frac{cq^{n_{(i)}}}{(-cq^{n_{(i)}})_2}\Bigg) \\ &\quad\cdot\frac{cq^{y-1+s}}{(-cq^{y-1+s})_2}\Bigg(\sum_{\substack{y+1+s\leqq n_{(y+1)} \\ n_{(i)}+2\leqq n_{(i+1)}, y+1\leqq i\leqq x+y-2,}}\prod_{i=y+1}^{x+y-1}\frac{cq^{n_{(i)}}}{(-cq^{n_{(i)}})_2}\Bigg). \end{aligned} \tag{4.12}$$

We may evaluate the two sums on the right side of (4.12) by setting $N=y+1+s$, $k=x-1$ in (4.7) and $M=s$, $k=y-1$ in (4.8). This yields

$$\begin{aligned} {}_q^cF_{2x,2y}(q^s) &= q^{-\binom{x+y-1}{2}}(q)_{(x+y-1)}\frac{q^{\binom{y-1}{2}}}{(-cq^5)_{(y-1)}}\frac{1}{(q)_{(y-1)}} \\ &\quad\cdot\frac{cq^{y-1+s}}{(-cq^{y-1+s})_2}c^{(x-1)}\frac{q^{(y+1+s)(x-1)+2\binom{x-1}{2}}}{(-cq^{y+1+s})_{(x-1)}}\frac{1}{(q)_{(x-1)}} \end{aligned}$$

(4.13)
$$= \frac{c^x q^{\binom{x}{2}+sx}}{(-cq^s)_{(x+y)}} \frac{(q)_{(x+y-1)}}{(q)_{(x-1)}(q)_{(y-1)}}$$
$$= c^x q^{\binom{x}{2}} \frac{\Gamma_q(x+y)}{\Gamma_q(x)\Gamma_q(y)} \frac{(q^s)^{(x-1)}}{(-cq^s)_{(x+y)}} [q^s(1-q)].$$

Since ${}_q^cF_{2x,2y}(q^s)$ is a probability, we have

(4.14)
$$1 = c^x q^{\binom{x}{2}} \frac{(q)_{(x+y-1)}}{(q)_{(x-1)}(q)_{(y-1)}} \sum_{s=-\infty}^{\infty} \frac{(q^x)^s}{(-cq^s)_{(x+y)}}.$$

Observe that

(4.15)
$$c^x q^{\binom{x}{2}} = \frac{(-c)_\infty(-c^{-1}q)_\infty}{(-cq^x)_\infty(-c^{-1}q^{1-x})_\infty}$$

holds, but only when x is an integer. Solving for the sum on the right side of (4.14), we obtain

(4.16)
$$\sum_{s=-\infty}^{\infty} \frac{(-c)_s}{(-cq^{x+y})_s}(q^x)^s = \frac{(-cq^x)_\infty(-c^{-1}q^{1-x})_\infty(q)_\infty(q^{x+y})_\infty}{(-cq^{x+y})_\infty(-c^{-1}q)_\infty(q^x)_\infty(q^y)_\infty}.$$

Thus Ramanujan's sum (1.4) holds for $a=-c$, $b=-cq^{x+y}$, $t=q^x$. Since

(4.17)
$$(a)_{-n} = \frac{(a)_\infty}{(aq^{-n})_\infty} = \frac{1}{(1-aq^{-n})\cdots(1-aq^{-1})} = \frac{q^{\binom{n+1}{2}}}{(-a)^n}\frac{1}{(qa^{-1})_n},$$

we have

(4.18)
$$\sum_{n=-\infty}^{\infty} \frac{(a)_n}{(b)_n} t^n = \sum_{n=-\infty}^{\infty} \frac{(a)_{-n}}{(b)_{-n}} t^{-n} = \sum_{n=-\infty}^{\infty} \frac{(qb^{-1})_n}{(qa^{-1})_n}(ba^{-1}t^{-1})^n.$$

This shows the analyticity at $b=0$ used by Ismail [12]. Both sides of (1.4) are analytic at any value of a which does not give a pole. Fix $t=q^x$, where x is a positive integer. Two applications of Theorem 1 establish Ramanujan's sum (1.4) in this case. Now fix a and t. By (4.18), (1.4) holds for $ba^{-1}t^{-1}=q^x$, which is $b=atq^x$. The full result (1.4) now follows by a third application of Theorem 1.

Using

(4.19)
$$\int_0^\infty f(s)\,d_qs = (1-q)\sum_{n=-\infty}^{\infty} q^n f(q^n),$$

we may write (4.14) as the q-analogue

(4.20)
$$\int_0^\infty s^{(x-1)}\frac{(-csq^{x+y})_\infty}{(-cs)_\infty}\,d_qs = \frac{(-cq^x)_\infty(-c^{-1}q^{1-x})_\infty}{(-c)_\infty(-c^{-1}q)_\infty}\frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}$$

of the beta integral (1.10). Three applications of Theorem 1 establish (4.20) for $\operatorname{Re}(x)>0$, $\operatorname{Re}(y)>0$, and all c for which the integrand has no poles. See Askey [5].

5. The $q-\chi^2$ distribution. The F distribution normally occurs as the ratio of two χ^2 distributions, each of which is divided by its degrees of freedom. It is well known (we may use the Central Limit Theorem) that $\chi^2(n)/n$ converges in distribution to 1 as n tends to ∞. Hence we may recover the χ^2 distribution from the F distribution by letting either the numerator or denominator degrees of freedom tend to ∞. Setting $y=\infty$ in (4.13), we have the probability mass function

$$(5.1)\qquad {}^{c}_{q}\chi^2_{2x}(q^s)=\frac{c^x q^{\binom{x}{2}+sx}}{(-cq^s)_\infty}\frac{1}{(q)_{(x-1)}},$$

which provides a q-analogue of the χ^2 distribution with $2x$ degrees of freedom. Since

$$(5.2)\qquad {}^{c}_{q}F_{2x,2y}(q^s)={}^{c^{-1}q^{-(x+y)}}_{\qquad\quad q}F_{2y,2x}(q^{1-s}),$$

we obtain essentially the same q-analogue of the χ^2 distribution by taking $x=\infty$.

To obtain a q-analogue of a Poisson process with c events per unit time, we let the events occur independently at integer powers of q with

$$(5.3)\qquad \text{Prob}\,[\text{event at } q^t]=\frac{cq^t}{(1+cq^t)}.$$

The following theorem interprets our $q-\chi^2$ distribution probabilistically.

THEOREM 3.

$$(5.4)\qquad {}^{c}_{q}\chi^2_{2x}(q^t)=\text{Prob}\,[x\textit{th event at } q^t].$$

Proof. We may, of course, use (1.8) to sum over all of the possible occurrences of the first $x-1$ events. We proceed by induction on x. For $x=1$, we have

$$(5.5)\qquad \begin{aligned}\text{Prob}\,[\text{first event at } q^t]&=\text{Prob}\,[\text{event at } q^t]\prod_{n=t+1}^{\infty}\text{Prob}\,[\text{no event at } q^n]\\ &=\frac{cq^t}{(1+cq^t)}\prod_{n=t+1}^{\infty}\frac{1}{(1+cq^n)}=\frac{cq^t}{(-cq^t)_\infty},\end{aligned}$$

as required. We have

$$(5.6)\qquad \begin{aligned}&\text{Prob}\,[x\text{th event at } q^t]\\ &\quad=\text{Prob}\,[\text{event at } q^t]\sum_{n=t+1}^{\infty}\text{Prob}\,[(x-1)\text{st event at } q^n]\\ &\qquad\cdot\prod_{j=t+1}^{n-1}\text{Prob}\,[\text{no event at } q^j]\\ &\quad=\frac{cq^t}{(1+cq^t)}\sum_{n=t+1}^{\infty}\frac{c^{(x-1)}q^{\binom{x-1}{2}+n(x-1)}}{(-cq^n)_\infty}\frac{1}{(q)_{(x-2)}}\prod_{j=t+1}^{n-1}\frac{1}{(1+cq^j)}\\ &\quad=\frac{c^x q^{\binom{x-1}{2}+t}}{(-cq^t)_\infty}\frac{1}{(q)_{(x-2)}}\sum_{n=t+1}^{\infty}q^{n(x-1)}.\end{aligned}$$

The result (5.4) follows using the well-known sum (put $a=q$ in the q-binomial theorem (1.2)) of a geometric series. □

The limiting case $b=0$ of Ramanujan's sum (1.4) is

$$(5.7)\qquad \sum_{n=-\infty}^{\infty}(a)_n t^n=\frac{(at)_\infty(qa^{-1}t^{-1})_\infty(q)_\infty}{(t)_\infty(qa^{-1})_\infty},\qquad |t|<1.$$

Since ${}^{c}_{q}\chi^2_{2x}(q^t)$ sums to 1, (5.7) holds for $a=-c$, $t=q^x$. By Theorem 1, it holds for $t=q^x$, where x is a positive integer. Observe that we have only used the sum of a geometric series to obtain this result.

The χ^2 distribution may be obtained by rescaling the beta distribution on $(0, y)$ and letting y tend to ∞. We find that

$$(5.8)\qquad {}_qbe_{2y,\infty}(q^n)=(q^y)_\infty\frac{q^{yn}}{(q)_n},\qquad n\geqq 0,$$

is also a q-analogue of the χ^2 distribution. We close by showing that our q-F distribution also arises as a ratio of (5.1) and (5.8).

THEOREM 4. *If q^t has the $q-\chi^2$ distribution* (5.1) *and q^n has the distribution* (5.8), *then the ratio q^{t-n} has the q-F distribution* (4.13).

Proof. A simple application of the q-binomial theorem (1.2) gives

$$(5.9)\qquad \begin{aligned}\operatorname{Prob}[q^{t-n}=q^s]&=\sum_{n=0}^{\infty}{}_q^c\chi^2_{2x}(q^{s+n})\,{}_qf_{2y}(q^n)\\&=\sum_{n=0}^{\infty}\frac{c^xq^{\binom{x}{2}+(s+n)x}}{(-cq^{s+n})_\infty}\frac{1}{(q)_{(x-1)}}(q^y)_\infty\frac{q^{yn}}{(q)_n}\\&=\frac{c^xq^{\binom{x}{2}+sx}}{(-cq^s)_\infty}\frac{(q^y)_\infty}{(q)_{(x-1)}}\sum_{n=0}^{\infty}\frac{(-cq^s)_n}{(q)_n}q^{n(x+y)}\\&=\frac{c^xq^{\binom{x}{2}+sx}}{(-cq^s)_{(x+y)}}\frac{(q)_{(x+y-1)}}{(q)_{(x-1)}(q)_{(y-1)}},\end{aligned}$$

as required. □

REFERENCES

[1] G. E. ANDREWS, *On Ramanujan's summation of* ${}_1\psi_1(a; b, z)$, Proc. Amer. Math. Soc., 22 (1969), pp. 552–553.

[2] ———, *On a transformation of bilateral series with applications*, Proc. Amer. Math. Soc., 25 (1970), pp. 554–558.

[3] ———, *The theory of partitions*, in Encyclopedia of Mathematics and Its Applications, Vol. 2, Addison-Wesley, Reading, MA, 1976.

[4] G. E. ANDREWS AND R. ASKEY, *A simple proof of Ramanujan's summation of the* ${}_1\psi_1$, Aequationes Math., 18 (1978), pp. 333–337.

[5] R. ASKEY, *The q-gamma and q-beta functions*, Applicable Anal., 8 (1978), pp. 125–141.

[6] ———, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.

[7] R. ASKEY AND M. ISMAIL, *The very well poised* ${}_6\psi_6$, Proc. Amer. Math. Soc., 77 (1979), pp. 218–222.

[8] W. FELLER, *An Introduction to Probability Theory and its Applications, Vol.* 2, John Wiley, New York, 1971.

[9] R. A. FISHER, *On the similarity of the distribution found for the test of significance in harmonic analysis, and in Steven's problem in geometric probability*, Ann. Eugenics, 10 (1940), pp. 14–17.

[10] C. F. GAUSS, *Werke, Vol.* 2, Konigliche Gesellschaft der Wissenschaften, Gottingen, 1863.

[11] W. HAHN, *Beitrage zur Theorie der Heineschen Reihen*, Math. Nachr., 2 (1949), pp. 340–379.

[12] M. E.-H. ISMAIL, *A simple proof of Ramanujan's* ${}_1\psi_1$ *sum*, Proc. Amer. Math. Soc., 63 (1977), pp. 185–186.

[13] F. H. JACKSON, *On q-definite integrals*, Quart. J. Pure and Appl. Math., 41 (1910), pp. 193–203.

[14] M. JACKSON, *On Lerch's transcendent and the basic bilateral hypergeometric series* ${}_2\psi_2$, J. London Math. Soc., 25 (1950), pp. 189–196.

[15] M. G. KENDALL AND A. STUART, *The Advanced Theory of Statistics, Vol.* 2, Hafner, New York, 1973.

[16] G. POLYA, *On the number of certain lattice polygons*, J. Combin. Theory, 6 (1969), pp. 102–105.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

004

NEW INEQUALITIES FOR THE ZEROS OF JACOBI POLYNOMIALS*

LUIGI GATTESCHI†

Abstract. It is shown that certain asymptotic approximations are upper or lower bounds for the zeros $\theta_{n,k}(\alpha,\beta)$ of Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$. The procedure for deriving these bounds is based on the Sturm comparison theorem. Numerical examples are given to illustrate the sharpness of the new inequalities.

Key words. Jacobi polynomials, zeros, inequalities, Sturm comparison theorem

AMS(MOS) subject classifications. Primary 33A65; secondary 34C10, 64D20

1. Introduction. The purpose of this paper is to show that certain asymptotic approximations for the zeros of Jacobi polynomials are in fact inequalities.

The main tool that we need is the well-known Sturm comparison theorem in the following form given by Szegö [9, p. 19].

THEOREM 1.1 (Sturm's comparison theorem). *Let $f(x)$ and $F(x)$ be functions continuous in $x_0<x<X_0$ with $f(x)\leqq F(x)$. Let the functions $y(x)$ and $Y(x)$, both not identically zero, satisfy the differential equations*

$$y''+f(x)y=0, \qquad Y''+F(x)Y=0, \tag{1.1}$$

respectively. Let x' and x'', $x'<x''$, be two consecutive zeros of $y(x)$. Then the function $Y(x)$ has at least one zero in the interval $x'<x<x''$ provided $f(x)\not\equiv F(x)$ in $[x',x'']$.

The statement also holds for $x'=x_0[y(x_0+0)=0]$ if the additional condition

$$\lim_{x\to x_0+0}[y'(x)Y(x)-y(x)Y'(x)]=0 \tag{1.2}$$

is satisfied (similarly for $x''=X_0$).

Throughout this paper we denote by $x_{n,k}\equiv x_{n,k}(\alpha,\beta)$ the zeros, in decreasing order, of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$:

$$1>x_{n,1}>x_{n,2}>\cdots>x_{n,n}>-1,$$

and by $\theta_{n,k}\equiv\theta_{n,k}(\alpha,\beta)$,

$$0<\theta_{n,1}<\theta_{n,2}<\cdots<\theta_{n,n}<\pi,$$

the corresponding zeros of $P_n^{(\alpha,\beta)}(\cos\theta)$.

Let us recall the following asymptotic results and inequalities.

THEOREM 1.2 (Frenzen and Wong [3], [4]). *Let $\alpha>-\frac{1}{2}$, $\alpha+\beta\geqq-1$. Then as $n\to\infty$*

$$\theta_{n,k}=\frac{j_{\alpha,k}}{N}-\frac{1}{4N^2}\left[\left(\frac{1}{4}-\alpha^2\right)\left(\frac{2}{t}-\cot\frac{t}{2}\right)+\left(\frac{1}{4}-\beta^2\right)\tan\frac{t}{2}\right]+t^2O(n^{-3}) \tag{1.3}$$

where

$$N=n+\frac{\alpha+\beta+1}{2}, \tag{1.4}$$

$j_{\alpha,k}$ is the kth positive zero of the Bessel function $J_\alpha(x)$ and $t=j_{\alpha,k}/N$. The O-term is uniformly bounded for all values of $k=1,2,\cdots,[\gamma n]$, where $\gamma\in(0,1)$.

* Received by the editors April 2, 1986; accepted for publication (in revised form) September 17, 1986. This work was supported by the Consiglio Nazionale delle Ricerche of Italy and by the Ministero della Pubblica Istruzione of Italy.

† Dipartimento di Matematica dell'Università, Torino, Italy.

THEOREM 1.3 (Gatteschi and Pittaluga [6], [7]). *Let $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$. Then for all the zeros $\theta_{n,k}$ belonging to the interval $a<\theta<b$, with $0<a<b<\pi$, the following asymptotic expansion holds as $n\to\infty$:*

$$\theta_{n,k}=\phi_{n,k}+\frac{1}{4N^2}\left[\left(\frac{1}{4}-\alpha^2\right)\cot\frac{\phi_{n,k}}{2}-\left(\frac{1}{4}-\beta^2\right)\tan\frac{\phi_{n,k}}{2}\right]+O(n^{-4}) \tag{1.5}$$

where

$$\phi_{n,k}=\frac{2k+\alpha-1/2}{N}\frac{\pi}{2},$$

and N has the same meaning as in the previous theorem.

THEOREM 1.4 (Buell's inequalities; see Szegö [9, p. 125]). *Under the conditions $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$ and excluding the case $\alpha^2=\beta^2=\frac{1}{4}$, we have*

$$\frac{k+(\alpha+\beta-1)/2}{N}\pi<\theta_{n,k}<\frac{k}{N}\pi, \qquad k=1,2,\cdots,n, \tag{1.6}$$

whereas in the ultraspherical case $\alpha=\beta$

$$\theta_{n,k}>\frac{k+\alpha/2-1/4}{N}\pi, \qquad k=1,2,\cdots,\left[\frac{n}{2}\right]. \tag{1.7}$$

When $\alpha=\beta=-\frac{1}{2}$, $\alpha=\beta=\frac{1}{2}$, $\alpha=-\beta=-\frac{1}{2}$, $\alpha=-\beta=\frac{1}{2}$, we notice that

$$\theta_{n,k}=\frac{k-1/2}{n}\pi,\quad \frac{k}{n+1}\pi,\quad \frac{k-1/2}{n+1/2}\pi,\quad \frac{k}{n+1/2}\pi,$$

respectively.

THEOREM 1.5 (Gatteschi [5], [6]). *Let $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$. Then*

$$\frac{j_{\alpha,k}}{\nu^*}\leqq\theta_{n,k}\leqq\frac{j_{\alpha,k}}{\nu}, \qquad k=1,2,\cdots,\left[\frac{n}{2}\right] \tag{1.8}$$

where $j_{\alpha,k}$, as in Theorem 1.2, is the kth positive zero of $J_\alpha(x)$ and

$$\begin{aligned}\nu&=\left[N^2+\frac{1-\alpha^2-3\beta^2}{12}\right]^{1/2},\\ \nu^*&=\left[N^2+\frac{1}{4}-\frac{\alpha^2+\beta^2}{2}-\frac{1-4\alpha^2}{\pi^2}\right]^{1/2}.\end{aligned} \tag{1.9}$$

We remark (see Gatteschi [6, Thm. 3.1]) that the inequality

$$\theta_{n,k}\leqq\frac{j_{\alpha,k}}{\nu}, \tag{1.10}$$

under the same conditions for the parameters α and β, holds for all the zeros $\theta_{n,k}$ located in the interval $0<\theta<\pi$.

For the ultraspherical case $\alpha=\beta$, using Sturm's theorem, Ahmed, Muldoon and Spigler [2] have recently obtained the following interesting inequalities involving only elementary functions.

THEOREM 1.6. *Let $\alpha=\beta$ and let $x_{n,k}(\alpha)$, $k=1,2,\cdots,[n/2]$, be the zeros of the ultraspherical polynomial $P_n^{(\alpha,\alpha)}(x)$. Then, for $-\frac{1}{2}<\alpha<\frac{1}{2}$ and $k=1,2,\cdots,[n/2]$, we have*

$$x_{n,k}(\alpha)<\left[\frac{2n^2+4n+3}{2n^2+1+(2\alpha+1)(2n+1)}\right]^{1/2}\cos\frac{k\pi}{n+1}, \tag{1.11}$$

and

$$x_{n,k}(\alpha) > \left[\frac{2n^2+1}{2n^2+1+(2\alpha+1)(2n+1)}\right]^{1/2} \cos\frac{2k-1}{2n}\pi. \tag{1.12}$$

Ahmed, Muldoon and Spigler [2] have considered also values of α outside the range $-\frac{1}{2} < \alpha < \frac{1}{2}$, but the method used cannot be applied to obtain similar results for the general Jacobi case [8].

2. A lower bound for the zeros of $P_n^{(\alpha,\beta)}(\cos\theta)$. We shall assume throughout this paper that the parameters α and β satisfy the inequalities

$$-\tfrac{1}{2} \leqq \alpha \leqq \tfrac{1}{2}, \qquad -\tfrac{1}{2} \leqq \beta \leqq \tfrac{1}{2}. \tag{2.1.}$$

Moreover, we shall refer to the differential equation

$$\frac{d^2u}{d\theta^2} + \left[N^2 + \frac{1/4-\alpha^2}{4\sin^2\theta/2} + \frac{1/4-\beta^2}{4\cos^2\theta/2}\right]u = 0, \qquad N = n + \frac{\alpha+\beta+1}{2} \tag{2.2}$$

which is satisfied by

$$u(\theta) = \left(\sin\frac{\theta}{2}\right)^{\alpha+1/2}\left(\cos\frac{\theta}{2}\right)^{\beta+1/2} P_n^{(\alpha,\beta)}(\cos\theta). \tag{2.3}$$

Now we observe that the function

$$z(\theta) = \left(\frac{f}{f'}\right)^{1/2} J_\alpha[f(\theta)] \tag{2.4}$$

satisfies the differential equation

$$\frac{d^2z}{d\theta^2} + F(\theta)z = 0 \tag{2.5}$$

where

$$F(\theta) = \frac{1}{2}\frac{f'''}{f'} - \frac{3}{4}\left(\frac{f''}{f'}\right)^2 + \left(\frac{1}{4}-\alpha^2\right)\left(\frac{f'}{f}\right)^2 + f'^2. \tag{2.6}$$

In this section we assume that

$$f(\theta) = N\theta + \frac{1}{4N}\left[\left(\frac{1}{4}-\alpha^2\right)\left(\frac{2}{\theta} - \cot\frac{\theta}{2}\right) + \left(\frac{1}{4}-\beta^2\right)\tan\frac{\theta}{2}\right] \tag{2.7}$$

and make use of (2.5) as a comparison equation to derive, by means of Sturm's method, inequalities for the zeros $\theta_{n,k}(\alpha,\beta)$, $k = 1, 2, \cdots, n$, of $P_n^{(\alpha,\beta)}(\cos\theta)$.

LEMMA 2.1. *Let α and β satisfy* (2.1). *The function $F(\theta)$ defined by* (2.6) *and* (2.7) *is such that*

$$F(\theta) \geqq N^2 + \frac{1/4-\alpha^2}{4\sin^2\theta/2} + \frac{1/4-\beta^2}{4\cos^2\theta/2} \tag{2.8}$$

for $0 < \theta < \pi$. Here the equality sign holds if and only if $\alpha^2 = \beta^2 = \frac{1}{4}$.

For the proof we put

$$A = \tfrac{1}{4} - \alpha^2, \qquad B = \tfrac{1}{4} - \beta^2,$$

$$a(\theta) = \frac{2}{\theta} - \cot\frac{\theta}{2}, \qquad b(\theta) = \tan\frac{\theta}{2},$$

and we observe that

$$f'^2(\theta)=N^2+\frac{A}{4\sin^2\theta/2}-\frac{A}{\theta^2}+\frac{B}{4\cos^2\theta/2}+\frac{1}{16N^2}(Aa'+Bb')^2.$$

Hence

$$F(\theta)-N^2-\frac{A}{4\sin^2\theta/2}-\frac{B}{4\cos^2\theta/2}$$
$$=\frac{1}{2f'}F_1(\theta)+\frac{A(f'\theta+f)}{f^2\theta^2}F_2(\theta)+\frac{1}{16N^2}(Aa'+Bb')^2$$

where

$$F_1(\theta)=f'''-\frac{3}{2}\frac{f''^2}{f'},\qquad F_2(\theta)=f'\theta-f,$$

and it suffices to show that $F_1(\theta)\geqq 0$ and $F_2(\theta)\geqq 0$.

To this end we note that $a(\theta)$ and $b(\theta)$ are positive increasing functions for $0<\theta<\pi$ as are all their derivatives. Moreover, the following expansions hold [1, p. 75]:

$$a(\theta)=\frac{\theta}{6}+\frac{\theta^3}{360}+\cdots+\frac{(-1)^{n-1}2B_{2n}}{(2n)!}\theta^{2n-1}+\cdots\qquad(|\theta|<2\pi),$$

$$b(\theta)=\frac{\theta}{2}+\frac{\theta^3}{24}+\cdots+\frac{(-1)^{n-1}2(2^{2n}-1)B_{2n}}{(2n)!}\theta^{2n-1}+\cdots\qquad(|\theta|<\pi)$$

where B_{2n} is the $2n$th Bernoulli number.

It is readily seen that $F_2(\theta)\geqq 0$ for $0<\theta<\pi$. Indeed, we have

$$F_2(\theta)=\frac{A}{4N}(a'\theta-a)+\frac{B}{4N}(b'\theta-b)\geqq 0.$$

To study $F_1(\theta)$ we first consider the interval $0<\theta\leqq\pi/2$ where we have

$$Aa'''+Bb'''\geqq\frac{A}{60}+\frac{B}{4},\qquad Aa''+Bb''\leqq A\left(\frac{32}{\pi^3}-1\right)+B\leqq\frac{8}{\pi^3};$$

thus

$$F_1(\theta)=\frac{1}{4N}\left[Aa'''+Bb'''-\frac{3}{2}\frac{(Aa''+Bb'')^2}{4N^2+Aa'+Bb'}\right]$$
$$\geqq\frac{1}{4N}\left\{\frac{A}{60}+\frac{B}{4}-\frac{3}{N^2}\left[A\left(\frac{32}{\pi^3}-1\right)+B\right]\frac{1}{\pi^3}\right\}\geqq 0.$$

Similar considerations may be applied to

$$4N(4N^2+Aa'+Bb')F_1(\theta)$$

for the interval $\pi/2<\theta<\pi$. More precisely, it is straightforward to prove that

$$4N^2Aa'''-\frac{3}{2}A^2a''^2\geqq A\left[4N^2a'''\left(\frac{\pi}{2}\right)-\frac{3}{8}a''^2(\pi)\right]\geqq 0,$$

$$b'''b'-\frac{3}{2}b''^2=\frac{1}{8}\frac{1}{\cos^4\theta/2}>0,$$

and

$$a'b''' - 3a''b'' > \frac{1}{2\cos^2\theta/2}\left[\frac{1}{2}\left(1-\frac{8}{\pi^2}\right)\left(1+3\tan^2\frac{\theta}{2}\right)-\frac{12}{\pi^3}\tan\frac{\theta}{2}\right] > 0.$$

Having established both $F_1(\theta)$ and $F_2(\theta) \geqq 0$, Lemma 2.1 is proved and we can state the following main result of this section.

THEOREM 2.1. *Let* $-\frac{1}{2} \leqq \alpha \leqq \frac{1}{2}$, $-\frac{1}{2} \leqq \beta \leqq \frac{1}{2}$. *Then the Frenzen-Wong approximation for* $\theta_{n,k}(\alpha, \beta)$, *obtained by omitting the* $O(n^{-3})$ *term in* (1.3), *is in fact a lower bound*; *that is*

$$\theta_{n,k}(\alpha,\beta) \geqq \frac{j_{\alpha,k}}{N} - \frac{1}{4N^2}\left[\left(\frac{1}{4}-\alpha^2\right)\left(\frac{2}{t}-\cot\frac{t}{2}\right)+\left(\frac{1}{4}-\beta^2\right)\tan\frac{t}{2}\right] \tag{2.9}$$

where

$$N = n + \frac{\alpha+\beta+1}{2}, \qquad t = \frac{j_{\alpha,k}}{N}$$

and $k = 1, 2, \cdots, n$. *The equality sign in* (2.9) *holds when* $\alpha^2 = \beta^2 = \frac{1}{4}$.

Indeed, the validity of (2.8) enables us to apply Theorem 1.1 to the differential equations (2.2) and (2.5) and to compare the positive zeros of the function $z(\theta)$ defined by (2.4), i.e., the positive zeros $\tau_{n,k}$, $k = 1, 2, \cdots$,

$$\tau_{n,1} < \tau_{n,2} < \cdots < \tau_{n,n}$$

of the function $J_\alpha[f(\theta)]$, with the zeros

$$0, \theta_{n,1}, \theta_{n,2}, \cdots, \theta_{n,n}$$

of the function $u(\theta)$ defined by (2.3).

The condition (1.2), which is required when we apply Theorem 1.1 to the interval $[0, \theta_{n,1}]$, is satisfied. Hence, we conclude that

$$\tau_{n,k} \leqq \theta_{n,k}, \qquad k = 1, 2, \cdots, \tag{2.10}$$

where, since $f(\theta)$ is a continuous increasing function in $0 < \theta < \pi$,

$$N\tau_{n,k} + \frac{1}{4N}\left[\left(\frac{1}{4}-\alpha^2\right)\left(\frac{2}{\tau_{n,k}}-\cot\frac{\tau_{n,k}}{2}\right)+\left(\frac{1}{4}-\beta^2\right)\tan\frac{\tau_{n,k}}{2}\right] = j_{\alpha,k}. \tag{2.11}$$

We also have

$$j_{\alpha,k} \leqq f(\theta_{n,k}), \qquad k = 1, 2, \cdots, n.$$

Now, when we set

$$h(\theta) = \frac{j_{\alpha,k}}{N} - \frac{1}{4N^2}\left[\left(\frac{1}{4}-\alpha^2\right)\left(\frac{2}{\theta}-\cot\frac{\theta}{2}\right)+\left(\frac{1}{4}-\beta^2\right)\tan\frac{\theta}{2}\right],$$

it follows that the equation $\theta = h(\theta)$ has the solution $\theta = \tau_{n,k}$. Since the function $h(\theta)$ decreases and, from (1.10),

$$\theta_{n,k} \leqq \frac{j_{\alpha,k}}{\nu} \leqq \frac{j_{\alpha,k}}{N},$$

we obtain the following from (2.10):

$$\theta_{n,k} \geqq \tau_{n,k} = h(\tau_{n,k}) \geqq h(\theta_{n,k}) \geqq h\left(\frac{j_{\alpha,k}}{\nu}\right) \geqq h\left(\frac{j_{\alpha,k}}{N}\right),$$

which proves the theorem.

This theorem can be improved by using (1.10) and by observing that $\theta_{n,k} \geqq h(j_{\alpha,k}/\nu)$, $k = 1, 2, \cdots, n$.

THEOREM 2.2. *Let* $-\frac12 \leqq \alpha \leqq \frac12$, $-\frac12 \leqq \beta \leqq \frac12$. *Then for the zeros* $\theta_{n,k}(\alpha,\beta)$ *of* $P_n^{(\alpha,\beta)}(\cos\theta)$ *the following inequalities hold*:

$$\theta_{n,k}(\alpha,\beta) \geqq \frac{j_{\alpha,k}}{N} - \frac{1}{4N^2}\left[\left(\frac14-\alpha^2\right)\left(\frac{2}{\tau}-\cot\frac{\tau}{2}\right)+\left(\frac14-\beta^2\right)\tan\frac{\tau}{2}\right], \quad k=1,2,\cdots,n \tag{2.12}$$

where

$$N = n+\frac{\alpha+\beta+1}{2}, \quad \nu=\left(N^2+\frac{1-\alpha^2-3\beta^2}{12}\right)^{1/2}, \quad \tau=\frac{j_{\alpha,k}}{\nu}.$$

As before, the equality sign in (2.12) *holds when* $\alpha^2=\beta^2=\frac14$.

Using this result and the property

$$P_n^{(\alpha,\beta)}(\cos\theta) = (-1)^n P_n^{(\beta,\alpha)}[\cos(\pi-\theta)], \tag{2.13}$$

we can obtain upper bounds for $\theta_{n,k}(\alpha,\beta)$, $k=1,2,\cdots,n$.

COROLLARY 2.1. *Let* $-\frac12 \leqq \alpha \leqq \frac12$, $-\frac12 \leqq \beta \leqq \frac12$. *Then*

$$\begin{aligned}\theta_{n,k}(\alpha,\beta) &= \pi - \theta_{n,n-k+1}(\beta,\alpha)\\ &\leqq \pi - \frac{j_{\beta,n-k+1}}{N} + \frac{1}{4N^2}\left[\left(\frac14-\beta^2\right)\left(\frac{2}{\bar\tau}-\cot\frac{\bar\tau}{2}\right)+\left(\frac14-\alpha^2\right)\tan\frac{\bar\tau}{2}\right],\end{aligned} \tag{2.14}$$

$$k=1,2,\cdots,n$$

where

$$N = n+\frac{\alpha+\beta+1}{2}, \quad \bar\nu=\left(N^2+\frac{1-\beta^2-3\alpha^2}{12}\right)^{1/2}, \quad \bar\tau=\frac{j_{\beta,n-k+1}}{\bar\nu},$$

the equality sign holding when $\alpha^2=\beta^2=\frac14$.

Inequalities (2.12) and (2.14) give very sharp results for the zeros which are close to π, when we apply (2.12), or close to zero, when we apply (2.14). Table 1 provides lower and upper bounds, obtained using (2.12) and (2.14), in the case $n=10$, $\alpha=\frac13$ and $\beta=0$.

TABLE 1
Zeros of $P_{10}^{(1/3,0)}(\cos\theta)$.

k	Lower bound	Exact value	Upper bound
1	0.2720 2843	0.2720 2854	0.2721 5052
2	0.5653 8156	0.5653 8183	0.5653 9792
3	0.8594 3899	0.8594 3944	0.8594 4435
4	1.1536 6321	1.1536 6395	1.1536 6610
5	1.4479 3628	1.4479 3750	1.4479 3864
6	1.7422 0569	1.7422 0784	1.7422 0853
7	2.0364 2307	2.0364 2737	2.0364 2780
8	2.3305 0083	2.3305 1150	2.3305 1177
9	2.6241 7391	1.6242 1351	2.6242 1367
10	2.9157 9656	2.9161 9539	2.9161 9545

Should we be interested in obtaining sharper numerical results, it is convenient to use the inequalities (2.12) and (2.14) jointly with the following simpler ones

$$\begin{aligned}&\theta_{n,k}(\alpha,\beta) \leqq j_{\alpha,k}\left(N^2+\frac{1-\alpha^2-3\beta^2}{12}\right)^{-1/2},\\ &\theta_{n,k}(\alpha,\beta) \geqq \pi - j_{\beta,n-k+1}\left(N^2+\frac{1-\beta^2-3\alpha^2}{12}\right)^{-1/2} \quad (k=1,2,\cdots,n)\end{aligned} \tag{2.15}$$

obtained by means of the inequality (1.10) and the property (2.13). Indeed, the first of (2.15) may be better than (2.14) for the first few values of k, while the second of (2.15) may be better than (2.12) for the last few values of k. Thus, for example, by using (2.15) we obtain

$$\theta_{10,1}(\tfrac{1}{3},0)<0.2720\,2892, \qquad \theta_{10,2}(\tfrac{1}{3},0)<0.5653\,8602,$$

$$\theta_{10,8}(\tfrac{1}{3},0)>2.3305\,0366, \qquad \theta_{10,9}(\tfrac{1}{3},0)>2.6242\,1163,$$

$$\theta_{10,10}(\tfrac{1}{3},0)>2.9161\,9528,$$

which are better than those given in Table 1.

The sharpness of the results furnished by applying inequalities (2.12) and (2.14) together with (2.15) is shown in Fig. 1 for $n=16$, $\alpha=-\frac{1}{3}$, $\beta=\frac{1}{3}$ and $\alpha=0$, $\beta=\frac{1}{4}$. More precisely, in the Fig. 1 we can see the digits of accuracy

$$\rho_n(\alpha,\beta;k)=-\log_{10}\frac{U_k-L_k}{U_k},$$

obtained by means of the upper bound U_k and the lower bound L_k of the kth zero $\theta_{n,k}(\alpha,\beta)$.

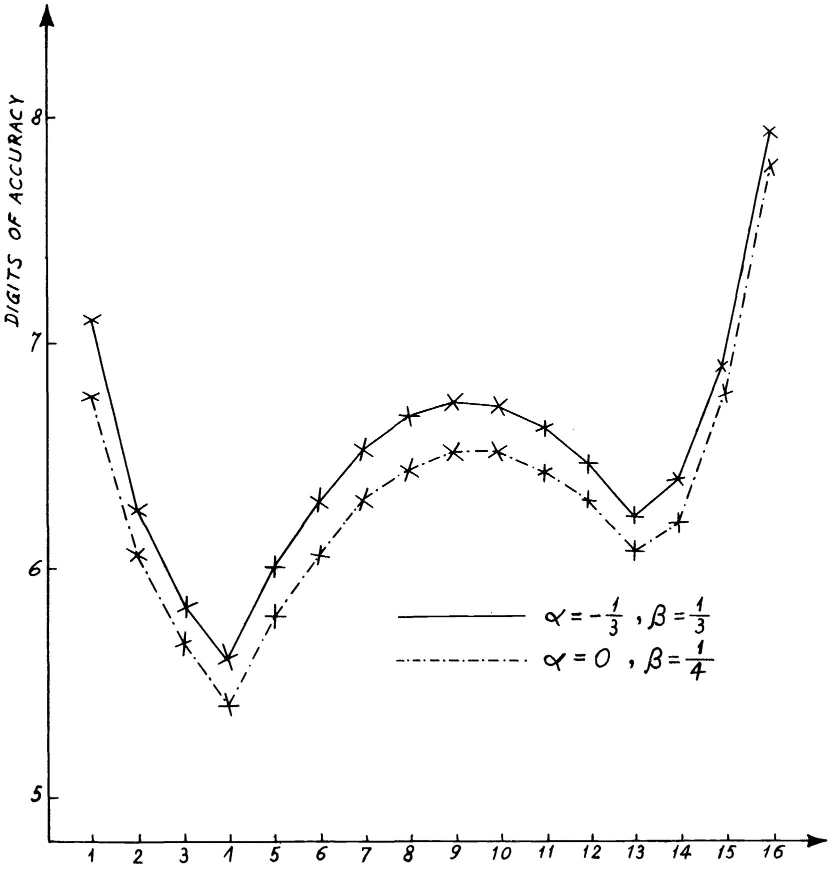


FIG. 1. $\rho_{16}(\alpha,\beta;k)$ *versus* $k=1(1)16$.

In the ultraspherical case $\alpha=\beta$ Theorem 2.2 and the first of inequalities (2.15) yield the following corollary.

COROLLARY 2.2. *Let $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$ and let $\theta_{n,k}(\alpha)$ be the zeros of the ultraspherical polynomial $P_n^{(\alpha,\alpha)}(\cos\theta)$. Then*

$$\frac{j_{\alpha,k}}{N}-\frac{1-4\alpha^2}{8N^2}\left(\frac{\nu}{j_{\alpha,k}}-\cot\frac{j_{\alpha,k}}{\nu}\right)\leqq\theta_{n,k}(\alpha)\leqq\frac{j_{\alpha,k}}{\nu}\tag{2.16}$$

where

$$N=n+\alpha+\frac{1}{2},\quad \nu=\left(N^2+\frac{1-4\alpha^2}{12}\right)^{1/2},\quad k=1,2,\cdots,n.$$

The equality signs in (2.16) *hold when* $|\alpha|=\frac{1}{2}$.

3. Other separation results for the zeros of $P_n^{(\alpha,\beta)}(\cos\theta)$ and upper bounds in the ultraspherical case. We now use the differential equation

$$\frac{d^2z}{d\theta^2}+G(\theta)z=0,\tag{3.1}$$

with

$$G(\theta)=\frac{1}{2}\frac{g'''}{g'}-\frac{3}{4}\left(\frac{g''}{g'}\right)^2+g'^2,$$

satisfied by

$$z=(g')^{-1/2}\cos g(\theta).$$

Further, we assume that

$$g(\theta)=N\theta-\left(\alpha+\frac{1}{2}\right)\frac{\pi}{2}-\frac{1}{4N}\left[\left(\frac{1}{4}-\alpha^2\right)\cot\frac{\theta}{2}-\left(\frac{1}{4}-\beta^2\right)\tan\frac{\theta}{2}\right],\tag{3.2}$$

suggested by (1.5), and we show that

$$G(\theta)\geqq N^2+\frac{1/4-\alpha^2}{4\sin^2\theta/2}+\frac{1/4-\beta^2}{4\cos^2\theta/2},\qquad -\frac{1}{2}\leqq\alpha,\quad\beta\leqq\frac{1}{2},\tag{3.3}$$

with equality sign when $\alpha^2=\beta^2=\frac{1}{4}$.

For the proof we set

$$A=\tfrac{1}{4}-\alpha^2,\qquad B=\tfrac{1}{4}-\beta^2$$

and we study the sign of

$$\frac{1}{2}\frac{g'''}{g'}-\frac{3}{4}\left(\frac{g''}{g'}\right)^2+g'^2-N^2-\frac{A}{4\sin^2\theta/2}-\frac{B}{4\cos^2\theta/2},$$

when $A\geqq0$, $B\geqq0$, $g(\theta)$ is defined by (3.2) and $0<\theta<\pi$.

We find that

$$g'^2-N^2-\frac{A}{4\sin^2\theta/2}-\frac{B}{4\cos^2\theta/2}=\frac{1}{64N^2}\left(\frac{A}{\sin^2\theta/2}+\frac{B}{\cos^2\theta/2}\right)^2\geqq0$$

and

$$\frac{1}{2}\frac{g'''}{g'}-\frac{3}{4}\left(\frac{g''}{g'}\right)^2=\frac{1}{2g'}\left(g'''-\frac{3}{2}\frac{g''^2}{g'}\right)$$

$$=\frac{1}{32Ng'}\left[\frac{A}{\sin^4\theta/2}r(\theta)+\frac{B}{\cos^4\theta/2}s(\theta)+6ABt(\theta)\right]$$

where

$$t(\theta)=(8N^2\sin^2\theta/2\cos^2\theta/2+A\cos^2\theta/2+B\sin^2\theta/2)^{-1},$$
$$r(\theta)=3-2\sin^2\theta/2-3At(\theta)\cos^4\theta/2,$$
$$s(\theta)=3-2\cos^2\theta/2-3Bt(\theta)\sin^4\theta/2.$$

When we observe that $t(\theta)>0$ and that

$$r(\theta)>3-2\sin^2\theta/2-3\cos^2\theta/2=\sin^2\theta/2,$$
$$s(\theta)>3-2\cos^2\theta/2-3\sin^2\theta/2=\cos^2\theta/2,$$

the proof of (3.3) immediately follows.

The application of Sturm's method to the differential equations (3.1) and (2.2) requires an accurate study of the distribution of the zeros $\psi_{n,k}\equiv\psi_{n,k}(\alpha,\beta)$ of the function

$$z_n^{(\alpha,\beta)}(\theta)=[g'(\theta)]^{-1/2}\cos g(\theta). \tag{3.4}$$

If we exclude not only the case $\alpha^2=\beta^2=\frac{1}{4}$, but also the cases in which only one of the two parameters α and β is $\pm\frac{1}{2}$, it is easy to see that $z_n^{(\alpha,\beta)}(\theta)$ has infinitely many zeros

$$\cdots<\psi_{n,-2}<\psi_{n,-1}<\psi_{n,0}<\psi_{n,1}<\psi_{n,2}<\cdots, \tag{3.5}$$

lying in the interval $0<\theta<\pi$. These zeros can be obtained by solving the equations

$$\begin{gathered}N\theta-\left(\alpha+\frac{1}{2}\right)\frac{\pi}{2}-\frac{1}{4N}\left(A\cot\frac{\theta}{2}-B\tan\frac{\theta}{2}\right)=(2k-1)\frac{\pi}{2},\\ A=\tfrac{1}{4}-\alpha^2,\quad B=\tfrac{1}{4}-\beta^2,\quad k=0,\pm1,\pm2,\cdots,\end{gathered} \tag{3.6}$$

with respect to θ; that is they are (see Fig. 2) the abscissae of the intersections of the

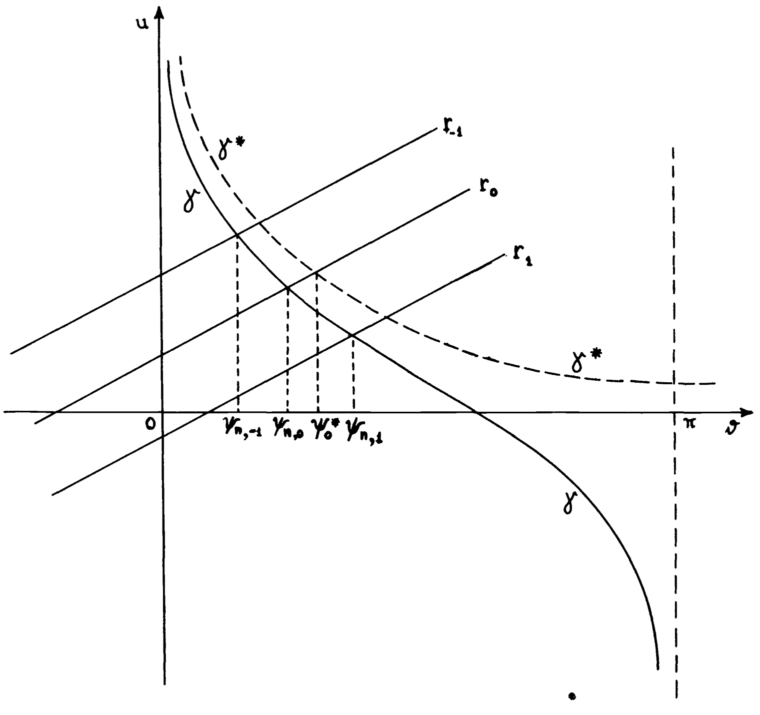


FIG. 2

straight lines r_k

$$u = N\theta - \left(2k + \alpha - \frac{1}{2}\right)\frac{\pi}{2}, \qquad k = 0, \pm 1, \pm 2, \cdots,$$

with the curve γ

$$u = \frac{1}{4N}\left(A \cot\frac{\theta}{2} - B \tan\frac{\theta}{2}\right), \qquad 0 < \theta < \pi.$$

LEMMA 3.1. *The function $z_n^{(\alpha,\beta)}(\theta)$ satisfies the identity*

$$z_n^{(\alpha,\beta)}(\pi - \theta) = (-1)^n z_n^{(\beta,\alpha)}(\theta). \tag{3.7}$$

Consequently, for the zeros $\psi_{n,k}(\alpha, \beta)$ we have

$$\psi_{n,k}(\alpha, \beta) = \pi - \psi_{n,n-k+1}(\beta, \alpha), \qquad k = 0, \pm 1, \pm 2, \cdots. \tag{3.8}$$

The proof is readily obtained by means of (3.4).

LEMMA 3.2. *The zeros $\psi_{n,1}, \psi_{n,2}, \cdots, \psi_{n,n}$* of *$z_n^{(\alpha,\beta)}(\theta)$ belong to the same interval*

$$\frac{1 + (\alpha + \beta - 1)/2}{N}\pi < \theta < \frac{n}{N}\pi, \tag{3.9}$$

as do (*in view of the inequalities of Buell* (1.6)) *the zeros $\theta_{n,k}$ of $P_n^{(\alpha,\beta)}(\cos\theta)$, while the other zeros of $z_n^{(\alpha,\beta)}(\theta)$, i.e., $\cdots < \psi_{n,-2} < \psi_{n,-1} < \psi_{n,0}$ and $\psi_{n,n+1} < \psi_{n,n+2} < \cdots$, lie outside this interval* (3.9).

For the proof we first observe that

$$A \cot\frac{\theta}{2} - B \tan\frac{\theta}{2} < A\frac{2}{\theta}.$$

Then, if ψ_0^* denotes the abscissa (see Fig. 2) of the intersection of the straight line r_0 with the curve γ^*

$$u = \frac{A}{2N\theta}, \qquad \theta > 0,$$

i.e.,

$$\psi_0^* = \frac{1}{2N}\left[\left(\alpha - \frac{1}{2}\right)\frac{\pi}{2} + \sqrt{\left(\alpha - \frac{1}{2}\right)^2\frac{\pi^2}{4} + 2A}\right],$$

it is easy to see that, for $-\frac{1}{2} < \alpha < \frac{1}{2}$ and $-\frac{1}{2} < \beta < \frac{1}{2}$,

$$\psi_{n,0} < \psi_0^* < \frac{\alpha + \beta + 1}{2N}\pi < \psi_{n,1}.$$

By applying Lemma 3.1 we have analogously

$$\psi_{n,n} < \frac{n}{N}\pi < \psi_{n,n+1}.$$

The above results can be extended to the cases $\alpha = \pm\frac{1}{2}$ or $\beta = \pm\frac{1}{2}$ taking into account that, instead of the sequence (3.5), we have one of the two sequences

$$\psi_{n,1} < \psi_{n,2} < \cdots < \psi_{n,n} < \psi_{n,n+1} < \cdots, \text{ or}$$

$$\cdots < \psi_{n,-2} < \psi_{n,-1} < \psi_{n,0} < \psi_{n,1} < \cdots < \psi_{n,n}.$$

We now apply Theorem 1.1 to the differential equations (2.2) and (3.1) relatively to the interval (3.9). Inequality (3.3) readily furnishes the following separation result.

THEOREM 3.1. *Let $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$. Then, between any two consecutive zeros $\theta_{n,h}(\alpha,\beta)$, $h=1,2,\cdots,n$, of the Jacobi polynomial $P_n^{(\alpha,\beta)}(\cos\theta)$ there is at least one zero $\psi_{n,k}(\alpha,\beta)$, $k=1,2,\cdots,n$, of the function $z_n^{(\alpha,\beta)}(\theta)$ defined by* (3.4).

We notice that this theorem does not allow us to derive inequalities satisfied by the zeros $\theta_{n,k}$ for general α and β, $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$ and $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$. Indeed, it assures us only that if $\psi_{n,1}\leqq\theta_{n,1}$ or $\psi_{n,n}\geqq\theta_{n,n}$ then, for each $k=1,2,\cdots,n-1$, either $\theta_{n,k}<\psi_{n,k+1}<\theta_{n,k+1}$ or $\theta_{n,k}<\psi_{n,k}<\theta_{n,k+1}$; otherwise there is one of the intervals $[\theta_{n,h},\theta_{n,h+1}]$, $h=1,2,\cdots,n-1$, which contains two ψ-zeros and all the others contain exactly one ψ-zero in their interior. A more precise result can be obtained for the ultraspherical case $\alpha=\beta$.

THEOREM 3.2. *Let $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$ and let $\theta_{n,k}(\alpha)\equiv\theta_{n,k}(\alpha,\alpha)$, $k=1,2,\cdots,n$, be the zeros of the ultraspherical polynomial $P_n^{(\alpha,\alpha)}(\cos\theta)$. Then*

$$\theta_{n,k}(\alpha)\leqq\psi_{n,k}(\alpha),\qquad k=1,2,\cdots,\left[\frac{n}{2}\right]\tag{3.10}$$

where $\psi_{n,k}(\alpha)$ is the root lying in $0<\theta<\pi/2$ of the equation

$$N\theta-\frac{1}{2N}\left(\frac{1}{4}-\alpha^2\right)\cot\theta=\left(2k+\alpha-\frac{1}{2}\right)\frac{\pi}{2},\tag{3.11}$$

with $N=n+\alpha+\frac{1}{2}$. The equality sign in (3.10) *holds only when $\alpha^2=\frac{1}{4}$.*

For the proof we exclude the trivial case $\alpha^2=\frac{1}{4}$ and observe that, according to Lemma 3.1, the zeros $\psi_{n,k}(\alpha)$ are symmetric with respect to $\pi/2$ and that the same property holds for the zeros $\theta_{n,k}(\alpha)$. Hence for n even, the "central" interval

$$\theta_{n,n/2}(\alpha)<\theta<\theta_{n,n/2+1}(\alpha),$$

must contain, by virtue of Theorem 3.1, exactly the two zeros $\psi_{n,n/2}(\alpha)$ and $\psi_{n,n/2+1}(\alpha)$, and consequently each of the other $n-2$ intervals

$$\theta_{n,k}(\alpha)<\theta<\theta_{n,k+1}(\alpha),\qquad k=1,2,\cdots,n-1,\quad k\neq\frac{n}{2},$$

contains exactly one zero $\psi_{n,k}(\alpha)$. For n odd, we have $\psi_{n,(n+1)/2}(\alpha)=\theta_{n,(n+1)/2}(\alpha)=\pi/2$ and each interval

$$\theta_{n,k}(\alpha)<\theta<\theta_{n,k+1}(\alpha),\qquad k=1,2,\cdots,n-1,$$

contains exactly one zero $\psi_{n,k}(\alpha)$, $k\neq(n+1)/2$.

As a consequence of Theorem 3.2 we can obtain an exact statement concerning the location of the zeros of Jacobi polynomials with α or $\beta=\pm\frac{1}{2}$. Indeed, it is well known that the following formulas hold [9, p. 59]:

$$P_{2n}^{(\alpha,\alpha)}(\cos\theta)=\frac{\Gamma(2n+\alpha+1)\Gamma(n+1)}{\Gamma(n+\alpha+1)\Gamma(2n+1)}P_n^{(\alpha,-1/2)}(\cos 2\theta),$$

$$P_{2n+1}^{(\alpha,\alpha)}(\cos\theta)=\frac{\Gamma(2n+\alpha+2)\Gamma(n+1)}{\Gamma(n+\alpha+1)\Gamma(2n+2)}\cos\theta P_n^{(\alpha,1/2)}(\cos 2\theta).$$

Now, observe that similar relationships are valid for the function $z_n^{(\alpha,\beta)}(\theta)$ defined by (3.4). Specifically we have

$$z_{2n}^{(\alpha,\alpha)}(\theta)=2^{-1/2}z_n^{(\alpha,-1/2)}(2\theta),\qquad z_{2n+1}^{(\alpha,\alpha)}(\theta)=2^{-1/2}z_n^{(\alpha,1/2)}(2\theta)$$

where $0<\theta<\pi/2$. Hence, applying Theorem 3.2 and recalling Lemma 3.1, we obtain the following result.

COROLLARY 3.1. *Let* $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$. *Then*

$$\begin{aligned}\theta_{n,k}(\alpha, \pm\tfrac{1}{2})&\leqq\psi_{n,k}(\alpha, \pm\tfrac{1}{2}),\\ \theta_{n,k}(\pm\tfrac{1}{2}, \beta)&\geqq\psi_{n,k}(\pm\tfrac{1}{2}, \beta),\end{aligned}\qquad k=1,2,\cdots n, \tag{3.12}$$

where $\psi_{n,k}(\alpha, \pm\frac{1}{2})$ *and* $\psi_{n,k}(\pm\frac{1}{2}, \beta)$ *are the roots, in* $0<\theta<\pi$, *of the equations*

$$\left[n+(\alpha+1)/2\pm\frac{1}{4}\right]\theta-\frac{1/4-\alpha^2}{4[n+(\alpha+1)/2\pm1/4]}\cot\frac{\theta}{2}=\left(2k+\alpha-\frac{1}{2}\right)\frac{\pi}{2},$$

$$\left[n+(\beta+1)/2\pm\frac{1}{4}\right]\theta+\frac{1/4-\beta^2}{4[n+(\beta+1)/2\pm1/4]}\tan\frac{\theta}{2}=\left(2k\pm\frac{1}{2}-\frac{1}{2}\right)\frac{\pi}{2},$$

respectively. The equality signs in (3.12) *hold when* $\alpha=\pm\frac{1}{2}$ *or* $\beta=\pm\frac{1}{2}$.

Another interesting consequence of Theorem 3.2 can be obtained by observing that $\psi_{n,k}(\alpha)$ is the solution of the equation $\theta=h(\theta)$ in the interval $0<\theta<\pi/2$, where

$$h(\theta)=\frac{2k+\alpha-1/2}{N}\frac{\pi}{2}+\frac{1}{2N^2}\left(\frac{1}{4}-\alpha^2\right)\cot\theta,$$

and that, from (3.10) and (1.7),

$$\psi_{n,k}(\alpha)\geqq\theta_{n,k}(\alpha)\geqq\phi_{n,k}(\alpha)=\frac{2k+\alpha-1/2}{N}\frac{\pi}{2}.$$

Since $h(\theta)$ is monotonic decreasing, we have

$$\theta_{n,k}(\alpha)\leqq\psi_{n,k}(\alpha)=h(\psi_{n,k}(\alpha))\leqq h(\phi_{n,k}(\alpha));$$

that is, the following final result holds.

COROLLARY 3.2. *Let* $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$ *and let* $\theta_{n,k}(\alpha)$, $k=1,2,\cdots,[n/2]$, *be the zeros of* $P_n^{(\alpha,\alpha)}(\cos\theta)$. *Then*

$$\phi_{n,k}(\alpha)\leqq\theta_{n,k}(\alpha)\leqq\phi_{n,k}(\alpha)+\frac{1}{2N^2}\left(\frac{1}{4}-\alpha^2\right)\cot\phi_{n,k}(\alpha) \tag{3.13}$$

where

$$N=n+\alpha+\frac{1}{2},\qquad \phi_{n,k}(\alpha)=\frac{2k+\alpha-1/2}{N}\frac{\pi}{2}.$$

The equality signs in (3.13) *hold only when* $|\alpha|=\frac{1}{2}$.

This corollary establishes that the approximations obtained by omitting the O-term in the Gatteschi-Pittaluga formula (1.5) are upper bounds for $\theta_{n,k}(\alpha)$, $k=1,2,\cdots,[n/2]$, in the ultraspherical case.

These bounds are very sharp, except for the early values of k, and numerical comparisons have shown that they are better than those derived from the Ahmed, Muldoon and Spigler inequality (1.12).

Corollary 2.1 provides lower bounds of a nonelementary type, that is in terms of the zeros of the Bessel function $J_\alpha(x)$. For elementary lower bounds we can use (3.13), that is the Buell inequality (1.7). Moreover, we can use also, as soon as it is convenient, the inequality

$$\theta_{n,k}(\alpha) \geqq \arccos x^*_{n,k}(\alpha) \tag{3.14}$$

where $x^*_{n,k}(\alpha)$ is defined by

$$x^*_{n,k}(\alpha) = \left[\frac{2n^2+4n+3}{2n^2+1+(2\alpha+1)(2n+1)}\right]^{1/2} \cos\frac{k\pi}{n+1}$$

for the values of k and α such that the right-hand side of (1.11) is $\leqq 1$. Ahmed, Muldoon and Spigler [2] have observed that (3.14) is sometimes sharper than (1.7).

In Table 2 the exact values of the zeros $\theta_{16,k}(0)$, $k = 1, 2, \cdots, 8$, of the Legendre polynomial $P_{16}(\cos\theta)$ are compared with the upper bounds given by (3.13), the lower bounds given by the nonelementary inequality (2.16) and the lower bounds obtained from the larger of (3.13) or (3.14). The asterisk indicates the case where the lower bound in (3.14) is better than the one in (3.13).

TABLE 2

Bounds for the zeros $\theta_{16,k}(0)$.

k	Lower bound (3.13) or (3.14)	Lower bound (2.16)	Exact value	Upper bound (3.13)
1	0.1427 9967	0.1457 2467	0.1457 2468	0.1459 9303
2	0.3331 9922	0.3344 9861	0.3344 9864	0.3345 2581
3	0.5235 9878	0.5243 8659	0.5243 8664	0.5243 9402
4	0.7139 9833	0.7145 2518	0.7145 2525	0.7145 2820
5	0.9043 9789	0.9047 5743	0.9047 5753	0.9047 5895
6	1.0947 9744	1.0950 3327	1.0950 3340	1.0950 3414
7	1.2851 9699	1.2853 3126	1.2853 3144	1.2853 3181
8	1.4756 3699*	1.4756 4003	1.4756 4028	1.4756 4039

Acknowledgments. It is a pleasure to thank Professor Roderick Wong for several helpful discussions and for the hospitality of the Department of Mathematics and Astronomy of the University of Manitoba where part of this work was carried out.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, Applied Mathematics Series, 55, National Bureau of Standards, Washington, DC, 1964.

[2] S. AHMED, M. E. MULDOON AND R. SPIGLER, *Inequalities and numerical bounds for zeros of ultraspherical polynomials*, this Journal, 17 (1986), pp. 1000-1007.

[3] C. L. FRENZEN AND R. WONG, *A uniform asymptotic expansion of the Jacobi polynomials with error bounds*, Canad. J. Math., 37 (1985), pp. 979-1007.

[4] ———, *The Lebesgue constants for Jacobi series*, Rend. Semin. Mat. Univ. Politec. Torino, Fasc. spec., Special Functions: Theory and Computation, (1985), pp. 117-148.

[5] L. GATTESCHI, *Una nuova disuguaglianza per gli zeri dei polinomi di Jacobi*, Atti Accad. Sci Torino Cl. Sci. Fis. Mat. Natur., 103 (1968-69), pp. 259-265.

[6] L. GATTESCHI, *On the zeros of Jacobi polynomials and Bessel functions*, Rend. Semin. Mat. Univ. Politec. Torino, Fasc. spec., Special Functions: Theory and Computation, (1985), pp. 149–177.

[7] L. GATTESCHI AND G. PITTALUGA, *An asymptotic expansion for the zeros of Jacobi polynomials*, in Mathematical Analysis, J. M. Rassias, ed., Teubner-Texte zür Math., Bd. 79, 1985, pp. 70-86.

[8] M. E. MULDOON, private communication.

[9] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications, Vol. 23, 4th ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

005

NEW PROPERTIES OF THE ZEROS OF A JACOBI POLYNOMIAL IN RELATION TO THEIR CENTROID*

ÁRPÁD ELBERT† AND ANDREA LAFORGIA‡

Abstract. In this paper we investigate the sign of $(d/dx)^i P_n^{(\alpha,\beta)}(x)$, $(i=0,1,\cdots,n)$ at the point $x=s_n=(\beta-\alpha)/(2n+\alpha+\beta)$, where $P_n^{(\alpha,\beta)}(x)$ is the Jacobi polynomial of degree n. As an application we establish new inequalities for the zeros $x_{nk}^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}(x)$. The asymptotic property $x_{nk}^{(\alpha,\beta)}\to s_n$, as the parameters α, β tend to ∞ in a particular way, is shown.

Key words. Jacobi polynomials, zeros of Jacobi polynomials, asymptotics

AMS(MOS) subject classification. Primary 33A65

1. Introduction and preliminaries. The ultraspherical polynomials $P_n^{(\lambda)}(x)$ satisfy the symmetry relation [3, p. 80]

$$(1.1)\qquad P_n^{(\lambda)}(-x)=(-1)^n P_n^{(\lambda)}(x),\qquad n=1,2,\cdots,\quad \lambda>-\tfrac{1}{2}$$

which implies

$$(1.2)\qquad P_{2k-1}^{(\lambda)}(0)=0,\quad \frac{d}{dx}P_{2k}^{(\lambda)}(x)\bigg|_{x=0}=0,\quad k=1,2,\cdots,\quad \lambda>-\tfrac{1}{2}.$$

Thus the value $x=0$ plays a particular role in studying the properties of the ultraspherical polynomials. The usefulness of properties (1.2) can be seen for example in applications of Sturmian comparison theorems where the value $x=0$ is used as the initial point. Unfortunately the properties (1.1) and (1.2) cannot be extended to the more general case of Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$, $\alpha,\beta>-1$. (In the common notation we have $P_n^{(\lambda)}(x)=P_n^{(\lambda-1/2,\lambda-1/2)}(x)$.) Thus in Jacobi's case the question arises naturally to find a value s_n which would play a similar role as $x=0$ in (1.2).

For $\alpha,\beta>-1$ and $n=1,2,\cdots$ let s_n be defined by

$$(1.3)\qquad s_n=s_n^{(\alpha,\beta)}=\frac{\beta-\alpha}{2n+\alpha+\beta}.$$

We know that $P_n^{(\alpha,\beta)}(x)$ has single zeros $x_{nk}=x_{nk}^{(\alpha,\beta)}(k=1,2,\cdots,n)$ in $(-1,1)$ denoted in decreasing order, i.e., $1>x_{n1}>x_{n2}>\cdots>x_{nn}>-1$.

We observe that the zeros x_{nk} and the value s_n in (1.3) are connected by

$$(1.4)\qquad s_n=\frac{1}{n}\sum_{k=1}^{n}x_{nk}.$$

The meaning of (1.4) is that s_n is the *centroid* of the zeros of $P_n^{(\alpha,\beta)}(x)$. In fact the polynomial $P_n^{(\alpha,\beta)}(x)$ can be written as

$$\begin{aligned}P_n^{(\alpha,\beta)}(x)&=k_n^{(\alpha,\beta)}(x-x_{n1})(x-x_{n2})\cdots(x-x_{nn})\\&=k_n^{(\alpha,\beta)}[x^n-(x_{n1}+x_{n2}+\cdots+x_{nn})x^{n-1}+\cdots]\\&=k_n^{(\alpha,\beta)}(x^n+r_n^{(\alpha,\beta)}x^{n-1}+\cdots).\end{aligned}$$

* Received by the editors April 21, 1986; accepted for publication May 20, 1986. This research was supported by the Consiglio Nazionale delle Ricerche of Italy.

† Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest, Hungary.

‡ Dipartimento di Matematica, Ed Applicazioni, via Archirafi 34, 90123 Palermo, Italy.

Therefore [1, p. 169]

$$x_{n1}+x_{n2}+\cdots+x_{nn}=-r_n^{(\alpha,\beta)}=n\frac{\beta-\alpha}{2n+\alpha+\beta}$$

which proves (1.4).

Actually we have by [1, 10.8 (16), (17)]

$$2P_1^{(\alpha,\beta)}(x)=(2+\alpha+\beta)x-\beta+\alpha,$$

$$\frac{d}{dx}P_2^{(\alpha,\beta)}(x)=\frac{1}{2}(\alpha+\beta+3)P_1^{(\alpha+1,\beta+1)}(x)=\frac{1}{2}(\alpha+\beta+3)[(4+\alpha+\beta)x+\alpha-\beta],$$

and it is not difficult to see that

$$P_1^{(\alpha,\beta)}(s_1)=0,\qquad \frac{d}{dx}P_2^{(\alpha,\beta)}(s_2)=0$$

which resembles (1.2) for $n=1$ and $n=2$.

Now we can formulate the main results, which shows that the situation changes completely for $n\geqq 3$.

THEOREM 1.1. *Let $n=1,2,\cdots$, $\alpha,\beta>1$, $\alpha\neq\beta$. Let s_n be defined by* (1.4) *and put $k=[(n-1)/2]$, $\gamma_k=8k^2/(k-1)(2k-1)$. Then the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ satisfied the relations*

$$\left(\frac{d}{dx}\right)^n P_n^{(\alpha,\beta)}(x)\bigg|_{x=s_n}>0,\quad \left(\frac{d}{dx}\right)^{n-1} P_n^{(\alpha,\beta)}(x)\bigg|_{x=s_n}=0,$$

$$(-1)^i\left(\frac{d}{dx}\right)^{n-2i} P_n^{(\alpha,\beta)}(x)\bigg|_{x=s_n}>0,\quad (-1)^{i+1}(\beta-\alpha)\left(\frac{d}{dx}\right)^{n-2i-1} P_n^{(\alpha,\beta)}(x)\bigg|_{x=s_n}>0, \tag{1.5}$$

for $i=1$ and for $i=2,3,\cdots,k$, $|\alpha-\beta|<\sqrt{\gamma_k}$. Moreover if n is even then also the inequality

$$(-1)^{n/2}P_n^{(\alpha,\beta)}(s_n)>0$$

holds for $\alpha,\beta>0$ and $|\alpha-\beta|\leqq\sqrt{\gamma_{k+1}}$.

Theorem 1.1 will be proved in § 2. In § 3 some applications and consequences of our theorem will be given.

2. The proof of Theorem 1.1. For the proof of the theorem we shall need the following result.

LEMMA 2.1. *For real variables a,b let the functions δ_i, $\varepsilon_i (i=1,2,\cdots)$ be defined by the recurrence relations*

$$\varepsilon_1=0,\qquad \delta_1=ab(a+b),$$

$$\varepsilon_{i+1}=\delta_i+ab\left(\frac{a+b}{2i-1}-1\right)\varepsilon_i, \tag{2.1}$$

$$\delta_{i+1}=\left[ab\left(\frac{a+b}{2i}-1\right)-(a-b)^2\right]\delta_i-(a-b)^2ab\left(\frac{a+b}{2i-1}-1\right)\varepsilon_i.$$

Let $j=1,2,3,\cdots$; suppose $a,b\geqq 2j$, for $j=2,3,\cdots$ also $|a-b|\leqq\sqrt{\gamma_j}$, where γ_j has the same meaning as in Theorem 1.1. *Then*

$$\delta_i>\frac{2j-1}{2i-1}(j-i)(a-b)^2\varepsilon_i,\quad \varepsilon_{i+1}>0,\quad \textit{for } i=1,2,\cdots,j. \tag{2.2}$$

Proof. It is clear that the function $\varepsilon_i=\varepsilon_i(a, b)$ and $\delta_i=\delta_i(a, b)$ are polynomials in the variables a, b.

The proof of (2.2) will be carried out by complete induction.

By (2.1) we obtain

$$\delta_1=ab(a+b)>(2j-1)(j-1)(a-b)^2\varepsilon_1=0,$$

$$\varepsilon_2=\delta_1+ab(a+b-1)\varepsilon_1=\delta_1>0,$$

which proves also the validity of (2.2) at $j=1$.

From now on we may assume $j\geqq 2$. Suppose the validity of (2.2) for a fixed $i\in\{1,2,\cdots,j-1\}$. Then we have to prove that

$$\delta_{i+1}>\frac{2j-1}{2i+1}(j-i-1)(a-b)^2\varepsilon_{i+1}$$

which can be written in the form

$$\delta_{i+1}>\sigma_{i+1}\Delta^2\varepsilon_{i+1} \tag{2.3}$$

where the notation

$$\sigma_i=\frac{2j-1}{2i-1}(j-i), \qquad \Delta=|a-b|$$

have been used. By (2.1) the inequality (2.3) is equivalent to

$$\left[ab\left(\frac{a+b}{2i}-1\right)-\Delta^2-\sigma_{i+1}\Delta^2\right]\delta_i>\Delta^2ab\left(\frac{a+b}{2i-1}-1\right)(1+\sigma_{i+1})\varepsilon_i. \tag{2.4}$$

First we prove that the quantity in the brackets is positive. Since $a, b\geqq 2j$, we have

$$ab\left(\frac{a+b}{2i}-1\right)\geqq 4j^2\left(\frac{2j}{i}-1\right)$$

and $\Delta^2\leqq\gamma_j$, hence it is sufficient to show that

$$(2i+1)\left(\frac{2j}{i}-1\right)-\frac{2}{(j-1)(2j-1)}[(2j-1)(j-i-1)+2i+1]>0.$$

Since

$$\frac{(2j-i)(2i+1)}{i}>4j-2i+1$$

then we need only to prove that

$$4j-2i+1-\frac{2}{(j-1)(2j-1)}[(2j-1)(j-i-1)+2i+1]>0, \qquad i=1,2,\cdots,j-1.$$

The expression on the left-hand side is linear in i, thus we have to check this inequality only at $i=1$ and $i=j-1$. Actually at $i=0$ we have

$$4j-1-\frac{2}{(j-1)(2j-1)}>0$$

and at $i=j-1$

$$2j+3-\frac{2}{j-1}>0,$$

respectively. Therefore the coefficient of δ_i in (2.4) is positive. By induction we know that $\delta_i > \sigma_i\Delta^2\varepsilon_i \geqq 0$; thus instead of (2.4) it is sufficient to prove the inequality

$$\sigma_i\left(\frac{a+b}{2i}-1\right)-(1+\sigma_{i+1})\left(\frac{a+b}{2i-1}-1\right)\geqq\sigma_i(1+\sigma_{i+1})\frac{\Delta^2}{ab},\qquad i=1,2,\cdots,j-1,$$

or equivalently

$$\begin{aligned}(a+b-4j+4)[(2j-1)(j+i)-2i(2i+1)]+(j-i-1)[4j(2j-1)+8(j-1)i-8i^2]&\\ \geqq 2i(2j-1)(j-i)[(2j-1)(j-i-1)+2i+1]\frac{\Delta^2}{ab},\qquad i=1,2,\cdots,j-1.&\end{aligned}\tag{2.5}$$

Since the coefficient of $(a+b-4j+4)$ is positive and $a, b\geqq 2j, \Delta^2\leqq\gamma_j=8j^2/(2j-1)(j-1)$, we have only to show that

$$\begin{aligned}Q(i)=4[(2j-1)(j+i)-2i(2i+1)]+(j-i-1)[4j(2j-1)+8(j-1)i-8i^2]&\\ -4i\frac{j-i}{j-1}[(2j-1)(j-i-1)+2i+1]\geqq 0.&\end{aligned}$$

By straightforward calculations $Q(i)$ can be written as

$$Q(i)=4\frac{j-i-1}{j-1}[i(j-i-1)+j(4j^2-j-2)]$$

which is clearly positive for $i=1,2,\cdots,j-2$, while for $i=j-1$, $Q(j-1)=0$. Thus we have proved the inequality $\delta_{i+1}>\sigma_{i+1}\Delta^2\varepsilon_{i+1}$. Since by induction step we know that $\varepsilon_{i+1}>0$, hence $\delta_{i+1}>0$.

To complete the proof of Lemma 2.1 we have still to show that $\varepsilon_{i+2}>0$. But this is true by the recurrence relation (2.1). This completes the proof of Lemma 2.1.

Now we can pass over to the proof of Theorem 1.1.

From [3, p. 68] we know that

$$P_n^{(\alpha,\beta)}(x)=2^{-n}\sum_{m=0}^{n}\binom{n+\alpha}{m}\binom{n+\beta}{n-m}(x-1)^{n-m}(x+1)^m.\tag{2.6}$$

Let us consider the positive variables a, b as

$$a=n+\alpha,\qquad b=n+\beta;\tag{2.7}$$

then by using (1.3) and (2.6) we get

$$s_n=\frac{b-a}{a+b},\qquad P_n^{(\alpha,\beta)}(x)=2^{-n}\sum_{m=0}^{n}\binom{a}{m}\binom{b}{n-m}(x-1)^{n-m}(x+1)^m.\tag{2.8}$$

From [3, p. 63]

$$\begin{aligned}\frac{d}{dx}P_n^{(\alpha,\beta)}(x)&=\frac{1}{2}(n+\alpha+\beta+1)P_{n-1}^{(\alpha,\beta)}(x)\\ &=2^{-n}(n+\alpha+\beta+1)\cdot\sum_{m=0}^{n-1}\binom{a}{m}\binom{b}{n-m-1}(x-1)^{n-m-1}(x+1)^m,\end{aligned}$$

it follows

$$\left(\frac{d}{dx}\right)^l P_n^{(\alpha,\beta)}(x)=2^{-n}(n+\alpha+\beta+1)\cdots(n+\alpha+\beta+l) \tag{2.9}$$
$$\cdot\sum_{m=0}^{n-l}\binom{a}{m}\binom{b}{n-m-l}(x-1)^{n-m-l}(x+1)^m,\quad l=0,1,\cdots,n.$$

For our purposes it is convenient to use the notation

$$c_h=c_h(a,b)=2^{-h}(a+b)^h\sum_{m=0}^{h}\binom{a}{m}\binom{b}{h-m}(s-1)^{h-m}(s+1)^m,\quad h=0,1,\cdots. \tag{2.10}$$

By (2.8) we have $s-1=-2a/(a+b)$, $s+1=2b/(a+b)$; hence

$$c_h=\sum_{m=0}^{h}\binom{a}{m}\binom{b}{h-m}(-a)^{h-m}b^m,\qquad h=0,1,\cdots.$$

It is clear that the function $c_h(a,b)$ is a polynomial in the variables a, b and moreover

$$c_0=1,\quad c_1=0,\quad c_2=\frac{-1}{2}ab(a+b). \tag{2.11}$$

The functions c_h satisfy a recurrence relation which we are going to find. To this end we consider the generating function

$$\phi(x)=\sum_{h=0}^{\infty}c_hx^h. \tag{2.12}$$

We claim that

$$\phi(x)=(1-ax)^b(1+bx)^a, \tag{2.13}$$

so the power series in (2.12) is convergent for $|x|<\min\{1/a, 1/b\}$. Indeed

$$(1-ax)^b=\sum_{k=0}^{\infty}\binom{b}{k}(-ax)^k,\qquad (1+bx)^a=\sum_{m=0}^{\infty}\binom{a}{m}(bx)^m;$$

therefore

$$(1-ax)^b(1+bx)^a=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\binom{a}{m}\binom{b}{k}(-a)^kb^mx^{k+m}$$
$$=\sum_{h=0}^{\infty}x^h\sum_{m=0}^{h}\binom{a}{m}\binom{b}{h-m}(-a)^{h-m}b^m=\sum_{h=0}^{\infty}c_hx^h=\phi(x).$$

Taking the logarithmic derivative of $\phi(x)$ in (2.13) we get

$$\frac{\phi'}{\phi}=\frac{-ab(a+b)x}{(1-ax)(1+bx)},$$

or equivalently,

$$[1+(b-a)x-abx^2]\phi'=-ab(a+b)x\phi.$$

Equating the coefficients of the terms x^m in both sides we obtain

$$(m+1)c_{m+1}=-(b-a)mc_m+ab(m-1-a-b)c_{m-1},\qquad m=1,2,\cdots. \tag{2.14}$$

This follows directly from [1, 10.8 (14), (17)]. Now because of (2.11) the values of c_0 and c_1 are known, hence the sequence $\{c_n\}_{n=0}^{\infty}$ is well defined.

Replacing in (2.14) c_m by

$$\begin{aligned}(2i-1)c_{2i-1}&=(-1)^i(b-a)\varepsilon_i, \qquad i=1,2,\cdots,\\ 2ic_{2i}&=(-1)^i\delta_i,\end{aligned}\tag{2.15}$$

we obtain for ε_i and δ_i the recurrence relations (2.1) given in Lemma 2.1. By (2.9), (2.10), (2.15) the connection between the derivatives of $P_n^{(\alpha,\beta)}(x)$ at $x=s_n$ and the functions ε_i and δ_i is the following:

$$(-1)^i\left(\frac{d}{dx}\right)^{n-2i}P_n^{(\alpha,\beta)}(x)\Bigg|_{x=s_n}=\frac{2^{2i-n-1}}{i}(a+b)^{-2i}\delta_i\prod_{j=1}^{n-2i}(n+\alpha+\beta+j),$$

$$(-1)^{i+1}(\beta-\alpha)\left(\frac{d}{dx}\right)^{n-2i-1}P_n^{(\alpha,\beta)}(x)\Bigg|_{x=s_n}$$

$$=\frac{2^{2i-n+1}}{2i+1}(\beta-\alpha)^2(a+b)^{-2i-1}\varepsilon_{i+1}\prod_{j=1}^{n-2i-1}(n+\alpha+\beta+j),$$

where, in view of (2.7), the relation $b-a=\beta-\alpha$ was used. In order to prove Theorem 1.1 we need to show that $\varepsilon_2, \varepsilon_3, \cdots, \varepsilon_{k+1}, \delta_1, \delta_2, \cdots, \delta_k$ are positive.

When n is odd, i.e., $n=2k+1$, applying Lemma 2.1 with $j=k$ we have that $\varepsilon_2, \cdots, \varepsilon_{k+1}$ and $\delta_1, \cdots, \delta_k$ are positive. Finally when n is even, i.e., $n=2k+2$ we have as before the desired sign of the derivatives in (1.5) as in the previous case.

It remains only to investigate the inequality $(-1)^{n/2}P_n^{(\alpha,\beta)}(s_n)>0$, which is equivalent to the relation $\delta_{k+1}>0$. But for $\alpha, \beta \geqq 0$ we can apply the lemma with $j=k+1$ leading to $\delta_{k+1}>0$.

The proof of Theorem 1.1 is complete.

3. Further results. We are concerned in this section with some consequences of Theorem 1.1 and related problems. The first result is the following.

THEOREM 3.1. *Let $x_{ni}^{(\alpha,\beta)}$ denote the ith zero of $P_n^{(\alpha,\beta)}(x)$ in decreasing order and suppose $|\alpha-\beta|<\sqrt{\gamma_k}$, where γ_k has been defined in Theorem 1.1. Then*

(i) *for* $\beta>\alpha>-1$ $\quad x_{2k+1,k+2}^{(\alpha,\beta)}<s_{2k+1}^{(\alpha,\beta)}<x_{2k+1,k+1}^{(\alpha,\beta)}$,

(ii) *for* $\alpha>\beta>-1$ $\quad x_{2k+1,k+1}^{(\alpha,\beta)}<s_{2k+1}^{(\alpha,\beta)}<x_{2k+1,k}^{(\alpha,\beta)}$,

(iii) *for* $\alpha, \beta>0$ $\quad x_{2k,k+1}^{(\alpha,\beta)}<s_{2k}^{(\alpha,\beta)}<x_{2k,k}^{(\alpha,\beta)}$.

Proof. In the interval $(x_{n,i+1}^{(\alpha,\beta)}, x_{ni}^{(\alpha,\beta)})\equiv(x_{i+1}, x_i)$ the polynomial $P_n^{(\alpha,\beta)}(x)$ has the same sign as $(-1)^i$. Moreover applying Theorem 1.1 with $\beta>\alpha$ we get that $(-1)^{k+1}P_{2k+1}^{(\alpha,\beta)}(s_{2k+1})$ is positive. Hence s_{2k+1} belongs to one of the intervals

$$\cdots, (x_{k+4}, x_{k+3}), (x_{k+2}, x_{k+1}), (x_k, x_{k-1}), \cdots.$$

Letting $\beta\to\alpha$ we get $x_{k+1}\to x_{2k+1,k+1}^{(\alpha,\alpha)}=0=s_{2k+1}^{(\alpha,\alpha)}$, thus by continuity argumentation we obtain the only possibility $s_{2k+1}^{(\alpha,\beta)}\in(x_{k+2}^{(\alpha,\beta)}, x_{k+1}^{(\alpha,\beta)})$ which proves the first part of Theorem 3.1. The second part can be proved similarly.

For $\alpha>\beta>-1$ we have $(-1)^kP_{2k+1}^{(\alpha,\beta)}(s_{2k+1}^{(\alpha,\beta)})>0$, thus s_{2k+1} belongs to one of the intervals $\cdots(x_{k+3}, x_{k+2}), (x_{k+1}, x_k), (x_{k-1}, x_{k-2}), \cdots$ and again by continuity argumentation we have the conclusion of part (ii) of Theorem 3.1.

Finally when n is even we have from Theorem 1.1 $(-1)^k P_{2k}^{(\alpha,\beta)}(s_{2k})>0$ and s_{2k} belongs to one of the intervals

$$\cdots, (x_{2k,k+3}, x_{2k,k+2}), (x_{2k,k+1}, x_{2k,k}), (x_{2k,k-1}, x_{2k,k-2}), \cdots.$$

Now, letting $\beta \to \alpha$ we get $s_{2k}^{(\alpha,\beta)} \to 0$.

Moreover by symmetry properties of the ultraspherical polynomials we know that $x_{2k,k}^{(\alpha,\alpha)}>0$ and $x_{2k,k+1}^{(\alpha,\alpha)}<0$. Therefore as before by continuity argumentation we obtain $s_{2k}^{(\alpha,\beta)} \in (x_{2k,k+1}^{(\alpha,\beta)}, x_{2k,k}^{(\alpha,\beta)})$. The proof of Theorem 3.1 is complete.

As an immediate consequence of Theorem 3.1 we have the following result.

COROLLARY 3.1. *For* $|\alpha-\beta|<\sqrt{\gamma_k}$ *the following inequalities*

$$\begin{aligned}
&x_{2k+1,k+1}^{(\alpha,\beta)} > \frac{\beta-\alpha}{4k+\alpha+\beta+2}, \qquad \beta>\alpha>-1,\\
&x_{2k+1,k+1}^{(\alpha,\beta)} < \frac{\beta-\alpha}{4k+\alpha+\beta+2}, \qquad \alpha>\beta>-1,\\
&x_{2k,k}^{(\alpha,\beta)} > \frac{\beta-\alpha}{4k+\alpha+\beta}, \qquad \beta>\alpha>0,\\
&x_{2k,k+1}^{(\alpha,\beta)} < \frac{\beta-\alpha}{4k+\alpha+\beta}, \qquad \alpha>\beta>0
\end{aligned} \tag{3.1}$$

hold.

Remark 3.1. The bounds given in Corollary 3.1 are more stringent than the ones which we could obtain by applying the results of Stieltjes [3, p. 121]

$$\frac{\partial}{\partial\beta} x_{ni}^{(\alpha,\beta)}>0, \qquad \frac{\partial}{\partial\alpha} x_{ni}^{(\alpha,\beta)}<0.$$

Remark 3.2. We do not believe that the inequalities (3.1) are the best ones. Indeed in the particular case $\alpha=-\frac{1}{2}, \beta=\frac{1}{2}$ the Jacobi polynomial $P_n^{(-1/2,1/2)}(x)$ has the form [3, p. 60]

$$P_n^{(-1/2,1/2)}(\cos\theta)=\frac{1\cdot 3\cdot \cdots \cdot(2n-1)}{2\cdot 4\cdot \cdots \cdot 2n}\,\frac{\cos((2n+1)/2)\,\theta}{\cos\theta/2}.$$

Thus for $n=2k+1$ we get

$$x_{2k+1,k+1}^{(-1/2,1/2)}=\cos\frac{2k+1}{4k+3}\pi=\sin\frac{\pi}{2(4k+3)},$$

while $s_{2k+1}=1/(4k+2)$. So we lead to the following question: Let $-1<\alpha<\beta$. For which values of $\mu_1, \mu_2>0$ do the inequalities

$$\mu_1\frac{\beta-\alpha}{4k+\alpha+\beta+2}<x_{2k+1,k+1}^{(\alpha,\beta)}<\mu_2\frac{\beta-\alpha}{4k+\alpha+\beta+2}$$

hold? By Theorem 3.1 this is true with $\mu_1=1$. We suspect that $\mu_2=\pi/2$. It is clear that a similar question can be considered for even n.

Concerning the zeros of Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ under the condition $-\frac{1}{2}\leqq\alpha\leqq\frac{1}{2}$, $-\frac{1}{2}\leqq\beta\leqq\frac{1}{2}$ we have the inequalities [3, p. 125]

$$\frac{2i+\alpha+\beta-1}{2n+\alpha+\beta+1}\pi<\theta_{ni}^{(\alpha,\beta)}<\frac{2i}{2n+\alpha+\beta+1}\pi, \qquad i=1,2,\cdots,n, \tag{3.2}$$

where, as usual

$$x_{ni}^{(\alpha,\beta)}=\cos\,\theta_{ni}^{(\alpha,\beta)}=\cos\,\theta_{ni}.$$

The results of Theorem 3.1 enable us to improve several bounds in (3.2). The results are given in what follows.

THEOREM 3.2. *For* $-\frac{1}{2}\leqq\alpha<\beta\leqq\frac{1}{2}$ *the function* $P_{2k+1}^{(\alpha,\beta)}(\cos\,\theta)$ *has exactly k zeros such that*

$$\frac{\pi}{2}<\theta_{2k+1,k+2}<\theta_{2k+1,k+3}<\cdots\theta_{2k+1,2k+1}<\pi.$$

Moreover the inequalities

$$\theta_{2k+1,i}<\frac{\pi}{2}-\tau+\frac{2(i-k-1)}{4k+\alpha+\beta+3}\pi,\qquad i=k+1,\cdots,2k+1,$$

hold, where

$$\tau=\arcsin\,((\beta-\alpha)/(4+\alpha+\beta+2)).$$

Proof. By the lower bound in (3.2) we obtain for $n=2k+1$, $i=k+2$

$$\theta_{2k+1,k+2}>\frac{2k+\alpha+\beta+3}{4k+\alpha+\beta+3}\pi$$

and the right-hand side is larger than $\pi/2$, because $\alpha+\beta+3>0$. Therefore the values $\theta_{2k+1,i}(i=k+2,k+3,\cdots,2k+1)$ lie in $(\pi/2,\pi)$.

On the other hand by [3, p. 125, (6.3.3)] we have also the inequality

$$\theta_{2k+1,i+1}-\theta_{2k+1,i}\leqq\frac{2\pi}{4k+\alpha+\beta+3},\qquad i=1,2,\cdots,2k,\tag{3.3}$$

where equality occurs only for $|\alpha|=\frac{1}{2}$, $|\beta|=\frac{1}{2}$.

For $\beta-\alpha\leqq1<\sqrt{\gamma_k}$ we can apply Theorem 3.1 leading to the inequality

$$x_{2k+1,k+1}=\cos\,\theta_{2k+1,k+1}>\frac{\beta-\alpha}{4k+\alpha+\beta+2}=\cos\left(\frac{\pi}{2}-\tau\right),$$

hence

$$\theta_{2k+1,k+1}^{(\alpha,\beta)}<\frac{\pi}{2}-\tau.$$

We conclude that there are exactly k zeros in $(\pi/2,\pi)$.

By repeated applications of inequality (3.3) for $i=k+1,k+2,\cdots,2k$ we obtain the desired inequality for $\theta_{2k+1,i+1}$ which completes the proof of Theorem 3.2.

In the case $|\alpha|\geqq\frac{1}{2}$, $|\beta|\geqq\frac{1}{2}$ similar results are established in the following theorem.

THEOREM 3.3. *For* $|\alpha|\geqq\frac{1}{2}$, $|\beta|\geqq\frac{1}{2}$ *and* $-1<\alpha<\beta<\alpha+\sqrt{\gamma_k}$ *the function* $P_{2k+1}^{(\alpha,\beta)}(\cos\,\theta)$ *has at least* $k+1$ *zeros in* $(0,\pi/2)$ *such that the inequalities*

$$\theta_{2k+1,i}<\frac{\pi}{2}-\tau-\frac{2(k+1-i)}{4k+\alpha+\beta+3}\pi,\qquad i=1,2,\cdots,k+1$$

hold, where τ *is the same as in Theorem* 3.2.

Proof. The proof runs on the same lines as the proof of Theorem 3.2. We observe that following the same arguments given in [3, p. 125] for the proof of inequality (3.3) we obtain

$$\theta_{2k+1,i+1}-\theta_{2k+1,i}\geqq\frac{2\pi}{4k+\alpha+\beta+3},\qquad i=1,2,\cdots,2k,|\alpha|\geqq\tfrac{1}{2},\qquad|\beta|\geqq\tfrac{1}{2}$$

(where equality occurs if and only if $|\alpha|=|\beta|=\frac{1}{2}$) and the conclusion of Theorem 3.3 follows from these inequalities using Theorem 3.1.

Remark 3.3. For $|\alpha|, |\beta| \geqq \frac{1}{2}$, $\alpha>\beta$ a similar argumentation gives that the function $P^{(\alpha,\beta)}_{2k+1}(\cos\theta)$ has exactly k zeros $\theta_{2k+1,i}(i=k+2, \cdots 2k+1)$ in $(\pi/2, \pi)$ which satisfy

$$\theta_{2k+1,i}>\frac{\pi}{2}+\tau+\frac{2(i-k-1)}{4k+\alpha+\beta+3}\pi, \quad i=k+2, \cdots 2k+1, \quad \alpha-\beta<\sqrt{\gamma_k}.$$

In the case $n=2k$ the following result can be established.

THEOREM 3.4. *For $0<\alpha<\beta\leqq\frac{1}{2}$ the last $k+1$ zeros of the function $P^{(\alpha,\beta)}_{2k}(\cos\theta)$ satisfy the inequalities*

$$\theta_{2k,k}<\frac{\pi}{2}<\theta_{2k,k+1}<\cdots<\theta_{2k,2k}<\pi,$$

$$\theta_{2k,i}<\frac{\pi}{2}-\nu+\frac{2(i-k)}{4k+\alpha+\beta+1}\pi, \qquad i=k+1, \cdots, 2k$$

where

$$\nu=\arcsin((\beta-\alpha)(4k+\alpha+\beta)).$$

Proof. From (3.2) we obtain

$$\theta_{2k,k+1}>\frac{2k+\alpha+\beta+1}{4k+\alpha+\beta+1}\pi>\frac{\pi}{2}.$$

Therefore the values $\theta_{2k,i}(i=k+1, k+2, \cdots 2k)$ lie in $(\pi/2, \pi)$.

On the other hand

$$\theta_{2k,k}<\frac{2k}{4k+\alpha+\beta+1}\pi<\frac{\pi}{2}$$

and we conclude that in $(\pi/2, \pi)$ there are *exactly* k zeros of $P^{(\alpha,\beta)}_{2k}(\cos\theta)$. Now by Corollary 3.1 with $\beta>\alpha$ we obtain

$$x_{2k,k}=\cos\theta_{2k,k}>\frac{\beta-\alpha}{4k+\alpha+\beta}=\cos\left(\frac{\pi}{2}-\nu\right), \qquad \beta>\alpha>0,$$

hence

$$\theta_{2k,k}<\frac{\pi}{2}-\nu.$$

Using this inequality in [3, (6.3.3)], by repeated applications, we get

$$\theta_{2k,i}<\frac{\pi}{2}-\nu+\frac{2(i-k)}{4k+\alpha+\beta+1}\pi, \qquad i=k+1, \cdots, 2k$$

which is the conclusion of the theorem.

Now we turn our attention to another interesting property of Jacobi polynomial $P^{(\alpha,\beta)}_n(x)$ which is also related to the *centroid* s_n of its zeros.

THEOREM 3.5. *Suppose $\alpha, \beta>-1$ and $\mu\geqq1$. Denote by $x_1(\mu), x_2(\mu)\cdots x_n(\mu)$ the zeros of the Jacobi polynomial $P^{((\mu-1)n+\mu\alpha, (\mu-1)n+\mu\beta)}_n(x)$ in decreasing order. Then the asymptotic relations*

$$\lim_{\mu\to\infty} x_i(\mu)=s_n=-\frac{\beta-\alpha}{2n+\alpha+\beta}, \qquad i=1,2, \cdots, n \tag{3.4}$$

hold.

Proof. Using (2.7) and (2.8) we have to consider the polynomial

$$Q_n(x;\mu)=2^nP_n^{((\mu-1)n+\mu\alpha,\,(\mu-1)n+\mu\beta)}(x)$$
$$=\sum_{m=0}^{n}\binom{\mu a}{m}\binom{\mu b}{n-m}(x-1)^{n-m}(x+1)^m.$$

It is clear that $Q_n(x;\mu)$ is also a polynomial with respect to μ which can be written in the form

$$Q_n(x;\mu)=\mu^nR_{nn}(a,b;x)+\mu^{n-1}R_{n,n-1}(a,b;x)+\cdots+R_{n0}(a,b;x),$$

where

$$R_{nn}(a,b;x)=\sum_{m=0}^{n}\frac{a^mb^{n-m}}{m!(n-m)!}(x-1)^{n-m}(x+1)^m$$
$$=\frac{1}{n!}[a(x+1)+b(x-1)]^n=\frac{(a+b)^n}{n!}(x-s)^n.$$

Substituting $x=x_i(\mu)$ in $Q_n(x;\mu)$ and letting $\mu\to\infty$, we find the desired result.

Remark 3.4. We observe that for $\alpha=\beta$, Theorem 3.4 gives the known result [1, p. 203, (6)] that the zeros of ultraspherical polynomial $P_n^{(\lambda)}(x)$ tend to zero if $\lambda\to\infty$. This result together with the classical Stieltjes' theorem [3, p. 121]

$$\frac{\partial}{\partial\lambda}x_{ni}^{(\lambda)}<0,\qquad i=1,2,\cdots,\left[\frac{n}{2}\right]$$

gives that $x_{ni}^{(\lambda)}$ tends to zero monotonically.

This property suggests the following question: Do the zeros $x_i(\mu)(i=1,2,\cdots n)$ of the polynomial $P_n^{((\mu-1)n+\alpha n,\,(\mu-1)n+\beta n)}(x)$ tend monotonically to zero if $\mu\to\infty$?

We conjecture that this property is true.

Remark 3.5. Finally we observe that (3.4) can be thought of as complementing the result of Moak, Saff and Varga proved in [2] where for $n\to\infty$ the behavior of the smallest and largest zeros of the Jacobi polynomials was studied.

REFERENCES

[1] A. Erdélyi et al., *Higher Transcendental Functions, Vol.* 2, McGraw-Hill, New York, 1953.

[2] D. S. Moak, E. B. Saff and R. S. Varga, *On the zeros of Jacobi polynomials* $P_n^{(\alpha_n,\beta_n)}(x)$, Trans. Amer. Math. Soc., 249 (1979), pp. 159-162.

[3] G. Szegö, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloquium Publications, Vol. 23, Amer. Math. Soc., Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

006

EXPLICIT FORMULA RELATING THE JACOBI, HAHN AND BERNSTEIN POLYNOMIALS*

Z. CIESIELSKI†

Abstract. A new connection between Jacobi and Hahn polynomials is given.

Key words. orthogonal polynomials, Jacobi polynomials, Hahn polynomials, Bernstein polynomials

AMS(MOS) subject classification. 33A65

The real linear space of all algebraic polynomials of degree not exceeding n is denoted by Π_n. Clearly, $\dim \Pi_n = n+1$.

The Bernstein polynomials

$$N_{i,n}(x)=\binom{n}{i}\left(\frac{1+x}{2}\right)^i\left(\frac{1-x}{2}\right)^{n-i}, \qquad i=0,\cdots,n$$

form a basis in Π_n.

For $\alpha, \beta > -1$ the Jacobi polynomials are defined by the Rodrigues formula

$$P_n^{(\alpha,\beta)}(x)=\frac{(-1)^n}{2^n n!(1-x)^\alpha(1+x)^\beta}D^n[(1-x^2)^n(1-x)^\alpha(1+x)^\beta], \qquad n=0,1,\cdots$$

where $D=d/dx$. The set $(P_n^{(\alpha,\beta)})_0^\infty$ is orthogonal with respect to the scalar product

$$\int_{-1}^1 f(x)g(x)(1-x)^\alpha 1+x)^\beta\,dx.$$

Since

$$\Pi_n=\operatorname{span}[P_0^{(\alpha,\beta)},\cdots,P_n^{(\alpha,\beta)}]$$

it follows that for each $j=0,\cdots,n$ there is a unique vector $(h_{j,n}(0),\cdots,h_{j,n}(n))$ such that

$$P_j^{(\alpha,\beta)}=\sum_{i=0}^n h_{j,n}(i)N_{i,n}.$$

We denote by $h_{j,n}$ the polynomial $h_{j,n}(\cdot)\in\Pi_n$ interpolating the points $(0,h_{j,n}(0)),\cdots,(n,h_{j,n}(n))$.

Our aim is to show that $(h_{j,n})_{j=0}^n$ are proportional to the Hahn polynomials.

For given $\alpha,\beta>-1$ and integer $n\geqq 0$ the jth Hahn polynomial $h_{j,n}^{(\alpha,\beta)}$ is defined by the Rodriques formula [3]

$$h_{j,n}^{(\alpha,\beta)}(x)=\frac{(-1)^j}{j!}\frac{1}{\rho(x)}\nabla^j_{\rho_j}(x),$$

where $\nabla f(x)=f(x)-f(x-1)$ and

$$\rho(x)=\frac{\Gamma(\alpha+1+n-x)\Gamma(\beta+1+x)}{\Gamma(1+n-x)\Gamma(1+x)},$$

$$\rho_j(x)=\frac{\Gamma(\alpha+1+n-x)\Gamma(\beta+1+x+j)}{\Gamma(1+n-(x+j))\Gamma(1+x)}.$$

* Received by the editors April 15, 1986; accepted for publication July 2, 1986.

† Instytut Matematyczny PAN, 81-825, Sopot, Poland.

For $j=0,\cdots,n$, we have $h_{j,n}^{(\alpha,\beta)}\in\Pi_j$ and these polynomials are orthogonal with respect to the scalar product

$$\sum_{i=0}^{n} f(i)g(i)\rho(i)$$

with the weight ρ.

THEOREM. *Let $\alpha,\beta>-1$ and the integer $n\geqq 0$ be given. Then*

$$(n)_jP_j^{(\alpha,\beta)}=\sum_{i=0}^{n} h_{j,n}^{(\alpha,\beta)}(i)N_{i,n},\qquad j=0,\cdots,n$$

where $(n)_j=n(n-1)\cdots(n-j+1)$.

For the proof we introduce the following singular operators:

$$(Lf)(x)=(1-x^2)(D^2f)(x)+[\beta-\alpha-(\alpha+\beta+2)x](Df)(x),$$
$$(l_nf)(i)=i(n+\alpha+1-i)(\nabla\Delta f)(i)+[(\beta+1)n-(\alpha+\beta+2)i](\Delta f)(i),$$

where $(\Delta f)(i)=f(i+1)-f(i)$.

LEMMA. *Let $H\in\Pi_n$ and let*

$$H=\sum_{i=0}^{n} h(i)N_{i,n}.$$

Then

$$LH=\sum_{i=0}^{n}(l_nh)(i)N_{i,n}.$$

Proof. One checks directly the identity $(i=0,\cdots,n)$

$$DN_{i,n}=\frac{n}{2}[N_{i-1,n-1}-N_{i,n-1}],$$

where $N_{-1,n-1}=N_{n,n-1}=0$. It gives the formulas

$$DH=\frac{n}{2}\sum_{i=0}^{n-1}\Delta h(i)N_{i,n-1},$$

$$D^2H=\frac{n(n-1)}{4}\sum_{i=0}^{n-2}\Delta^2h(i)N_{i,n-2}.$$

This and the definition of $N_{i,n}$ imply that

$$x(DH)(x)=\left[\frac{1+x}{2}-\frac{1-x}{2}\right](DH)(x)$$
$$=\frac{1}{2}\sum_{i=0}^{n}[i(\nabla h)(i)-(n-i)(\Delta h)(i)]N_{i,n}(x)$$

and

$$(1-x^2)(D^2H)(x)=\sum_{i=0}^{n} i(n-i)(\nabla\Delta)h(i)N_{i,n}(x).$$

Moreover, since

$$N_{i,n-1}=\frac{i+1}{n}N_{i+1,n}+\frac{n-i}{n}N_{i,n}$$

we have

$$DH=\frac{1}{2}\sum_{i=0}^{n}[i(\nabla h)(i)+(n-i)(\Delta h)(i)]N_{i,n}.$$

Now, $\nabla\Delta=\Delta-\nabla$ and therefore

$$\begin{aligned}i(n-i)\nabla\Delta+\frac{\beta-\alpha}{2}i\nabla+\frac{\beta-\alpha}{2}(n-i)\Delta-\frac{\alpha+\beta+2}{2}i\nabla+\frac{\alpha+\beta+2}{2}(n-i)\Delta\\ =i(n-i)\nabla\Delta+(\alpha+1)i(\nabla\Delta-\Delta)+(\beta+1)(n-i)\Delta\\ =i(n-i+\alpha+1)\nabla\Delta+[(\beta+1)n-(\alpha+\beta+2)i]\Delta,\end{aligned}$$

whence we infer the thesis of the lemma.

Proof of the theorem. It is known [3] that $j(j+\alpha+\beta+1), j=0,\cdots,n$, are simple eigenvalues for both operators $L:\Pi_n\to\Pi_n$ and $l_n:\Pi_n\to\Pi_n$.

Their corresponding eigenvectors are $P_j^{(\alpha,\beta)}, j=0,\cdots,n$, and $h_{j,n}^{(\alpha,\beta)}, j=0,\cdots,n$, respectively. Applying the lemma to $H=P_j^{(\alpha,\beta)}$we find

$$0=LH-j(j+\alpha+\beta+1)H=\sum_{i=0}^{n}(l_nh-j(j+\alpha+\beta+1)h)(i)N_{i,n},$$

whence

$$l_nh=j(j+\alpha+\beta+1)h,$$

and therefore $h=Ch_j^{(\alpha,\beta)}$ i.e.

$$P_j^{(\alpha,\beta)}=C\sum_{i=0}^{n}h_{j,n}^{(\alpha,\beta)}(i)N_{i,n}.$$

Using the Rodriques formulas defining $P_j^{(\alpha,\beta)}$ and $h_{j,n}^{(\alpha,\beta)}$ we find that $C\cdot(n)_j=1$, and this completes the proof.

The theorem in case of $\alpha=\beta=0$ was established earlier in [1] by a different method not extendable to the general case of $\alpha,\beta>-1$.

Consequences and applications of the theorem will be discussed elsewhere.

Remark. After this note was submitted to this Journal, Richard Askey and George Gasper kindly communicated to the author that the theorem can also be obtained with the help of the hypergeometric representations of the Jacobi and Hahn polynomials.

REFERENCES

[1] Z. Ciesielski and J. Domsta, *The degenerate B-splines as basis in the space of algebraic polynomials*, Ann. Polon. Math., 46 (1985), pp. 71-79.

[2] W. Hahn, *Über Orthogonal Polynome, die q-Differenzengleichungen genügen*, Math. Nachr., 2 (1949), pp. 4-34.

[3] A. F. Nikiforov, S. K. Suslov and V. B. Uvarov, *Classical Orthogonal Polynomials of Discrete Variable*, Nauka, ed., Moscow, 1985. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

007

MULTILATERAL SUMMATION THEOREMS FOR ORDINARY AND BASIC HYPERGEOMETRIC SERIES IN $U(n)$*

R. A. GUSTAFSON†

Abstract. In this paper we prove generalizations of ${}_2H_2$, ${}_5H_5$, ${}_1\Psi_1$, and ${}_6\Psi_6$ summation theorems for hypergeometric series in $U(n)$. This includes a further generalization of Milne's ${}_1\Psi_1$ summation theorem for basic hypergeometric series in $U(n)$. These results are mostly obtained by use of contour integration together with Milne's $U(n)$ generalizations of the Gauss, ${}_5F_4$ and ${}_6\varphi_5$ summation theorems.

Key words. hypergeometric series in $U(n)$, summation theorems, contour integrals, basic hypergeometric series, bilateral series

AMS(MOS) subject classifications. 33A75, 33A30, 33A35

1. Introduction and statement of results. In 1976 Holman, Biedenharn and Louck [12] and later Holman [11] defined interesting multivariable generalizations of the classical hypergeometric series and well-poised hypergeometric series, which they called hypergeometric series in $U(n)$. These definitions arose out of a problem in mathematical physics of finding analogues of the Wigner $(3-j)$ and Racah $(6-j)$ coefficients of angular momentum theory for the higher dimensional unitary groups. It was realized that many of the fundamental identities for the $3-j$ and $6-j$ coefficients were consequences of classical transformation and summation theorems for hypergeometric series. Holman [11] showed that, conversely, the corresponding identities for higher dimensional multiplicity-free Wigner and Racah coefficients implied generalizations of classical summation theorems to the setting of hypergeometric series in $U(n)$.

There were several important summation and transformation theorems for classical hypergeometric series for which Holman did not provide generalizations for hypergeometric series in $U(n)$. For example, he did not prove generalizations of the nonterminating Gauss summation theorem, nor a $U(n)$ generalization of Whipple's well-poised ${}_7F_6$ transformation. Furthermore, there remained the question of defining a q-analogue of the hypergeometric series in $U(n)$ whose properties would generalize those of the classical basic hypergeometric series. The importance of doing so was discussed by Andrews in [2].

Recently, Milne has proved a nonterminating $U(n)$ Gauss summation theorem [20] (see Theorem 2.7 below) and the present author gave a generalization of Whipple's transformation for hypergeometric series in $U(n)$ [10]. Milne has also defined a basic analogue of the hypergeometric series in $U(n)$ and proved a number of important properties for them [15]-[20]. Using a $U(n)$ generalization of the q-binomial theorem [15] that he found and an analytic continuation argument similar to that in Ismail [13], Milne [18] proved a $U(n)$ generalization of Ramanujan's ${}_1\Psi_1$, summation theorem for bilateral basic hypergeometric series. A specialization of Milne's ${}_1\Psi_1$ theorem gives the Macdonald identities for affine root systems of type $A_l^{(1)}$ [14], [15] in the same way as a specialization of the classical ${}_1\Psi_1$ theorem gives the Jacobi triple product identity. Milne also proved a $U(n)$ generalization of a terminating version of the well-poised ${}_6\varphi_5$ summation theorem [17] and used analytic continuation to obtain a nonterminating $U(n)$ ${}_6\varphi_5$ summation theorem [19]. This theorem is also important in that it clarifies the concept of "well-poised" hypergeometric series in the setting of hypergeometric series in $U(n)$.

* Received by the editors July 23, 1986; accepted for publication February 9, 1987.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843.

In this paper we prove $U(n)$ generalizations of some of the classical summation theorems for bilateral hypergeometric series and also for basic bilateral hypergeometric series. The method of proof is to begin with a summation theorem for one-sided (basic) hypergeometric series in $U(n)$ and use contour integration to inductively prove their multilateral generalizations. We also obtain a further generalization of Milne's ${}_1\Psi_1$ summation theorem as a consequence of a $U(n)$ generalization of the well-poised ${}_6\Psi_6$ summation theorem.

In order to define hypergeometric series we must first define the rising factorial (or Pochhammer symbol):

$$(c)_n=\frac{\Gamma(c+n)}{\Gamma(c)} \tag{1.1}$$

for $n\in\mathbb{Z}$, $c\in\mathbb{C}$.

If $n>0$, then

$$(c)_n=(c)(c+1)\cdots(c+n-1), \tag{1.2a}$$

$$(c)_{-n}=\frac{(-1)^n}{(1-c)(2-c)\cdots(n-c)} \tag{1.2b}$$

and

$$(c)_0=1. \tag{1.2c}$$

For the q-analogue, let q be a complex number with $|q|<1$ and for $a\in\mathbb{C}$, $n>0$, define

$$[a]_n=(a;\,q)_n=(1-a)(1-aq)\cdots(1-aq^{n-1}) \tag{1.3a}$$

and

$$[a]_\infty=\prod_{k=0}^{\infty}(1-aq^k). \tag{1.3b}$$

For $n\in\mathbb{Z}$, define

$$[a]_n=\frac{[a]_\infty}{[aq^n]_\infty}. \tag{1.3c}$$

The definition (1.3c) agrees with (1.3a) for $n>0$. It follows from (1.3c) that, for $n>0$, we have

$$[a]_{-n}=(a;\,q)_{-n}=\frac{1}{(1-a/q)(1-a/q^2)\cdots(1-a/q^n)}$$

$$=\frac{(-1)^n q^{n(n+1)/2}}{[q/a]_n a^n} \tag{1.4a}$$

and

$$[a]_0=1. \tag{1.4b}$$

An ordinary bilateral hypergeometric series with A numerator parameters, B denominator parameters, and variable z is defined by

$${}_AH_B\begin{bmatrix} a_1, a_2, \cdots, a_A; & \\ b_1, b_2, \cdots, b_B; & z\end{bmatrix}=\sum_{n=-\infty}^{\infty}\frac{(a_1)_n(a_2)_n\cdots(a_A)_n}{(b_1)_n(b_2)_n\cdots(b_B)_n}z^n \tag{1.5}$$

where $a_1,\cdots,a_A,b_1,\cdots,b_B,z\in\mathbb{C}$. Similarly, the basic bilateral hypergeometric series is defined by

$${}_A\Psi_B\begin{bmatrix} a_1, a_2, \cdots, a_A; & \\ b_1, b_2, \cdots, b_B; & q, z\end{bmatrix}=\sum_{n=-\infty}^{\infty}\frac{[a_1]_n[a_2]_n\cdots[a_A]_n}{[b_1]_n[b_2]_n\cdots[b_B]_n}z^n \tag{1.6}$$

where $|q|<1$ and $a_1,\cdots,a_A,\ b_1,\cdots,b_B,\ z\in\mathbb{C}$.

If one of the denominator parameters, e.g., b_1, equals one, then the terms in the bilateral series (1.5) vanish for $n<0$ and the series (1.5) reduces to the ordinary hypergeometric series

$$ {}_AF_{B-1}\left[\begin{matrix} a_1,\cdots,a_A; \\ b_2,\cdots,b_B; \end{matrix}\; z\right]. $$

Similarly, if b_1 equals q in (1.6), then the basic bilateral series (1.6) reduces to the one-sided basic hypergeometric series

$$ {}_A\varphi_{B-1}\left[\begin{matrix} a_1,\cdots,a_A; \\ b_2,\cdots,b_B; \end{matrix}\; z\right]. $$

Summation theorems for bilateral hypergeometric series date back to 1907, when Dougall [9] discovered the following pair of identities. He proved

$$ (1.7)\qquad {}_2H_2\left[\begin{matrix} a,b; \\ c,d; \end{matrix}\; 1\right]=\Gamma\left[\begin{matrix} c,d,1-a,1-b,c+d-a-b-1 \\ c-a,d-a,c-b,d-b \end{matrix}\right] $$

where $\operatorname{Re}(c+d-a-b-1)>0$ and

$$ \Gamma\left[\begin{matrix} a_1,\cdots,a_l \\ b_1,\cdots,b_k \end{matrix}\right]\equiv\frac{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_l)}{\Gamma(b_1)\Gamma(b_2)\cdots\Gamma(b_k)} $$

for $a_1,\cdots,a_l,b_1,\cdots,b_k\in\mathbb{C}$. He also proved

$$ \begin{aligned} &{}_5H_5\left[\begin{matrix} 1+\tfrac12 a,b,c,d,e; \\ \tfrac12 a,1+a-b,1+a-c,1+a-d,1+a-e; \end{matrix}\; 1\right] \\ (1.8)\qquad &=\Gamma\left[\begin{matrix} 1-b,1-c,1-d,1-e,1+a-b,1+a-c,1+a-d, \\ 1+a,1-a,1+a-b-c,1+a-b-d,1+a-b-e, \end{matrix}\right. \\ &\qquad\qquad\left.\begin{matrix} 1+a-e,1+2a-b-c-d-e \\ 1+a-c-d,1+a-c-e,1+a-d-e \end{matrix}\right], \end{aligned} $$

where, for convergence, we assume $\operatorname{Re}(3+4a-2b-2c-2d-2e)>0$. The ${}_2H_2$ identity (1.7) generalizes the classical Gauss summation theorem and the ${}_5H_5$ identity generalizes the well-poised ${}_5F_4$ summation theorem.

For basic bilateral series there are two summation theorems which are particularly important. They are Ramanujan's ${}_1\Psi_1$ summation theorem and W. N. Bailey's well-poised ${}_6\Psi_6$ summation theorem [8]. We have

$$ (1.9)\qquad {}_1\Psi_1\left[\begin{matrix} a; \\ b; \end{matrix}\; q,z\right]=\Pi\left[\begin{matrix} b/a,az,q/az,q; \\ q/a,b/az,b,z; \end{matrix}\; q\right] $$

where $|b/a|<|z|<1$ and

$$ \Pi\left[\begin{matrix} a_1,\cdots,a_l; \\ b_1,\cdots,b_k; \end{matrix}\; q\right]=\frac{[a_1]_\infty[a_2]_\infty\cdots[a_l]_\infty}{[b_1]_\infty[b_2]_\infty\cdots[b_k]_\infty} $$

for $a_1,\cdots,a_l,b_1,\cdots,b_k\in\mathbb{C}$ and $|q|<1$. Similarly

$$ \begin{aligned} &{}_6\Psi_6\left[\begin{matrix} q\sqrt{a},-q\sqrt{a},b,c,d,e; \\ \sqrt{a},-\sqrt{a},aq/b,aq/c,aq/d,aq/e; \end{matrix}\; q,a^2q/bcde\right] \\ (1.10)\qquad &=\Pi\left[\begin{matrix} aq,aq/bc,aq/bd,aq/be,aq/cd,aq/ce,aq/de,q,q/a; \\ q/b,q/c,q/d,q/e,aq/b,aq/c,aq/d,aq/e,a^2q/bcde; \end{matrix}\; q\right] \end{aligned} $$

where $|a^2q/bcde|<1$.

These basic bilateral summation theorems have many important applications. Both the q-binomial theorem and Jacobi's triple product formula can be obtained from (1.9) by appropriate specializations and limits. Also the ${}_1\Psi_1$ theorem can be used to prove some interesting integral identities (see Askey [4], [5]) which have applications to orthogonal polynomials (e.g., the Rogers q-ultra spherical polynomials [6]). Some of the applications of the ${}_6\Psi_6$ identity to number theory and the theory of partitions are discussed in Andrews' article [1]. These include a proof of Ramanujan's partition function congruence $p(5n+4)\equiv 0 \bmod 5$ and also formulas for the number of representations of an integer as a sum of 2, 4 or 8 squares. Both of the identities (1.9) and (1.10) also play an important role in the analytic theory of partitions [3].

Generalizations of the ${}_2H_2$, ${}_5H_5$, ${}_1\Psi_1$, and ${}_6\Psi_6$ identities for bilateral hypergeometric series in $U(n)$ will be stated below. It is hoped that these $U(n)$ generalizations might prove useful in higher dimensional analogues of the problems just mentioned.

In § 2 the following generalization of the ${}_2H_2$ summation theorem is proved.

THEOREM 1.11. *Let $n \geqq 1$ be an integer and* $\operatorname{Re}\left(\sum_{i=1}^{n+1}(\beta_i-\alpha_i)\right) > n$. *Then*

$$\sum_{y_1,\cdots,y_n=-\infty}^{\infty} \prod_{1\leqq i<j\leqq n}\left(\frac{(z_i+y_i)-(z_j+y_j)}{z_i-z_j}\right)\prod_{i=1}^{n+1}\prod_{k=1}^{n}\frac{(\alpha_i+z_k)_{y_k}}{(\beta_i+z_k)_{y_k}} = \frac{\Gamma\left(-n+\sum_{i=1}^{n+1}(\beta_i-\alpha_i)\right)\prod_{i=1}^{n+1}\prod_{k=1}^{n}\Gamma(1-\alpha_i-z_k)\Gamma(\beta_i+z_k)}{\prod_{i,j=1}^{n+1}\Gamma(\beta_i-\alpha_j)\prod_{1\leqq i<j\leqq n}\Gamma(1-z_i+z_j)\Gamma(1+z_i-z_j)} \tag{1.12}$$

where $z_i\neq z_j$ for $1\leqq i<j\leqq n$, and α_i+z_k and $1-\beta_i-z_k$ are not positive integers for $1\leqq i\leqq n+1$ and $1\leqq k\leqq n$.

In § 3 we prove a $U(n)$ generalization of the ${}_5H_5$ summation theorem for each $n\geqq 2$. The case $n=2$ is equivalent to the ${}_5H_5$ identity (1.8).

THEOREM 1.13. *With* $\operatorname{Re}\left(\sum_{i=1}^{n}(b_i-a_i)\right)>n-1$ *we have*

$$\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty} \prod_{1\leqq i<j\leqq n}\left(\frac{(z_i+y_i)-(z_j+y_j)}{z_i-z_j}\right)\prod_{i,k=1}^{n}\frac{(a_i-z_i+z_k)_{y_k}}{(b_i-z_i+z_k)_{y_k}} = \frac{\Gamma\left(1-n+\sum_{i=1}^{n}(b_i-a_i)\right)\prod_{i,k=1}^{n}\Gamma(1-a_i+z_i-z_k)\Gamma(b_i-z_i+z_k)}{\prod_{i,k=1}^{n}\Gamma(b_i-a_k-z_i+z_k)\prod_{1\leqq i<j\leqq n}\Gamma(1+z_i-z_j)\Gamma(1-z_i+z_j)} \cdot\frac{1}{\Gamma\left(1-\sum_{i=1}^{n}a_i\right)\Gamma\left(1-n+\sum_{i=1}^{n}b_i\right)} \tag{1.14}$$

where $z_i\neq z_j$ for $1\leqq i<j\leqq n$, and $a_i-z_i+z_j$ and $1-b_i+z_i-z_j$ are not positive integers for $1\leqq i,j\leqq n$.

In § 4 we prove a $U(n)$ generalization of the ${}_6\Psi_6$ summation theorem for each $n\geqq 2$. Again the case $n=2$ is equivalent to identity (1.10).

THEOREM 1.15. *With $|q|<1$ and $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<1$, we have*

$$\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty}\ \prod_{1\leqq i<j\leqq n}\left(\frac{z_iq^{y_i}-z_jq^{y_j}}{z_i-z_j}\right)\prod_{i,k=1}^{n}\frac{[a_iz_k/z_i]_{y_k}}{[b_iz_k/z_i]_{y_k}}$$

$$(1.16)\qquad =\frac{[q]_\infty^{n-1}\prod_{i,k=1}^{n}[(b_iz_k)/(a_kz_i)]_\infty\prod_{1\leqq i<j\leqq n}[qz_i/z_j]_\infty[qz_j/z_i]_\infty}{\left[q^{1-n}\prod_{i=1}^{n}(b_i/a_i)\right]_\infty\prod_{i,k=1}^{\infty}[(qz_i)/(a_iz_k)]_\infty[b_iz_k/z_i]_\infty}$$

$$\cdot\left[q\Big/\prod_{i=1}^{n}a_i\right]_\infty\left[q^{1-n}\prod_{i=1}^{n}b_i\right]_\infty,$$

where $z_i\neq z_j$ for $1\leqq i<j\leqq n$, and $a_iz_j/z_i\neq q^l$ and $q^{-1}b_iz_j/z_i\neq q^{-l}$ for a positive integer l and $1\leqq i,j\leqq n$.

In § 5 we prove a $U(n)$ generalization of the ${}_1\Psi_1$ summation theorem for $n\geqq 2$. The classical ${}_1\Psi_1$ identity (1.9) corresponds to the case $n=1$ and is used, along with Theorem 1.15, to prove the general result for $n\geqq 2$. The argument here is similar to that in [18].

THEOREM 1.17. *Let $|q|<1$ and $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<|t|<1$. Then*

$$\sum_{y_1,\cdots,y_n=-\infty}^{\infty}\ \prod_{1\leqq i<j\leqq n}\left(\frac{z_iq^{y_i}-z_jq^{y_j}}{z_i-z_j}\right)\prod_{i,k=1}^{n}\frac{[a_iz_k/z_i]_{y_k}}{[b_iz_k/z_i]_{y_k}}\,t^{(y_1+\cdots+y_n)}$$

$$(1.18)\qquad =\frac{\left[t\prod_{i=1}^{n}a_i\right]_\infty\left[q\left(t\prod_{i=1}^{n}a_i\right)^{-1}\right]_\infty}{[t]_\infty\left[q^{1-n}t^{-1}\prod_{i=1}^{n}(b_i/a_i)\right]_\infty}\prod_{i,k=1}^{n}\frac{[(b_iz_k)/(a_kz_i)]_\infty[qz_i/z_k]_\infty}{[(qz_i)/(a_iz_k)]_\infty[b_iz_k/z_i]_\infty}$$

where $z_i\neq z_j$ and $1\leqq i<j\leqq n$, and $a_iz_j/z_i\neq q^l$ and $q^{-1}b_iz_j/z_i\neq q^{-l}$ for a positive integer l and $1\leqq i,j\leqq n$.

The special cases of Theorems 1.15 and 1.17 when $b_1=\cdots=b_n=b$ have previously been obtained by Milne as Theorems 1.24 and 1.15 of [18].

We finally remark that the method of contour integration used to prove the identities in the present paper has a long history, which is discussed in Slater [22]. Slater and Lakin [23] give a contour integration proof of the ${}_6\Psi_6$ identity (1.10). The major difficulty in extending their method is to find the correct integrand which will permit an inductive proof of a $U(n)$ multilateral summation theorem, with induction on the number of summation indices allowed to take on negative integral values. The starting points for the induction are Milne's generalizations of the Gauss, the ${}_5F_4$ and the ${}_6\varphi_5$ summation theorems [20], [19]. There are also certain trigonometric function and theta function identities, Lemmas 2.14, 3.14 and 4.14, which are used in the proofs of Theorems 1.13, 1.15 and 1.17 here.

2. A generalization of the ${}_2H_2$ summation theorem.

For $1\leqq l\leqq n$, let

$$f_l(s)=(-1)^{n-l}\frac{\sin\pi s}{\sin\pi(z_l-s)}\prod_{i=1}^{n+1}\frac{\Gamma(1-\beta_i-s)}{\Gamma(1-\alpha_i-s)}$$

$$(2.1)\qquad\cdot\sum_{y_1,\cdots,y_{l-1}=0}^{\infty}\ \sum_{y_{l+1},\cdots,y_n=-\infty}^{\infty}\ \prod_{\substack{i=1\\ i\neq l}}^{n}(z_i+y_i-s)$$

$$\cdot\prod_{\substack{1\leqq i<j\leqq n\\ i,j\neq l}}((z_i+y_i)-(z_j+y_j))\prod_{i=1}^{n+1}\prod_{\substack{k=1\\ k\neq l}}^{n}\frac{(\alpha_i+z_k)_{y_k}}{(\beta_i+z_k)_{y_k}}$$

where $\beta_i = 1 - z_i$ for $i = 1, \cdots, l-1$, and $s \in \mathbb{C}$, and we assume that the series converges absolutely. We also assume that $(\beta_i + z_k)$ is not a negative integer or zero for $1 \leqq k \leqq n$, $1 \leqq i \leqq n+1$, and $(\alpha_i + z_k)$ is not a positive integer for $l \leqq k \leqq n$, $1 \leqq i \leqq n+1$.

$f_l(s)$ can also be written as

$$f_l(s) = \frac{\sin \pi s}{\sin \pi(z_l - s)} \prod_{i=1}^{n+1} \frac{\Gamma(1-\beta_i - s)}{\Gamma(1-\alpha_i - s)} \sum_{\sigma \in S_n} \varepsilon(\sigma) \sum_{y_1, \cdots, y_{l-1}=0}^{\infty} \sum_{y_{l+1}, \cdots, y_n = -\infty}^{\infty} \cdot \prod_{i=1}^{n} (c_i)^{n-\sigma(i)} \prod_{i=1}^{n+1} \prod_{\substack{k=1 \\ k \neq l}}^{n} \frac{(\alpha_i + z_k)_{y_k}}{(\beta_i + z_k)_{y_k}}, \tag{2.2}$$

where $c_i = z_i + y_i$ for $1 \leqq i \leqq n$, $i \neq l$, and $c_l = -s$ and $\varepsilon(\sigma)$ equals the sign of the permutation $\sigma \in S_n$. S_n is the symmetric group on $\{1, \cdots, n\}$.

We then obtain

$$\begin{aligned} f_l(s) = & \frac{\sin(\pi s)}{\sin(\pi(\alpha_l - s))} \prod_{i=1}^{n+1} \frac{\Gamma(1-\beta_i - s)}{\Gamma(1-\alpha_i - s)} \\ & \cdot \sum_{\sigma \in S_n} \varepsilon(\sigma)(-s)^{n-\sigma(l)} \prod_{k=1}^{l-1} \left[\sum_{y_k=0}^{\infty} \prod_{i=1}^{n+1} \frac{(\alpha_i + z_k)_{y_k}}{(\beta_i + z_k)_{y_k}} (z_k + y_k)^{n-\sigma(k)} \right] \\ & \cdot \prod_{k=l+1}^{n} \left[\sum_{y_k=-\infty}^{\infty} \prod_{i=1}^{n+1} \frac{(\alpha_i + z_k)_{y_k}}{(\beta_i + z_k)_{y_k}} (z_k + y_k)^{n-\sigma(k)} \right] \end{aligned} \tag{2.3}$$

with assumptions as above (cf. Lemma 2.1 of [18]).

The series

$$\sum_{y_k=-\infty}^{\infty} \prod_{i=1}^{n+1} \frac{(\alpha_i + z_k)_{y_k}}{(\beta_i + z_k)_{y_k}} (z_k + y_k)^{n-\sigma(k)}$$

for $l < k \leqq n$ and the similar series for $1 \leqq k < l$ converges absolutely whenever $\operatorname{Re}(\sigma(k) - n + \sum_{i=1}^{n+1} (\beta_i - \alpha_i)) > 1$. The proof of this fact is standard and is given for the ordinary ${}_2F_1$ hypergeometric series on p. 46 of [21].

It follows that the original series expression (2.1) for $f_l(s)$ converges absolutely whenever

$$\operatorname{Re}\left(\sum_{i=1}^{n+1} (\beta_i - \alpha_i) \right) > n. \tag{2.4}$$

For $1 < l < n$, we consider the contour integral $(-1/2\pi i) \int_C f_l(s)\, ds$, where $C = C(w)$, $w \in \mathbb{R}$, $w > 0$, is the sum of the directed line segments going from $-iw - w$ to $-iw + w$, from $-iw + w$ to $iw + w$, from $iw + w$ to $iw - w$, and from $iw - w$ to $-iw - w$. We will let $w \to \infty$ through the values w_0, $w_0 + 1$, $w_0 + 2, \cdots$, where w_0 is chosen so that the contours $C(w)$ avoid the poles of $f_l(s)$.

We will need asymptotic estimates for the integrand $f_l(s)$ when $\operatorname{Re}(s) \geqq 0$ and when $\operatorname{Re}(s) \leqq 0$. We use the reflection formula $[\Gamma(z)\Gamma(1-z)]^{-1} = \pi^{-1} \sin(\pi z)$, as well as Sterling's formula

$$\Gamma(a+s) = \sqrt{2\pi} s^{a+s-1/2} e^{-s} \left(1 + O\left(\frac{1}{s} \right) \right)$$

as $s \to \infty$ in $S_\theta = \{s\colon |\arg s| < \theta\}$, $0 < \theta < \pi$, and the estimates

$$|\sin \pi s| = O(e^{\pi |\operatorname{Im} s|})$$

in the entire plane, and

$$|\sin \pi(a-s)|^{-1} = O(e^{-\pi|\mathrm{Im}\, s|})$$

in the whole plane excluding ε-neighborhoods of the poles.

These formulas together with (2.3) imply that

$$|f_l(s)| = O(s^{n-1+\mathrm{Re}\,(\sum_{i=1}^{n+1}(\alpha_i-\beta_i))}) \tag{2.5}$$

as $s \to \infty$ with $\mathrm{Re}\,(s) \geqq 0$ (or $\mathrm{Re}\,(s) \leqq 0$) excluding an ε-neighborhood of the poles. If $\mathrm{Re}\,(\sum_{i=1}^{n+1}(\beta_i-\alpha_i)) > n$, it follows that

$$\lim_{w\to\infty} -\frac{1}{2\pi i}\int_C f_l(s)\,ds = 0. \tag{2.6}$$

Before going further we shall state Milne's $U(n)$ generalization of the Gauss summation theorem [20], which is used to start the inductive proof of Theorem 1.11.

THEOREM 2.7. *Let* $\mathrm{Re}\,(\gamma_{n+1}-b+\sum_{i=1}^{n}(\gamma_i-z_i)) > 0$. *Then*

$$\begin{aligned} &\sum_{y_1,\cdots,y_n=0}^{\infty} \prod_{1\leqq i<j\leqq n}\left(\frac{(z_i+y_i)-(z_j+y_j)}{z_i-z_j}\right)\prod_{l=1}^{n}\left[\frac{(z_l-\gamma_{n+1})_{y_l}}{(z_l-b)_{y_l}}\prod_{k=1}^{n}\frac{(z_l-\gamma_k)_{y_l}}{(1+z_l-z_k)_{y_l}}\right] \\ &\qquad = \frac{\Gamma\left(\gamma_{n+1}-b+\sum_{i=1}^{n}(\gamma_i-z_i)\right)}{\Gamma(\gamma_{n+1}-b)}\prod_{i=1}^{n}\frac{\Gamma(z_l-b)}{\Gamma(\gamma_l-b)}, \end{aligned} \tag{2.8}$$

where we assume that z_i-z_j *and* z_i-b *are not negative integers or zero for* $1\leqq i\neq j\leqq n$.

Remark 2.9. Theorem 2.7 can also be obtained from a limit (by sending $z_n\to\infty$) of a $U(n)$ Dougall theorem ([10, Cor. 3.1]).

Proof of Theorem 1.11. We prove identity (1.12) by induction on l, $1\leqq l\leqq n$. Suppose (1.12) is true for all l satisfying $n\geqq l>t$, where t is an integer, $1\leqq t\leqq n$. We shall then prove (1.12) for $l=t$.

Setting $l=t$ and $\mathrm{Re}\,(\sum_{i=1}^{n+1}(\beta_i-\alpha_i))>n$ and assumptions as in (1.12), we compute the residues of the integral

$$\frac{-1}{2\pi i}\int_C f_l(s)\,ds. \tag{2.10}$$

First consider the residues at $s=1-\beta_m+y_l$, where $1\leqq m\leqq n+1$ and y_l is a nonnegative integer. We then sum over $y_l\geqq 0$:

$$\begin{aligned} &\frac{\sin(\pi\beta_m)}{\sin(\pi(z_l+\beta_m-1))}\prod_{\substack{i=1\\ i\neq m}}^{n+1}\frac{\Gamma(1-\beta_i-(1-\beta_m))}{\Gamma(1-\alpha_i-(1-\beta_m))} \\ &\quad\cdot\frac{1}{\Gamma(1-\alpha_m-(1-\beta_m))}\sum_{y_1,\cdots,y_l=0}^{\infty}\ \sum_{y_{l+1},\cdots,y_n=-\infty}^{\infty} \\ &\quad\cdot\prod_{\substack{1\leqq i<j\leqq n\\ i,j\neq l}}((z_i+y_i)-(z_j+y_j))\prod_{\substack{i=1\\ i\neq l}}^{n}((z_i+y_i)-(1-\beta_m+y_l)) \\ &\quad\cdot\prod_{i=1}^{n+1}\left[\frac{(\alpha_i+\beta_m+1)_{y_l}}{(\beta_i-\beta_m+1)_{y_l}}\prod_{k=1}^{l-1}\frac{(\alpha_i+z_k)_{y_k}}{(1-z_i+z_k)_{y_k}}\prod_{k=l+1}^{n}\frac{(\alpha_i+z_k)_{y_k}}{(\beta_i+z_k)_{y_k}}\right], \end{aligned} \tag{2.11}$$

recalling that $\beta_i=1-z_i$ for $1\leqq i<l$ and assuming temporarily that $\beta_i-\beta_m$ is not an integer for all $1\leqq i$, $m\leqq n+1$.

If $l = n$, then we can sum the series (2.11) by the $U(n)$ Gauss summation Theorem 2.7. If $l < n$, then we can sum the series (2.11) by the induction hypothesis for (1.12). Expression (2.11) now becomes

$$\begin{aligned}
&= \frac{\Gamma\left(-n+\sum_{i=1}^{n+1}(\beta_i-\alpha_i)\right)\prod_{i=1}^{n+1}\prod_{k=1,k\neq l}^{n}[\Gamma(\beta_i+z_k)\Gamma(1-z_k-\alpha_i)]}{\prod_{i,j=1}^{n+1}\Gamma(\beta_i-\alpha_j)\prod_{1\leqq i<j\leqq n, i,j\neq l}[\Gamma(1-z_i+z_j)\Gamma(z_i-z_j)]} \\
(2.12)\qquad &\cdot \frac{\prod_{i=1}^{n+1}\Gamma(\beta_i+1-\beta_m)\prod_{i=1,i\neq m}^{n+1}\Gamma(\beta_m-\beta_i)}{\prod_{j=1}^{l-1}[\Gamma(2-z_j-\beta_m)\Gamma(z_j+\beta_m-1)]\prod_{j=l+1}^{n}[\Gamma(\beta_m+z_j)\Gamma(1-\beta_m-z_j)]} \\
&\cdot \frac{\sin\pi(\beta_m)}{\sin\pi(z_l+\beta_m-1)}.
\end{aligned}$$

When $1 \leqq m < l$, then expression (2.11) vanishes. This is because then $z_m = 1-\beta_m$ and (2.11) becomes both symmetric and skew-symmetric under the transposition $y_m \leftrightarrow y_l$.

If we sum expression (2.12) over m, $l \leqq m \leqq n+1$, we obtain

$$\begin{aligned}
(2.13)\qquad &\frac{\Gamma\left(-n+\sum_{i=1}^{n+1}(\beta_i-\alpha_i)\right)\prod_{i=1}^{n+1}\prod_{k=1,k\neq l}^{n}[\Gamma(\beta_i+z_k)\Gamma(1-z_k-\alpha_i)]}{\prod_{i,j=1}^{n+1}\Gamma(\beta_i-\alpha_j)\prod_{1\leqq i<j\leqq n, i,j\neq l}[\Gamma(1-z_i+z_j)\Gamma(z_i-z_j)]} \\
&\cdot(-\pi)\sum_{m=l}^{n+1}\frac{\sin\pi(\beta_m)\prod_{j=l+1}^{n}\sin\pi(\beta_m+z_j)}{\sin\pi(\beta_m+z_l)\prod_{i=l,i\neq m}^{n+1}\sin\pi(\beta_m-\beta_i)}.
\end{aligned}$$

To finish the proof of (1.12) we will need the following

LEMMA 2.14. *With assumptions as above we have*

$$(2.15)\qquad \sum_{m=l}^{n+1}\frac{\sin\pi(\beta_m)\prod_{j=l+1}^{n}\sin\pi(\beta_m+z_j)}{\sin\pi(\beta_m+z_l)\prod_{i=l,i\neq m}^{n+1}\sin\pi(\beta_m-\beta_i)} = \frac{\sin\pi(z_l)\prod_{j=l+1}^{n}\sin\pi(z_l-z_j)}{\prod_{i=l}^{n+1}\sin\pi(\beta_i+z_l)}.$$

Proof. Using the identity $\sin\theta = (e^{i\theta}-e^{-i\theta})/2i$ and setting $e^{\pi i\beta_k} = g_k$ and $e^{\pi i z_j} = a_j$ for $l \leqq k \leqq n+1$ and $l \leqq j \leqq n$, then we are reduced to verifying

$$(2.16)\qquad \sum_{m=l}^{n+1}\frac{(g_m-g_m^{-1})\prod_{j=l+1}^{n}(g_m a_j-g_m^{-1}a_j^{-1})}{(g_m a_l-g_m^{-1}a_l^{-1})\prod_{i=l,i\neq m}^{n+1}(g_m g_i^{-1}-g_m^{-1}g_i)} = \frac{(a_l-a_l^{-1})\prod_{j=l+1}^{n}(a_l a_j^{-1}-a_l^{-1}a_j)}{\prod_{i=l}^{n+1}(g_i a_l-g_i^{-1}a_l^{-1})}.$$

Identity (2.16) can be proved directly using the Louck and Biedenharn lemma (see

[16, Thm. 1.20]). It can also be obtained by considering the residues of the contour integral

$$(2.17)\qquad \frac{1}{2\pi i}\int_D \frac{(z-1)\prod_{j=l+1}^{n}(za_j-a_j^{-1})}{(za_l-a_l^{-1})\prod_{i=l}^{n+1}(zg_i^{-1}-g_i)}\,dz,$$

where D is a sufficiently large circle containing the poles of the integrand and traversed in the counterclockwise direction. Q.E.D.

Returning to the contour integral (2.10), we consider its residues at $s=z_l+y_l$ where y_l is an integer. Summing these residues over y_l, $-\infty<y_l<\infty$, we obtain

$$(2.18)\qquad \begin{aligned} &\sin \pi(z_l)\prod_{i=1}^{n+1}\left[\frac{\Gamma(1-\beta_i-z_l)}{\Gamma(1-\alpha_i-z_l)}\right]\sum_{y_1,\cdots,y_{l-1}=0}^{\infty}\ \sum_{y_l,\cdots,y_n=-\infty}^{\infty}\\ &\cdot\prod_{1\leqq i<j\leqq n}((z_i+y_i)-(z_j+y_j))\prod_{i=1}^{n+1}\prod_{k=1}^{n}\frac{(\alpha_i+z_k)_{y_k}}{(\beta_i+z_k)_{y_k}}, \end{aligned}$$

where $\beta_i=1-z_i$ for $1\leqq i<l$. By an argument similar to that for expressions (2.2) and (2.3), the series (2.18) converges absolutely whenever $\operatorname{Re}\left(\sum_{i=1}^{n+1}(\beta_i-\alpha_i)\right)>n$.

The limit (2.6) implies that the limit of the sum of the residues of (2.10), i.e., the sum of (2.13) and (2.18), is zero; after substituting (2.15) into (2.13) and simplifying, we prove the inductive step $l=t$ for (1.12). The assumption made in (2.11) that $\beta_i-\beta_m$ is not an integer for $1\leqq i$, $m\leqq n+1$ can be dropped by continuity. This completes the proof of (1.12) by induction.

3. A generalization of the ${}_5H_5$ summation theorem. For $1\leqq l\leqq n-1$, let

$$(3.1)\qquad \begin{aligned} &h_l(s;y_1,\cdots,y_{l-1},y_{l+1},\cdots,y_{n-1})\\ &\quad=h_l(s;y)\\ &\quad=(-1)^{n-l-1}\frac{\sin\pi(z_n+z_l-s)}{\sin\pi(z_l-s)}(s-u)\prod_{\substack{i=1\\ i\neq l}}^{n-1}[(z_i+y_i-s)(z_i+y_i-u)]\\ &\qquad\cdot\prod_{i=1}^{n}\left[\frac{\Gamma(1-b_i+z_i-s)\,\Gamma(1-b_i+z_i-u)}{\Gamma(1-a_i+z_i-s)\,\Gamma(1-a_i+z_i-u)}\right]\\ &\qquad\cdot\prod_{\substack{1\leqq i,j\leqq n-1\\ i,j\neq l}}((z_i+y_i)-(z_j+y_j))\prod_{i=1}^{n}\prod_{\substack{j=1\\ j\neq l}}^{n-1}\frac{(a_i-z_i+z_j)_{y_j}}{(b_i-z_i+z_j)_{y_j}}, \end{aligned}$$

where we assume that $a_i-z_i+z_j$ for $1\leqq i\leqq n$, $l\leqq j\leqq n$, is not a positive integer and $b_i-z_i+z_j$ is not a negative integer or zero for $1\leqq i, j\leqq n$. We also assume $y_1,\cdots,y_{l-1}$ are nonnegative integers and $y_{l+1},\cdots,y_{n-1}$ integers and

$$(3.2)\qquad s+u+\sum_{\substack{i=1\\ i\neq l}}^{n-1} y_i=z_n+z_l.$$

For $1\leqq l\leqq n-1$, we consider the contour integral

$$\frac{-1}{2\pi i}\int_C h_l(s;y)\,ds,$$

where $C=C(w)$ is defined as in § 2. As in § 2 we will let $w\to\infty$ through values w_0, w_0+1, $w_0+2,\cdots$, where w_0 is chosen so that the contours $C(w)$ avoid the poles of $h_l(s;y)$ for all y.

By an argument similar to § 2 one shows that

$$|h_l(s;y)| = O(s^{2n-3+2\,\mathrm{Re}\,(\sum_{i=1}^{n}(a_i-b_i))}) \tag{3.3}$$

as $s\to\infty$ with $\mathrm{Re}\,(s)\geqq 0$ (or $\mathrm{Re}\,(s)\leqq 0$) excluding an ε-neighborhood of the poles.

It follows that if $\mathrm{Re}\,(\sum_{i=1}^{n}(b_i-a_i))>n-1$, then

$$\lim_{w\to\infty}\frac{-1}{2\pi i}\int_C h_l(s;y)\,ds=0. \tag{3.4}$$

Proof of Theorem 1.13. We shall prove the absolute convergence of the series in (1.14) for $\mathrm{Re}\,(\sum_{i=1}^{n}(b_i-a_i))>n-1$ in Lemma 3.19.

To prove (1.14) we make the induction hypothesis that $b_i=1$ for $1\leqq i\leqq l-1$. This will reduce the summation in (1.14) to $y_1,\cdots,y_{l-1}\geqq 0$ and $y_l,\cdots,y_n\in\mathbb{Z}$. The other terms will vanish.

For $l=n$ identity (1.14) is an immediate consequence of Theorem 7.30 of [19]. (Note that the proof of Theorem 7.30 is incomplete. However, the proof is correct in the case of a terminating series. The general result follows by an application of Carlson's theorem [7, p. 39].) Now suppose (1.14) is true for all l satisfying $n\geqq l>t$, where t is an integer, $1\leqq t\leqq n$. We shall then prove (1.14) for $l=t$.

Consider the residue of the integral $(-1/2\pi i)\int_C h_l(s;y)\,ds$ at $s=1-b_m+z_m+y_l$ for $1\leqq m\leqq n$ and $y_l\geqq 0$, and sum over $y_1,\cdots,y_l\geqq 0$ and $y_{l+1},\cdots,y_n\in\mathbb{Z}$, where $y_1+\cdots+y_n=0$. Abbreviating $(1-b_i+z_i)=c_i$ and $b_i-1-z_i+z_l+z_n=d_i$ for $1\leqq i\leqq n$, we obtain

$$\begin{aligned}
&(-1)^{n-l-1}\frac{\sin\pi(z_n+z_l-c_m)}{\sin\pi(z_l-c_m)}\prod_{\substack{i=1\\ i\neq m}}^{n}\Gamma(c_i-c_m)\\
&\cdot\prod_{i=1}^{n}\frac{\Gamma(c_i-d_m)}{\Gamma(1-a_i+z_i-c_m)\Gamma(1-a_i+z_i-d_m)}\\
&\cdot\sum_{y_1,\cdots,y_l=0}^{\infty}\ \sum_{\substack{y_{l+1},\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty}\ \prod_{\substack{1\leqq i<j\leqq n-1\\ i,j\neq l}}((z_i+y_i)-(z_j+y_j))\\
&\cdot\prod_{\substack{i=1\\ i\neq l}}^{n-1}[((z_i+y_i)-(c_m+y_l))((z_i+y_i)-(d_m+y_n))]((c_m+y_l)-(d_m+y_n))\\
&\cdot\prod_{i=1}^{n}\left[\frac{(a_i+z_i+c_m)_{y_l}}{(b_i-z_i+c_m)_{y_l}}\frac{(a_i-z_i+d_m)_{y_n}}{(b_i-z_i+d_m)_{y_n}}\prod_{\substack{j=1\\ j\neq l}}^{n-1}\frac{(a_i-z_i+z_j)_{y_j}}{(b_i-z_i+z_j)_{y_j}}\right].
\end{aligned} \tag{3.5}$$

We denote expression (3.5) by

$$\frac{\sin\pi(z_n-z_l-z_m+b_m-1)}{\sin\pi(z_l-z_m+b_m-1)}R_m. \tag{3.6}$$

Under the condition $\mathrm{Re}\,(\sum_{i=1}^{n}(b_i-a_i))>n-1$, it follows from Lemma 3.19 below that R_m is absolutely convergent.

If we consider the residue of $(-1/2\pi i)\int_C h_l(s;y)\,ds$ at $u=1-b_m+z_m+y_l$ where u satisfies (3.2) and sum over $y_1,\cdots,y_l\geqq 0$ and $y_{l+1},\cdots,y_n\in\mathbb{Z}$ with $y_1+\cdots+y_n=0$ as above, we obtain

$$\frac{\sin\pi(1-b_m+z_m)}{\sin\pi(1-b_m+z_m-z_n)}R_m. \tag{3.7}$$

Finally, observe that since $b_i = 1$ for $1 \leqq i \leqq l-1$, then R_m is both symmetric and skew-symmetric upon interchange of y_m and y_l for $1 \leqq m \leqq l-1$. Hence

$$R_m = 0 \quad \text{for } 1 \leqq m \leqq l-1. \tag{3.8}$$

Now consider the residue of $(-1/2\pi i) \int_C h_l(s; y)\, ds$ at $s = z_l + y_l$, where $y_l \in Z$, and sum over $y_1, \cdots, y_{l-1} \geqq 0$ and $y_l, \cdots, y_n \in \mathbb{Z}$ with $y_1 + \cdots + y_n = 0$. We obtain

$$\begin{aligned}&\frac{\sin(\pi z_n)}{\pi} \prod_{i=1}^{n} \left[\frac{\Gamma(1-b_i+z_i-z_l)}{\Gamma(1-a_i+z_i-z_l)} \frac{\Gamma(1-b_i+z_i-z_n)}{\Gamma(1-a_i+z_i-z_n)} \right] \\ &\cdot \sum_{y_1,\cdots,y_{l-1}=0}^{\infty} \sum_{\substack{y_l,\cdots,y_n=-\infty \\ y_1+\cdots+y_n=0}}^{\infty} \prod_{1 \leqq i<j \leqq n} ((z_i+y_i)-(z_j+y_j)) \\ &\cdot \prod_{i,j=1}^{n} \frac{(a_i-z_i+z_j)_{y_j}}{(b_i-z_i+z_j)_{y_j}}.\end{aligned} \tag{3.9}$$

We denote expression (3.9) by

$$\frac{\sin(\pi z_n)}{\pi} \prod_{i=1}^{n} \left[\frac{\Gamma(1-b_i+z_i-z_l)}{\Gamma(1-a_i+z_i-z_l)} \frac{\Gamma(1-b_i+z_i-z_n)}{\Gamma(1-a_i+z_i-z_n)} \right] B_l. \tag{3.10}$$

Again, it follows from Lemma 3.19 that B_l is absolutely convergent under the condition $\operatorname{Re}(\sum_{i=1}^{n}(b_i-a_i)) > n-1$.

Now the fact the series B_l and R_m for $1 \leqq m \leqq n$ are absolutely convergent and that $\lim_{w \to 0} -1/2\pi i \int_C h_l(s; y)\, ds = 0$ imply that

$$\begin{aligned}&\sum_{y_1,\cdots,y_{l-1} \geqq 0} \sum_{y_{l+1},\cdots,y_{n-1}=-\infty}^{\infty} \lim_{w \to \infty} \frac{-1}{2\pi i} \int_C h_l(s; y)\, ds \\ &= \sum_{m=1}^{n} \left[\frac{\sin \pi(z_n+z_l-z_m+b_m-1)}{\sin \pi(z_l-z_m+b_m-1)} + \frac{\sin \pi(1-b_m+z_m)}{\sin \pi(1-b_m+z_m-z_n)} \right] R_m \\ &+ \frac{\sin \pi(z_n)}{\pi} \prod_{i=1}^{n} \left[\frac{\Gamma(1-b_i+z_i-z_l)}{\Gamma(1-a_i+z_i-z_l)} \frac{\Gamma(1-b_i+z_i-z_n)}{\Gamma(1-a_i+z_i-z_n)} \right] B_l = 0\end{aligned} \tag{3.11}$$

where $R_m = 0$ for $1 \leqq m < l$.

We abbreviate $1-b_i+z_i = c_i$ and $b_i-1-z_i+z_l+z_n = d_i$ for $1 \leqq i \leqq n$ as above. By the induction hypothesis and writing $b_m - z_m$ as $1-c_m$, then R_m for $l \leqq m \leqq n$ can be summed. Using that $[\Gamma(u)\Gamma(1-u)]^{-1} = \pi^{-1} \sin \pi u$, we obtain for $l \leqq m \leqq n$

$$\begin{aligned}R_m = &\prod_{\substack{i=1 \\ i \neq m}}^{n} \Gamma(c_i-c_m) \prod_{i=1}^{n} \Bigg[\Gamma(c_i-d_m)\Gamma(b_i-z_i+c_m)\Gamma(b_i-z_i+d_m) \\ &\cdot \prod_{\substack{k=1 \\ k \neq l}}^{n-1} \Gamma(1-a_i+z_i-z_k)\Gamma(b_i-z_i+z_k) \Bigg] \\ &\cdot \prod_{\substack{1 \leqq i<j \leqq n-1 \\ i,j \neq l}} \pi^{-1} \sin \pi(z_i-z_j) \prod_{i=1}^{l} \pi^{-1} \sin \pi(z_i-c_m) \\ &\cdot \prod_{j=l+1}^{n-1} \pi^{-1} \sin \pi(c_m-z_j) \prod_{\substack{i=1 \\ i \neq l}}^{n-1} \pi^{-1} \sin \pi(z_i-d_m) \\ &\cdot \frac{\pi^{-1} \sin \pi(c_m-d_m)}{\prod_{i,k=1}^{n} \Gamma(b_i-a_k-z_i+z_k)} \frac{\Gamma\left(1-n+\sum_{i=1}^{n}(b_i-a_i)\right)}{\Gamma\left(1-\sum_{i=1}^{n} a_i\right)\Gamma\left(1-n+\sum_{i=1}^{n} b_i\right)}\end{aligned} \tag{3.12}$$

where we have used that $c_m+d_m=z_m+z_l$ to show that

$$\Gamma\left(1-\sum_{i=1}^{n}a_i-c_m-d_m+z_n+z_l\right)=\Gamma\left(1-\sum_{i=1}^{n}a_i\right)$$

and similarly for

$$\Gamma\left(1-n+\sum_{i=1}^{n}b_i\right).$$

From (3.12) it follows that

$$\begin{aligned}
&\sum_{m=1}^{n}\left[\frac{\sin\pi(z_n+z_l-z_m+b_m-1)}{\sin\pi(z_l-z_m+b_m-1)}+\frac{\sin\pi(1-b_m+z_m)}{\sin\pi(1-b_m+z_m-z_n)}\right]R_m\\
&\quad=\sum_{m=l}^{n}\left\{\left[\frac{\sin\pi(z_n+z_l-c_m)}{\sin\pi(z_l-c_m)}+\frac{\sin\pi(z_n+z_l-d_m)}{\sin\pi(z_l-d_m)}\right]\right.\\
&\qquad\left.\cdot\frac{\prod_{i=l+1}^{n-1}\sin\pi(z_i-c_m)\sin\pi(z_i-d_m)}{\prod_{i=l,i\neq m}^{n}\sin\pi(b_i-z_i+c_m)\sin\pi(b_i-z_i+d_m)}\right\}\\
&\qquad\cdot\frac{(-1)^{n-l-1}\pi^2\prod_{i=1}^{n}\prod_{k=1,k\neq l}^{n-1}\Gamma(1-a_i+z_i-z_k)\Gamma(b_i-z_i+z_k)}{\Gamma\left(1-\sum_{i=1}^{n}a_i\right)\Gamma\left(1-n+\sum_{i=1}^{n}b_i\right)\prod_{i,k=1}^{n}\Gamma(b_i-a_k-z_i+z_k)}\\
&\qquad\cdot\Gamma\left(1-n+\sum_{i=1}^{n}(b_i-a_i)\right)\prod_{\substack{1\leqq i<j\leqq n-1\\ i,j\neq l}}\pi^{-1}\sin\pi(z_i-z_j),
\end{aligned}\tag{3.13}$$

where we have used the induction hypothesis that $b_i=1$ for $1\leqq i\leqq l-1$.

To finish the proof of (1.14) we need the following:

LEMMA 3.14. *With assumptions as above we have*

$$\begin{aligned}
&\sum_{m=l}^{n}\left[\frac{\sin\pi(z_n+z_l-c_m)}{\sin\pi(z_l-c_m)}+\frac{\sin\pi(z_n+z_l-d_m)}{\sin\pi(z_l-d_m)}\right]\\
&\qquad\cdot\frac{\prod_{i=l+1}^{n-1}\sin\pi(z_i-c_m)\sin\pi(z_i-d_m)}{\prod_{\substack{i=l\\ i\neq m}}^{n}\sin\pi(b_i-z_i+c_m)\sin\pi(b_i-z_i+d_m)}\\
&\quad=\frac{\prod_{i=l+1}^{n-1}\sin\pi(z_i-z_l)\sin\pi(z_i-z_n)}{\prod_{i=l}^{n}\sin\pi(z_l-c_i)\sin\pi(d_i-z_l)}\cdot\sin\pi(z_n-z_l)\sin\pi(z_n).
\end{aligned}\tag{3.15}$$

Proof. The proof is very similar to that of Lemma 2.14. Setting $e^{\pi i a_j} = \alpha_j$, $e^{\pi i b_j} = \beta_j$, $e^{\pi i z_j} = \gamma_j$, $e^{\pi i c_j} = \tau_j$, and $e^{\pi i d_j} = \delta_j$ for $1 \leqq j \leqq n$, then we are reduced to verifying

$$\tag{3.16} \begin{aligned} &\sum_{m=l}^{n} \left[\frac{(\gamma_n \gamma_l \tau_m^{-1} - \gamma_n^{-1} \gamma_l^{-1} \tau_m)}{(\gamma_l \tau_m^{-1} - \gamma_l^{-1} \tau_m)} + \frac{(\gamma_n \gamma_l \delta_m^{-1} - \gamma_n^{-1} \gamma_l^{-1} \delta_m)}{(\gamma_l \delta_m^{-1} - \gamma_l^{-1} \delta_m)} \right] \\ &\qquad \cdot \frac{\prod_{i=l+1}^{n-1} (\gamma_i \tau_m^{-1} - \gamma_i^{-1} \tau_m)(\gamma_i \delta_m^{-1} - \gamma_i^{-1} \delta_m)}{\prod_{i \neq l, i = m}^{n} (\beta_i \tau_m \gamma_i^{-1} - \beta_i^{-1} \tau_m^{-1} \gamma_i)(\beta_i \delta_m \gamma_i^{-1} - \beta_i^{-1} \delta_m^{-1} \gamma_i)} \\ &= \frac{\prod_{i=l+1}^{n-1} (\gamma_i \gamma_l^{-1} - \gamma_i^{-1} \gamma_l)(\gamma_i \gamma_n^{-1} - \gamma_i^{-1} \gamma_n)}{\prod_{i=l}^{n} (\gamma_l \tau_i^{-1} - \gamma_l^{-1} \tau_i)(\delta_i \gamma_l^{-1} - \delta_i^{-1} \gamma_l)} (\gamma_n \gamma_l^{-1} - \gamma_n^{-1} \gamma_l)(\gamma_n - \gamma_n^{-1}). \end{aligned}$$

Identity (3.16) can be obtained by considering the residues of the contour integral

$$\tag{3.17} \begin{aligned} &\frac{1}{2\pi i} \int_D \frac{\prod_{i=l+1}^{n-1} [(\gamma_i - \gamma_i^{-1} s)(\gamma_i \gamma_n^{-1} \gamma_l^{-1} s - \gamma_i^{-1} \gamma_n \gamma_l)]}{\prod_{i=l}^{n} [(s \tau_i^{-1} - \tau_i)(\delta_i - \delta_i^{-1} s)]} \\ &\qquad \cdot \frac{(\gamma_n \gamma_l - \gamma_n^{-1} \gamma_l^{-1} s^2)(\gamma_n \gamma_l - \gamma_n^{-1} \gamma_l^{-1} s)}{(\gamma_l - \gamma_l^{-1} s)} \, ds, \end{aligned}$$

where D is a sufficiently large circle containing the poles of the integrand and traversed in the counterclockwise direction. Q.E.D.

Substituting identity (3.15) into (3.13) we obtain

$$\tag{3.18} \begin{aligned} &\sum_{m=l}^{n} \left[\frac{\sin \pi(z_n + z_l - z_m + b_m - 1)}{\sin \pi(z_l - z_m + b_m - 1)} + \frac{\sin \pi(1 - b_m + z_m)}{\sin \pi(1 - b_m + z_m - z_n)} \right] R_m \\ &= \frac{-\pi^{-1} \sin \pi(z_n) \prod_{i=1}^{n} \prod_{k=1, k \neq l}^{n-1} \Gamma(1 - a_i + z_i - z_k) \Gamma(b_i - z_i + k)}{\prod_{i,k=1}^{n} \Gamma(b_i - a_k - z_i + z_k) \prod_{1 \leqq i < j \leqq n-1, i, j \neq l} \Gamma(z_i - z_j) \Gamma(1 - z_i + z_j)} \\ &\quad \cdot \frac{\prod_{i=1}^{n} \Gamma(z_l - c_i) \Gamma(1 - z_l + c_i) \Gamma(d_i - z_l) \Gamma(1 - d_i + z_l)}{\Gamma\left(1 - \sum_{i=1}^{n} a_i\right) \Gamma\left(1 - n + \sum_{k=1}^{n} b_i\right) \Gamma(z_l - z_n) \Gamma(1 - z_l + z_n)} \\ &\quad \cdot \frac{\Gamma\left(1 - n + \sum_{i=1}^{n} (b_i - a_i)\right)}{\prod_{i=l+1}^{n-1} \Gamma(z_l - z_i) \Gamma(1 - z_l + z_i) \Gamma(z_i - z_n) \Gamma(1 - z_i + z_n)}. \end{aligned}$$

Now substituting identity (3.18) into (3.11) and using that $b_i = 1$, $1 \leqq i \leqq l - 1$, we obtain identity (1.14). Q.E.D.

LEMMA 3.19. *With assumptions as in Theorem* 1.13 *the series*

$$\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty} \prod_{1\leqq i<j\leqq n} ((z_i+y_i)-(z_j+y_j)) \prod_{i,k=1}^{n} \frac{(a_i-z_i+z_k)_{y_k}}{(b_i-z_i+z_k)_{y_k}} \tag{3.20}$$

converges absolutely whenever $\operatorname{Re}(\sum_{i=1}^n (b_i-a_i))>n-1$.

Proof. (cf. [18, Lemma 2.6].) The proof is similar to that for the convergence of the series (2.1). With notation as above and as in (2.3), we write

$$\begin{aligned}&\prod_{1\leqq i<j\leqq n} ((z_i+y_i)-(z_j+y_j)) \prod_{i,k=1}^{n} \frac{(a_i-z_i+z_k)_{y_k}}{(b_i-z_i+z_k)_{y_k}}\\ &\quad=\sum_{\sigma\in S_n} \varepsilon(\sigma) \prod_{k=1}^{n}\left[(z_{\sigma(k)}+y_{\sigma(k)})^{n-k}\prod_{i=1}^{n}\frac{(a_i-z_i+z_{\sigma(k)})_{y_{\sigma(k)}}}{(b_i-z_i+z_{\sigma(k)})_{y_{\sigma(k)}}}\right].\end{aligned} \tag{3.21}$$

Using the well-known identity

$$\Gamma(z)=\lim_{n\to\infty}\frac{(n-1)!\,n^z}{(z)_n} \tag{3.22}$$

[21, p. 11], one shows that

$$\lim_{|y_{\sigma(1)}|\to\infty} (z_{\sigma(1)}+y_{\sigma(1)})^{n-1} \prod_{i=1}^{n}\frac{(a_i-z_i+z_{\sigma(1)})_{y_{\sigma(1)}}}{(b_i-z_i+z_{\sigma(1)})_{y_{\sigma(1)}}}=0 \tag{3.23}$$

since $\operatorname{Re}(\sum_{i=1}^n (b_i-a_i))>n-1$. Hence

$$\left|(z_{\sigma(1)}+y_{\sigma(1)})^{n-1} \prod_{i=1}^{n}\frac{(a_i-z_i+z_{\sigma(1)})_{y_{\sigma(1)}}}{(b_i-z_i+z_{\sigma(1)})_{y_{\sigma(1)}}}\right|\leqq M \tag{3.24}$$

for some constant $M>0$ independent of $y_{\sigma(1)}$. It now follows that the series (3.20) converges absolutely whenever the following series converges:

$$\sum_{\sigma\in S_n} M \prod_{k=2}^{n}\left[\sum_{y_{\sigma(k)}=-\infty}^{\infty}\left|(z_{\sigma(k)}+y_{\sigma(k)})^{n-k}\prod_{i=1}^{n}\frac{(a_i-z_i+z_{\sigma(k)})_{y_{\sigma(k)}}}{(b_i-z_i+z_{\sigma(k)})_{y_{\sigma(k)}}}\right|\right]. \tag{3.25}$$

Since $\operatorname{Re}(\sum_{i=1}^n (b_i-a_i))>n-1$, then each interior series of (3.25) for $k=2,\cdots,n$ converges absolutely by a standard argument (see [21, p. 46]). Q.E.D.

4. A generalization of the ${}_6\Psi_6$ summation theorem.

In this section we consider basic hypergeometric series in $U(n)$, with base q such that $|q|<1$. Similar to § 3, let

$$\begin{aligned}&k_l(s;y_1,\cdots,y_{l-1},y_{l+1},\cdots,y_{n-1})=k_l(s;y)\\ &\qquad=(-1)^{n-l-1}\frac{[z_nz_l/s]_\infty[qs/(z_nz_l)]_\infty[z_nz_l/u]_\infty[qu/(z_nz_l)]_\infty}{[z_l/s]_\infty[qs/z_l]_\infty[z_l/u]_\infty[qu/z]_\infty}\\ &\qquad\cdot\left(z_n\prod_{i=1}^{n}(b_i/a_i)\right)^{\sum_{i=1,i\neq l}^{n-1}y_i}\prod_{\substack{i=1\\ i\neq l}}^{n-1}(z_iq^{y_i}-s)(z_iq^{y_i}-u)\\ &\qquad\cdot\prod_{i=1}^{n}\frac{[qa_i^{-1}z_is^{-1}]_\infty[qa_i^{-1}z_iu^{-1}]_\infty}{[qb_i^{-1}z_is^{-1}]_\infty[qb_i^{-1}z_iu^{-1}]_\infty}\prod_{\substack{1\leqq i<j\leqq n-1\\ i,j\neq l}}(z_iq^{y_i}-z_jq^{y_j})\prod_{i=1}^{n}\prod_{\substack{j=1\\ j\neq l}}^{n-1}\frac{[a_iz_j/z_i]_{y_j}}{[b_iz_j/z_i]_{y_j}},\end{aligned} \tag{4.1}$$

where we assume that $a_iz_j/z_i \neq q^k$ for $1 \leqq i \leqq n$, $l \leqq j \leqq n$, and $q^{-1}b_iz_j/z_i \neq q^{-k}$ for $1 \leqq i, j \leqq n$, where k is a positive integer. We also assume that $y_1, \cdots, y_{l-1}$ are nonnegative integers and $y_{l+1}, \cdots, y_{n-1}$ are integers and

$$\text{(4.2)} \qquad suq^{\sum_{i=1, i\neq l}^{n-1} y_i} = z_n z_l.$$

As in § 3 we consider the contour integral, for $1 \leqq l \leqq n-1$, $(-1/2\pi i) \int_C k_l(s; y)\, ds$, where $C = C_1(w) + C_2(w)$, for $w > 0$, and $C_1(w)$ is the circle of radius w^{-1} traversed in the clockwise direction and $C_2(w)$ is the circle of radius w traversed in the counterclockwise direction. We will let $w \to \infty$ through values w_0, $w_0|q|^{-1}$, $w_0|q|^{-2}, \cdots$, where w_0 is chosen so that the contours $C_1(w) + C_2(w)$ avoid the poles of $k_l(s; y)$ for all $y_1, \cdots, y_{l-1} \geqq 0$; and $y_{l+1}, \cdots, y_n \in \mathbb{Z}$. If for $a, z \in \mathbb{C}$ we define $\sigma(a, z) = [a/z]_\infty [qz/a]_\infty$, then

$$\text{(4.3)} \qquad \sigma(a, qz) = -(a/qz)\sigma(a, z).$$

From identity (4.3) one shows that for fixed $y_1, \cdots, y_{l-1}, y_{l+1}, \cdots, y_{n-1}$, we have

$$\text{(4.4)} \qquad |k_l(s; y)| = O\left(\left|q^{(2-n)|k|} \prod_{i=1}^n (b_i/a_i)\right|\right)$$

for $|q|^k \geqq |s| > |q|^{k+1}$ as $k \to \infty$ or $k \to -\infty$. It follows that if $|q^{1-n} \prod_{i=1}^n (b_i/a_i)| < 1$, then

$$\text{(4.5)} \qquad \lim_{w\to\infty} -\frac{1}{2\pi i} \int_C k_l(s; y)\, ds = 0.$$

Proof of Theorem 1.15. The proof of the absolute convergence for $|q^{1-n} \prod_{i=1}^n (b_i/a_i)| < 1$ of the series (1.16) is similar to the proof of Lemma 3.19 and will be given in Lemma 4.22.

As in the proof of Theorem 1.13, we make the induction hypothesis that $b_i = q$ for $1 \leqq i \leqq l-1$. This reduces the summation in (1.16) to $y_1, \cdots, y_{l-1} \geqq 0$ and $y_l, \cdots, y_n \in \mathbb{Z}$. The other terms vanish. Also, we shall multiply both sides of identity (1.16) by $\prod_{1 \leqq i<j \leqq n} (z_i - z_j)$. We are reduced to proving

$$\sum_{y_1, \cdots, y_{l-1}=0}^{\infty} \sum_{\substack{y_l, \cdots, y_n=-\infty \\ y_1+\cdots+y_n=0}}^{\infty} \prod_{1\leqq i<j\leqq n} (z_iq^{y_i} - z_jq^{y_j}) \prod_{i,k=1}^{n} \frac{[a_iz_j/z_i]_{y_j}}{[b_iz_j/z_i]_{y_j}}$$

$$\text{(4.6)} \qquad = \frac{[q]_\infty^{n-1} \prod_{i=1}^n z_i^{n-1} \prod_{i,k=1}^n [(b_iz_k)/(a_kz_i)]_\infty}{\prod_{i,k=1}^n [(qz_i)/(a_iz_k)]_\infty [b_iz_k/z_i]_\infty}$$

$$\cdot \prod_{1\leqq i<j\leqq n} [qz_i/z_j]_\infty [z_j/z_i]_\infty \frac{\left[q\Big/\prod_{i=1}^n a_i\right]_\infty \left[q^{1-n} \prod_{i=1}^n b_i\right]_\infty}{\left[q^{1-n} \prod_{i=1}^n (b_i/a_i)\right]_\infty}.$$

For $l = n$, identity (4.6) is an immediate consequence of Theorem 1.44 of [19]. Now suppose (4.6) is true for all l satisfying $n \geqq l > t$, where t is an integer, $1 \leqq t \leqq n$. We shall then prove (4.6) for $l = t$.

As in § 3 we consider the residue of the integral

$$\frac{-1}{2\pi i} \int_C k_l(s; y)\, ds \text{ at } s = z_m q^{y_l+1} b_m^{-1},$$

for $1 \leqq m \leqq n$ and $y_l \geqq 0$, and sum over $y_1, \cdots, y_l \geqq 0$ and $y_{l+1}, \cdots, y_n \in \mathbb{Z}$, where

$y_1+\cdots+y_n=0$. Abbreviating $qb_i^{-1}z_i=c_i$ and $b_iq^{-1}z_i^{-1}z_lz_n=d_i$, for $1\leqq i\leqq n$, we obtain

$$
\begin{aligned}
R_m={}&\frac{(-1)^{n-l-1}[z_nz_l/c_m]_\infty[qc_m/(z_nz_l)]_\infty[z_nz_l/d_m]_\infty[qd_m/(z_nz_l)]_\infty}{[q]_\infty[z_l/c_m]_\infty[qc_m/z_l]_\infty[z_l/d_m]_\infty[qd_m/z_l]_\infty}\\
&\cdot\prod_{i=1}^{n}\frac{[qz_i/(a_ic_m)]_\infty[qz_i/(a_id_m)]_\infty}{[c_i/d_m]_\infty}\prod_{\substack{i=1\\ i\neq m}}^{n}\frac{1}{[c_i/c_m]_\infty}\\
&\cdot\sum_{\substack{y_1,\cdots,y_l=0\\ y_1+\cdots+y_n=0}}^{\infty}\;\sum_{y_{l+1},\cdots,y_n=-\infty}^{\infty}c_mq^{y_l}\prod_{\substack{1\leqq i<j\leqq n-1\\ i,j\neq l}}(z_iq^{y_i}-z_jq^{y_j})\\
&\cdot\prod_{\substack{i=1\\ i\neq l}}^{n-1}[(z_iq^{y_i}-c_mq^{y_l})(z_iq^{y_i}-d_mq^{y_n})]\\
&\cdot\prod_{i=1}^{n}\prod_{\substack{j=1\\ j\neq l}}^{n-1}\frac{[a_iz_j/z_i]_{y_j}}{[b_iz_j/z_i]_{y_j}}\prod_{i=1}^{n}\frac{[a_ic_m/z_i]_{y_l}[a_id_m/z_i]_{y_n}}{[b_ic_m/z_i]_{y_l}[b_id_m/z_i]_{y_n}}.
\end{aligned}
\tag{4.7}
$$

Under the condition $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<1$, it follows as in Lemma 4.22 below that R_m is absolutely convergent.

If we consider the residue of $(-1/2\pi i)\int_C k_l(s;y)\,ds$ at $u=z_mq^{y_l+1}\cdot b_m^{-1}$, where u satisfies (4.2), and sum over $y_1,\cdots,y_l\geqq 0$ and $y_{l+1},\cdots,y_n\in\mathbb{Z}$ with $y_1+\cdots+y_n=0$, we obtain a similar series R'_m. The series R'_m is also absolutely convergent under the condition $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<1$. We have

$$
\begin{aligned}
R_m+R'_m={}&\frac{(-1)^{n-l-1}[z_nz_l/c_m]_\infty[qc_m/(z_nz_l)]_\infty[z_nz_l/d_m]_\infty[qd_m/(z_nz_l)]_\infty}{[q]_\infty[z_l/c_m]_\infty[qc_m/z_l]_\infty[z_l/d_m]_\infty[qd_m/z_l]_\infty}\\
&\cdot\prod_{i=1}^{n}\frac{[qz_i/(a_ic_m)]_\infty[qz_i/(a_id_m)]_\infty}{[c_i/d_m]_\infty}\prod_{\substack{i=1\\ i\neq m}}^{n}\frac{1}{[c_i/c_m]_\infty}\\
&\cdot\sum_{\substack{y_1,\cdots,y_l=0\\ y_1+\cdots+y_n=0}}^{\infty}\;\sum_{y_{l+1},\cdots,y_n=-\infty}^{\infty}(c_nq^{y_l}-d_mq^{y_n})\prod_{\substack{1\leqq i<j\leqq n-1\\ i,j\neq l}}(z_iq^{y_i}-z_jq^{y_j})\\
&\cdot\prod_{\substack{i=1\\ i\neq l}}^{n-1}[(z_iq^{y_i}-c_mq^{y_l})(z_iq^{y_i}-d_mq^{y_n})]\\
&\cdot\prod_{i=1}^{n}\prod_{\substack{j=1\\ j\neq l}}^{n-1}\frac{[a_iz_j/z_i]_{y_j}}{[b_iz_j/z_i]_{y_j}}\prod_{i=1}^{n}\frac{[a_ic_m/z_i]_{y_l}[a_id_m/z_i]_{y_n}}{[b_ic_m/z_i]_{y_l}[b_id_m/z_i]_{y_n}}.
\end{aligned}
\tag{4.8}
$$

By a symmetry argument similar to that for identity (3.8) of § 3, one shows that

$$R_m+R'_m=0\quad\text{for } 1\leqq m\leqq l-1. \tag{4.9}$$

Now considering the residues of $(-1/2\pi i)\int_C k_l(s;y)\,ds$ at $s=z_lq^{y_l}$ and $u=z_lq^{y_l}$ for $y_l\in\mathbb{Z}$ and summing over $y_1,\cdots,y_{l-1}\geqq 0$ and $y_l,\cdots,y_n\in\mathbb{Z}$ with $z_1+\cdots+y_n=0$, we obtain

$$
\begin{aligned}
B_l={}&\frac{[z_n]_\infty[q/z_n]_\infty[z_l]_\infty[q/z_l]_\infty}{[q]_\infty^2[z_l/z_n]_\infty[qz_n/z_l]_\infty}\prod_{i=1}^{n}\frac{[qa_i^{-1}z_iz_l^{-1}]_\infty[qa_i^{-1}z_iz_n^{-1}]_\infty}{[c_iz_l^{-1}]_\infty[c_iz_n^{-1}]_\infty}\\
&\cdot\sum_{\substack{y_1,\cdots,y_{l-1}=0\\ y_1+\cdots+y_n=0}}^{\infty}\;\sum_{y_l,\cdots,y_n=-\infty}^{\infty}\prod_{1\leqq i<j\leqq n}(z_iq^{y_i}-z_jq^{y_j})\prod_{i,j=1}^{n}\frac{[a_iz_j/z_i]_{y_j}}{[b_iz_j/z_i]_{y_j}}.
\end{aligned}
\tag{4.10}
$$

B_l is absolutely convergent under the condition $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<1$.

As in § 3, the fact that the series B_l and R_m, R'_m for $1 \leqq m \leqq n$ are absolutely convergent and that $\lim_{w\to\infty}(-1/2\pi i)\int_C k_l(s; y)\,ds = 0$ imply that

$$\sum_{y_1,\cdots,y_{l-1}\geqq 0}^{\infty}\ \sum_{y_{l+1},\cdots,y_{n-1}=-\infty}^{\infty} \lim_{w\to\infty}\frac{-1}{2\pi i}\int_C k_l(s; y)\,ds = B_l + \sum_{m=1}^{n}(R_m + R'_m) = 0. \tag{4.11}$$

By the induction hypothesis $R_m + R'_m$ for $l \leqq m \leqq n$ can be summed by (4.6):

$$\begin{aligned}
R_m + R'_m = {}& \frac{[q]_\infty^{n-2}[z_n z_l/c_m]_\infty[qc_m/z_n z_l]_\infty}{[z_l/c_m]_\infty[qc_m/z_l]_\infty}(c_m)^{n-l} \\
& \cdot \frac{[z_n z_l/d_m]_\infty[qd_m/z_n z_l]_\infty}{[z_l/d_m]_\infty[qd_m/z_l]_\infty}\prod_{i=1}^{n}\frac{[qz_i/a_i c_m]_\infty[qz_i/a_i d_m]_\infty}{[c_i/d_m]_\infty} \\
& \cdot \prod_{\substack{i=1\\ i\neq m}}^{n}\frac{1}{[c_i/c_m]_\infty}\prod_{i=1}^{n-1} z_i^{n-1}\prod_{i,k=1}^{n}[(b_i z_k)/(a_k z_i)]_\infty \\
& \cdot \frac{\displaystyle\prod_{1\leqq i<j\leqq n-1,i,j\neq l}[qz_i/z_j]_\infty[z_j/z_i]_\infty}{\displaystyle\prod_{i,k=1,k\neq l,n}^{n}[(qz_i)/(a_i z_k)]_\infty[b_i z_k/z_i]_\infty} \\
& \cdot \frac{\displaystyle\prod_{i=1}^{l-1}[qz_i/c_m]_\infty[c_m/z_i]_\infty\prod_{j=l+1}^{n-1}[qc_m/z_j]_\infty[z_j/c_m]_\infty}{\displaystyle\prod_{i=1}^{n}[qz_i/(a_i c_m)]_\infty[qz_i/(a_i d_m)]_\infty[b_i c_m/z_i]_\infty[b_i d_m/z_i]_\infty} \\
& \cdot \frac{\displaystyle\prod_{i=1,i\neq l}^{n-1}[qz_i/d_m]_\infty[d_m/z_i]_\infty\cdot\left[q\Big/\prod_{i=1} a_i\right]_\infty\left[q^{1-n}\prod_{i=1} b_i\right]_\infty}{\left[q^{1-n}\displaystyle\prod_{i=1}^{n}(b_i/a_i)\right]_\infty} \\
& \cdot [qc_m/d_m]_\infty[d_m/c_m]_\infty
\end{aligned} \tag{4.12}$$

$$\begin{aligned}
= {}& c_m^{n-l}(-d_m/c_m)(z_n z_l)^{l-1}\prod_{i=1}^{l-1} z_i^{-2}\prod_{i=1}^{n-1} z_i^{n-i} \\
& \cdot \frac{\displaystyle\prod_{j=l+1}^{n-1}[qc_m/z_j]_\infty[z_j/c_m]_\infty[qz_j/d_m]_\infty[d_m/z_j]_\infty}{\displaystyle\prod_{i=l,i\neq m}^{n}[c_i/c_m]_\infty[qc_m/c_i]_\infty[c_i/d_m]_\infty[qd_m/c_i]_\infty} \\
& \cdot \frac{[z_n z_l/c_m]_\infty[qc_m/(z_n z_l)]_\infty[z_n z_l/d_m]_\infty[qd_m/(z_n z_l)]_\infty}{[z_l/c_m]_\infty[qc_m/z_l]_\infty[z_l/d_m]_\infty[qd_m/z_l]_\infty} \\
& \cdot \frac{\displaystyle\prod_{i,k=1}^{n}[(b_i z_k)/(a_k z_i)]_\infty\prod_{1\leqq i<j\leqq n-1,i,j\neq l}[qz_i/z_j]_\infty[z_j/z_i]_\infty}{\left[q^{1-n}\displaystyle\prod_{i=1}^{n}(b_i/a_i)\right]_\infty\displaystyle\prod_{i=1}^{n}\prod_{k=1,k\neq l}^{n-1}[(qz_i)/(a_i z_k)]_\infty[b_i z_k/z_i]_\infty} \\
& \cdot \left[q\Big/\prod_{i=1} a_i\right]_\infty\left[q^{1-n}\prod_{i=1}^{n} b_i\right]_\infty[q]_\infty^{n-3}.
\end{aligned} \tag{4.13}$$

The following lemma will be used to compute the sum $\sum_{m=l}^{n}(R_m + R'_m)$.

LEMMA 4.14. *Assume that c_i, d_i, $z_i \in \mathbb{C}$, $l \leqq i \leqq n$, are defined as above and define*

$$E_m = c_m^{n-l}\left(\frac{-d_m}{c_m}\right)\frac{\prod\limits_{j=l+1}^{n-1}[qc_m/z_j]_\infty[z_j/c_m]_\infty[qz_j/d_m]_\infty[d_m/z_j]_\infty}{\prod\limits_{i=l,i\neq m}^{n}[c_i/c_m]_\infty[qc_m/c_i]_\infty[c_i/d_m]_\infty[qd_m/c_i]_\infty} \tag{4.15a}$$
$$\cdot\frac{[z_nz_l/c_m]_\infty[qc_m/(z_nz_l)]_\infty[z_nz_l/d_m]_\infty[qd_m/(z_nz_l)]_\infty}{[z_l/c_m]_\infty[qc_m/z_l]_\infty[z_l/d_m]_\infty[qd_m/z_l]_\infty},$$

for $l \leqq m \leqq n$, and

$$E_{n+1} = (z_l)^{n-l}\left(\frac{-z_n}{z_l}\right)\frac{\prod\limits_{j=l+1}^{n-1}[qz_l/z_j]_\infty[z_j/z_l]_\infty[qz_j/z_n]_\infty[z_n/z_j]_\infty}{\prod\limits_{i=1}^{n}[c_i/z_l]_\infty[qz_l/c_i]_\infty[c_i/z_n]_\infty[qz_n/c_i]_\infty} \tag{4.15b}$$
$$\cdot [z_n]_\infty[q/z_n]_\infty[z_l]_\infty[q/z_l]_\infty.$$

Then

$$E = \sum_{m=l}^{n+1} E_m = 0, \tag{4.16}$$

where it is assumed that the denominators of the E_m, $l \leqq m \leqq n+1$, do not vanish.

Proof. Consider E and E_m, $l \leqq m \leqq n+1$, as functions of the variables $c_l, \cdots, c_n$, $z_l, \cdots, z_n \in \mathbb{C}^*$ (where $\mathbb{C}^* = \mathbb{C} - \{0\}$). Observe that for $l \leqq s \leqq n$ and $l \leqq m \leqq n+1$, we have

$$\begin{aligned} &E_m(c_l, \cdots, c_{s-1}, qc_s, c_{s+1}, \cdots, c_n, z_l, \cdots, z_n) \\ &\quad = \left(\frac{c_s}{d_s}\right) E_m(c_l, \cdots, c_{s-1}, c_s, c_{s+1}, \cdots, z_n). \end{aligned} \tag{4.17}$$

We first show that $E(c_l, \cdots, c_n, z_l, \cdots, z_n)$ is holomorphic in the region $(\mathbb{C}^*)^{2(n-l+1)}$. The possible poles of E occur when $c_s = q^k c_t$, $c_s = q^k d_t$, $c_s = q^k z_l$ or $c_s = q^k z_n$ for $l \leqq s$, $t \leqq n$, $s \neq t$, and k an integer. It follows from (4.17) that if suffices to show that E has no poles for $k = 0$.

One now verifies that for $l \leqq s, t \leqq n$ and $s \neq t$, we have $\lim_{c_s \to c_t}(E_s + E_t)$ and $\lim_{c_s \to d_t}(E_s + E_t)$ exist and are finite. Since E_m for $l \leqq m \leqq n+1$ and $m \neq s, t$ is continuous as $c_s \to c_t$ and $c_s \to d_t$, it follows that E has no pole at $c_s = c_t$ or at $c_s = d_t$. Similarly for $l \leqq s \leqq n$ we verify that $\lim_{c_s \to z_l}(E_s + E_{n+1})$ exist and are finite, while E_m is continuous as $c_s \to z_l$ and $c_s \to z_n$ for $l \leqq m \leqq n$ and $m \neq s$. Hence E has no pole at $c_s = z_l$ or at $c_s = z_n$ and $E(c_l, \cdots, z_n)$ is holomorphic in the region $(\mathbb{C}^*)^{2(n-l+1)}$.

Now choose an integer s, $l \leqq s \leqq n$, and fix the variables $c_l, \cdots, c_{s-1}, c_{s+1}, \cdots, c_n$, $z_l, \cdots, z_n \in \mathbb{C}^*$. Let

$$M = \max_{|q| \leqq |c_s| \leqq 1} |E(c_l, \cdots, c_n, z_l, \cdots, z_n)|. \tag{4.18}$$

An application of identity (4.17) yields

$$\begin{aligned} &E(c_l, \cdots, c_{s-1}q^k c_s, c_{s+1}, \cdots, c_n, z_l, \cdots, z_n) \\ &\quad = q^{k(k-1)}\left(\frac{c_s}{d_s}\right)^k E(c_l, \cdots, c_{s-1}, c_s, c_{s+1}, \cdots, z_n) \end{aligned} \tag{4.19}$$

where k is an integer. By the maximum principle for holomorphic functions (in the variable c_s), we find

$$\max_{|q|^{h+1}\leqq|c_s|\leqq|q|^{-j}} |E(c_l, \cdots, z_n)| \leqq M \max\left(|q|^{h(h-1)}|z_n z_l|^{-h}, |q|^{j(j+1)}|z_n z_l|^{j}\right), \tag{4.20}$$

where j and h are nonnegative integers. It follows that for fixed $(c_l, \cdots, c_{s-1}, c_{s+1}, \cdots, c_n, z_l, \cdots, z_n)$ we have $\lim_{c_s\to 0} E=0$ and $\lim_{c_s\to\infty} E=0$. Hence by the maximum principle E must be identically 0. This completes the proof of Lemma 4.14.

Putting together expressions (4.11), (4.13), and (4.16), we obtain

$$\begin{aligned}
-\sum_{m=l}^{n}(R_m+R'_m) = B_l &= \frac{-z_n}{z_l}\prod_{i=1}^{n-1} z_i^{n-1}\prod_{i=1}^{l-1}\left(\frac{z_n z_l}{z_i^2}\right)[q]_\infty^{n-3} \\
&\cdot\frac{\prod_{j=l+1}^{n-1}[qz_l/z_j]_\infty[z_j/z_l]_\infty[qz_j/z_n]_\infty[z_n/z_j]_\infty}{\prod_{i=l}^{n}[c_i/z_l]_\infty[qz_l/c_i]_\infty[c_i/z_n]_\infty[qz_n/c_i]_\infty} \\
&\cdot\frac{\prod_{i,k=1}^{n}[(b_i z_k)/(a_k z_i)]_\infty \prod_{1\leqq i<j\leqq n-1, i,j\neq l}[qz_i/z_j]_\infty[z_j/z_i]_\infty}{\left[q^{1-n}\prod_{i=1}^{n}(b_i/a_i)\right]_\infty \prod_{i=1}^{n}\prod_{k=1,k\neq l}^{n-1}[(qz_i)/(a_i z_k)]_\infty[b_i z_k/z_i]_\infty} \\
&\cdot [z_n]_\infty[q/z_n]_\infty[z_l]_\infty[q/z_l]_\infty\left[q\Big/\prod_{i=1}^{n}a_i\right]_\infty\left[q^{1-n}\prod_{i=1}^{n}b_i\right]_\infty.
\end{aligned} \tag{4.21}$$

Equating expressions (4.10) and (4.21) for B_l, we obtain after a little algebra identity (4.6). Theorem 1.15 now follows by induction once we prove the previously promised Lemma 4.22 for the absolute convergence of the series (4.6).

LEMMA 4.22. *With assumptions as in Theorem* 1.15, *the series*

$$\sum_{\substack{y_1,\cdots,y_n=-\infty \\ y_1+\cdots+y_n=0}}^{\infty}\prod_{1\leqq i<j\leqq n}(z_i q^{y_i}-z_j q^{y_j})\prod_{i,k=1}^{n}\frac{[a_i z_k/z_i]_{y_k}}{[b_i z_k/z_i]_{y_k}} \tag{4.23}$$

converges absolutely whenever $|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)|<1$.

Proof. The proof is similar to that of Lemma 3.19 with some modification. We write

$$\begin{aligned}
&\prod_{1\leqq i<j\leqq n}(z_i q^{y_i}-z_j q^{y_j})\prod_{i,k=1}^{n}\frac{[a_i z_k/z_i]_{y_k}}{[b_i z_k/z_i]_{y_k}} \\
&\quad=\sum_{\sigma\in S_n}\varepsilon(\sigma)\prod_{k=1}^{n}(z_{\sigma(k)}q^{y_{\sigma(k)}})^{n-k}\prod_{i=1}^{n}\frac{[a_i z_{\sigma(k)}/z_i]_{y_{\sigma(k)}}}{[b_i z_{\sigma(k)}/z_i]_{y_{\sigma(k)}}}.
\end{aligned} \tag{4.24}$$

We have

$$\lim_{y_{\sigma(n)}\to\infty}\left|\prod_{i=1}^{n}\frac{[a_i z_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}{[b_i z_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}\right| = \prod_{i=1}^{n}\left|\frac{[a_i z_{\sigma(n)}/z_i]_\infty}{[b_i z_{\sigma(n)}/z_i]_\infty}\right| \tag{4.25a}$$

and

$$\lim_{y_{\sigma(n)}\to-\infty}\left|\prod_{i=1}^{n}\frac{[a_i z_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}{[b_i z_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}\right| = \lim_{y_{\sigma(n)_n}\to-\infty}\prod_{i=1}^{n}\left|\frac{b_i}{a_i}\right|^{-y_{\sigma(n)}}\left|\frac{[qz_i/(b_i z_{\sigma(n)})]_{-y_{\sigma(n)}}}{[qz_i/(a_i z_{\sigma(n)})]_{-y_{\sigma(n)}}}\right| = 0, \tag{4.25b}$$

since $|\prod_{i=1}(b_i/a_i)|<|q|^{n-1}$. Hence

$$\left|\prod_{i=1}^{n}\frac{[a_iz_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}{[b_iz_{\sigma(n)}/z_i]_{y_{\sigma(n)}}}\right|\leqq M \tag{4.26}$$

for some constant $M>0$ independent of $y_{\sigma(n)}$. It follows that the series (4.24) converges absolutely whenever the following series converges:

$$\sum_{\sigma\in S_n} M\prod_{i=1}^{n-1}\left[\sum_{y_{\sigma(n)}=-\infty}^{\infty}\left|[(z_{\sigma(k)}q^{y_{\sigma(k)}})^{n-k}\prod_{i=1}^{n}\frac{[a_iz_{\sigma(k)}/z_i]_{y_{\sigma(k)}}}{[b_iz_{\sigma(k)}/z_i]_{y_{\sigma(k)}}}\right|\right]. \tag{4.27}$$

Since $|\prod_{i=1}^{n}(b_i/a_i)|<|q|^{n-1}$, then each interior series for $k=1,\cdots,n-1$ converges absolutely by an application of the ratio test to the sums for $y_{\sigma(k)}\geqq 0$ and $y_{\sigma(k)}<0$. Q.E.D.

5. A generalization of the ${}_1\Psi_1$ summation theorem.

Proof of Theorem 1.17. The notation and assumptions below are as in the statement of Theorem 1.17.

First, by an argument virtually identical to that of Lemma 2.5 of [18], it follows that the series (1.18) converges for

$$\left|q^{1-n}\prod_{i=1}^{n}(b_i/a_i)\right|<|t|<1.$$

We rewrite the series in (1.18) as

$$\prod_{1\leqq i<j\leqq n}(z_i-z_j)^{-1}\sum_{M=-\infty}^{\infty}t^M\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1,\cdots,y_n=M}}^{\infty}\prod_{1\leqq i<j\leqq n}(z_iq^{y_i}-z_jq^{y_j})\prod_{i,k=1}^{n}\frac{[a_iz_k/z_i]_{y_k}}{[b_iz_k/z_i]_{y_k}}. \tag{5.1}$$

Set $y_1'=y_1-M$, $z_1'=z_1q^M$, $a_1'=a_1q^M$, $b_1'=b_1q^M$, and $y_l'=y_l$, $z_l'=z_l$, $a_l'=a_l$, $b_l'=b_l$ for $2\leqq l\leqq n$. Then expression (5.1) becomes

$$\begin{aligned}&=\prod_{1\leqq i<j\leqq n}(z_i-z_j)^{-1}\\ &\quad\cdot\sum_{M=-\infty}^{\infty}\left\{t^M\prod_{i=1}^{n}\frac{[a_iz_1/z_i]_M}{[b_iz_1/z_i]_M}\right.\\ &\qquad\cdot\left.\sum_{\substack{y_1',\cdots y_n'=-\infty\\ y_1'+\cdots+y_n'=0}}^{\infty}\prod_{1<i<j\leqq n}(z_i'q^{y_1'}-z_j'q^{y_1'})\prod_{i,k=1}^{n}\frac{[a_i'z_k'/z_i']_{y_k'}}{[b_i'z_k'/z_i']_{y_k'}}\right\}.\end{aligned} \tag{5.2}$$

Applying Theorem 1.15 and some algebra we obtain

$$\begin{aligned}&=\frac{[q]_\infty^{n-1}\prod_{i,k=1}^{n}[(b_iz_k)/(a_kz_i)]_\infty\prod_{1<i<j\leqq n}[qz_i/z_j]_\infty[qz_j/z_i]_\infty}{\left[q^{1-n}\prod_{i=1}^{n}(b_i/a_i)\right]_\infty\prod_{i,k=1}^{\infty}[(qz_i)/(a_iz_k)]_\infty[b_iz_k/z_i]_\infty}\\ &\quad\cdot\left[q\Big/\prod_{i=1}^{n}a_i\right]_\infty\left[q^{1-n}\prod_{i=1}^{n}b_i\right]_\infty\sum_{M=-\infty}^{\infty}\frac{\left[\prod_{i=1}^{n}a_i\right]_M}{\left[q^{1-n}\prod_{i=1}^{n}b_i\right]_M}t^M.\end{aligned} \tag{5.3}$$

Identity (1.18) and Theorem 1.17 now follow by an application of the classical ${}_1\Psi_1$ summation theorem (1.9). Q.E.D.

REFERENCES

[1] G. E. Andrews, *Applications of basic hypergeometric functions*, SIAM Rev., 16 (1974), pp. 441-484.

[2] ———, *Problems and prospects for basic hypergeometric functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191-224.

[3] ———, *The Theory of Partitions*, Vol. 2, in Encyl. of Math. and its Appl., G.-C. Rota, ed., Addison-Wesley, Reading, MA, 1976.

[4] R. Askey, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346-359.

[5] ———, *An elementary evaluation of a beta type integral*, Indian J. Pure Appl. Math., 14 (1983), pp. 892-895.

[6] R. Askey and M. Ismail, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, Birkhaüser, Basel, 1983, pp. 55-78.

[7] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge Math. Tract No. 32, Cambridge University Press, Cambridge, 1935. (Reprinted, Hafner, New York, 1964.)

[8] ———, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math., 7 (1936), pp. 105-115.

[9] J. Dougall, *On Vandermonde's theorem and some more general expansions*, Proc. Edinburgh Math. Soc., 25 (1907), pp. 114-132.

[10] R. Gustafson, *A Whipple's transformation for hypergeometric series in $U(n)$ and multivariable hypergeometric orthogonal polynomials*, this Journal, 18 (1987), pp. 495-530.

[11] W. J. Holman III, *Summation theorems for hypergeometric series in $U(n)$*, this Journal, 11 (1980), pp. 523-532.

[12] W. J. Holman III, L. C. Biedenharn and J. D. Louck, *On hypergeometric series well-poised in $SU(n)$*, this Journal, 7 (1976), pp. 529-541.

[13] M. E. H. Ismail, *A simple proof of Ramanujan's ${}_1\Psi_1$ sum*, Proc. Amer. Math. Soc., 63 (1977), pp. 185-186.

[14] I. G. Macdonald, *Affine root systems and Dedekind's η-function*, Invent. Math., 15 (1972), pp 91-143.

[15] S. C. Milne, *An elementary proof of the Macdonald identities for $A_l^{(1)}$*, Adv. in Math., 57 (1985), pp. 34-70.

[16] ———, *A q-analog of hypergeometric series well-poised in $SU(n)$ and invariant G-functions*, Adv. in Math., 58 (1985), pp. 1-60.

[17] ———, *A q-analog of the ${}_5F_4$ (1) summation theorem for hypergeometric series well-poised in $SU(n)$*, Adv. in Math., 57 (1985), pp. 14-33.

[18] ———, *A $U(n)$ generalization of Ramanujan's ${}_1\Psi_1$ summation*, J. Math. Anal. Appl., 118 (1986), pp. 263-277.

[19] ———, *Basic hypergeometric series very well poised in $U(n)$*, J. Math. Anal. Appl., to appear.

[20] ———, *A q-analog of the Gauss summation theorem for hypergeometric seris in $U(n)$*, preprint.

[21] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[22] L. J. Slater, *Generalized Hypergeometric Functions*, Cambridge University Press, London-New York, 1966.

[23] L. J. Slater and A. Lakin, *Two proof of the ${}_6\Psi_6$ summation theorem*, Proc. Edinburgh Math. Soc., 1956, pp. 116-121.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

008

ASYMPTOTIC BEHAVIOUR OF THE COEFFICIENTS OF SOME SEQUENCES OF POLYNOMIALS*

W. VAN ASSCHE†, G. FANO‡ AND F. ORTOLANI§

Abstract. The asymptotic behaviour of the Taylor coefficients of a sequence of polynomials $\{p_n(x); n=1,2,3,\cdots\}$ is given under the conditions that all the zeros are negative and that the limit $p_n(x)^{1/n} (n\to\infty)$ exists for $x>0$. The result is then applied to orthogonal polynomials and to the iterations of a polynomial.

Key words. coefficients of polynomials, zeros of polynomials, orthogonal polynomials

AMS(MOS) subject classifications. Primary 26C05; secondary 33A65, 42C05

1. Introduction. Suppose we are given a sequence of monic polynomials $\{p_n(x)= x^n+\cdots; n=1,2,3,\cdots\}$. Let us denote its Taylor series around $x=0$ by

$$(1)\qquad p_n(x)=\sum_{j=0}^{n} a_{j,n}x^{n-j} \qquad (a_{0,n}=1)$$

and its zeros by $\{x_{j,n}, j=1,2,\cdots,n\}$. The Taylor coefficients and the zeros are related by the elementary symmetric relations of Viète

$$a_{j,n}=\frac{(-1)^j}{j!}\sum_{\alpha_1\neq\alpha_2\neq\cdots\neq\alpha_j} x_{\alpha_1,n}x_{\alpha_2,n}\cdots x_{\alpha_j,n},$$

and if we suppose that all the zeros of p_n are negative then

$$(2)\qquad \begin{aligned} a_{j,n} &=\frac{1}{j!}\sum_{\alpha_1\neq\alpha_2\neq\cdots\neq\alpha_j} |x_{\alpha_1,n}x_{\alpha_2,n}\cdots x_{\alpha_j,n}| \\ &= \sum_{\alpha_1<\alpha_2<\cdots<\alpha_j} |x_{\alpha_1,n}x_{\alpha_2,n}\cdots x_{\alpha_j,n}| \end{aligned}$$

from which we easily deduce the positivity of the coefficients $\{a_{j,n}, j=0,1,2,\cdots,n\}$.

If one knows the asymptotic behaviour of the polynomials p_n when n increases to infinity, then a natural question is to ask for the asymptotic behaviour of its Taylor coefficients $a_{j,n}$ for increasing n. To cover as many limits as possible we let j tend to infinity together with n in such a way that $j/n\to d\in(0,1)$. This question is analogous to a problem in physics when one wants to obtain the free energy per unit volume of an infinite system from the grand partition function [7]. In particular the analysis of the thermodynamic limit of the famous Bardeen-Cooper-Schrieffer superconductivity state was studied in [5], [6] using standard techniques in statistical mechanics. The general mathematical problem of "dense sums" (like $a_{j,n}$ when j and n tend to infinity) was investigated in [5]. Let us define the distribution of the zeros by the following:

$$(3)\qquad F_n(t)=\frac{1}{n}\{\text{number of zeros of } p_n \text{ in } (-\infty,t]\}$$

* Received by the editors March 25, 1985; accepted for publication (in revised form) November 20, 1986.

† Katholieke Universiteit Leuven, Departement Wiskunde, Celestijnenlaan 200B, B-3030 Heverlee, Belgium. This author is a research assistant of the Belgian National Fund for Scientific Research.

‡ Dipartimento di Fisica, Via Paradiso 12, Ferrara, Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Italy.

§ Dipartimento di Fisica, Via Irnerio 46, Bologna, Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Italy.

then F_n makes a jump of size k/n at a zero $x_{j,n}$ of multiplicity k. From now on we suppose that all the zeros are in $(-A, -B]$ $(0<B<A)$ so that F_n is an increasing function with $F_n(-A)=0$ and $F_n(-B)=1$ (later on we will allow B to go to zero). We will devote all our attention to those sequences of polynomials for which the sequence $\{F_n, n=1,2,\cdots\}$ converges weakly to a distribution function F, i.e.,

$$\int_{-A}^{-B} f(t)\,dF_n(t) \to \int_{-A}^{-B} f(t)\,dF(t) \qquad (n\to\infty)$$

for every (bounded and) continuous function f on $[-A, -B]$. Since

$$\frac{1}{n}\log p_n(x) = \frac{1}{n}\sum_{j=1}^{n} \log(x-x_{j,n}), \qquad x>0$$

it follows that

$$\frac{1}{n}\log p_n(x) \to \int_{-A}^{-B} \log(x-y)\,dF(y), \qquad x>0 \tag{4}$$

and in fact (4) also implies the weak convergence of F_n to F, so that this weak convergence is equivalent to the asymptotic behaviour

$$\lim_{n\to\infty} \{p_n(x)\}^{1/n} = \exp\left\{\int_{-A}^{-B} \log(x-y)\,dF(y)\right\}, \qquad x>0. \tag{5}$$

2. The main result. This paper is devoted to the proof of the following theorem.

THEOREM 1. *Let $\{p_n(x), n=1,2,3,\cdots\}$ be a sequence of polynomials for which all the zeros are in $[-A, 0]$, $(0<A)$ and for which the distribution functions F_n given in* (3) *converge weakly to a distribution function F having no jump at zero. Then there exists a concave and differentiable function g on* $(0,1)$ *such that*

$$\frac{1}{n}\log a_{j,n} \to g(d), \tag{6}$$

$$\frac{a_{j,n}}{a_{j-1,n}} \to e^{g'(d)} = f(d) \tag{7}$$

when $n\to\infty$ and $j/n\to d\in(0,1)$. The inverse function of f is given by

$$f^{-1}(x) = -\int_{-A}^{0} \frac{y}{x-y}\,dF(y), \qquad x>0 \tag{8}$$

and

$$g(d) = -(1-d)\log f(d) + \int_{-A}^{0} \log[f(d)-y]\,dF(y), \qquad 0<d<1. \tag{9}$$

We will prove this theorem using the techniques of [5], [6], [7]. The result is not true if one only assumes weak convergence. The condition that the zeros are negative is needed to make all the $a_{j,n}$ positive. If $p_n(x)$ is an even polynomial then the coefficients of x^{2j+1} are zero and the left-hand side in (6) and (7) is not defined for many j's. Roughly speaking, the proof consists of two steps. In a first step we will show that the theorem holds for a special sequence of polynomials $\{q_n(x), n=1,2,\cdots\}$ satisfying the conditions. The second step consists in proving, by a perturbation argument, that the theorem remains to be true for any sequence of polynomials satisfying the conditions. We start by assuming that all the zeros lie in $[-A, -B]$ $(0<B<A)$. The inverse F^{-1} of the distribution function F is defined as

$$F^{-1}(y)=\inf\{x\in\mathbb{R}\,|\,F(x)\geqq y\}$$

making F^{-1} a left-continuous increasing function on $(0, 1]$. Define

$$y_{j,n}=F^{-1}\left(\frac{j}{n}\right),\qquad j=1,2,\cdots,n, \tag{10}$$

then every $y_{j,n}$ is negative and belongs to $(-A, -B]$. We will consider the sequence of polynomials $\{q_n(x),\ n=1,2,3,\cdots\}$ given by

$$q_n(x)=\prod_{j=1}^{n}(x-y_{j,n})=\sum_{j=0}^{n}b_{j,n}x^{n-j}\qquad(b_{0,n}=1). \tag{11}$$

By construction we find that these polynomials satisfy the conditions of Theorem 1.

LEMMA 2. *Let* $\{b_{j,n},\ 0\leqq j\leqq n,\ n=1,2,3,\cdots\}$ *be given by* (11), *then*

$$b_{k,n+m}\geqq\sum_{i+j=k}b_{i,n}b_{j,m}. \tag{12}$$

Proof. By (2) we have

$$b_{k,n+m}=\frac{1}{k!}\sum_{\alpha_1\neq\alpha_2\neq\cdots\neq\alpha_k}\left|F^{-1}\left(\frac{\alpha_1}{n+m}\right)F^{-1}\left(\frac{\alpha_2}{n+m}\right)\cdots F^{-1}\left(\frac{\alpha_k}{n+m}\right)\right|.$$

For every integer α there exist integers α' and β' such that

$$\frac{\alpha'-1}{n}\leqq\frac{\alpha}{n+m}<\frac{\alpha'}{n},\qquad\frac{\beta'-1}{m}\leqq\frac{\alpha}{n+m}<\frac{\beta'}{m}.$$

Replace systematically every number $\alpha/(n+m)$ by α'/n (when $\beta'/m\geqq\alpha'/n$) or by β'/m (when $\alpha'/n>\beta'/m$). This gives a one-to-one mapping from $(1/(n+m), 2/(n+m),\cdots,(n+m)/(n+m))$ to $(1/n, 2/n,\cdots,n/n, 1/m, 2/m,\cdots,m/m)$ and since F^{-1} is an increasing and negative function we find that $|F^{-1}|$ is a decreasing function so that by grouping together all the terms that have i members from $\{1/n, 2/n,\cdots,n/n\}$ and $k-i$ from $\{1/m, 2/m,\cdots,m/m\}$ we find that

$$\begin{aligned}b_{k,n+m}&\geqq\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\sum_{\alpha_1'\neq\alpha_2'\neq\cdots\neq\alpha_i'}\left|F^{-1}\left(\frac{\alpha_1'}{n}\right)\cdots F^{-1}\left(\frac{\alpha_i'}{n}\right)\right|\\&\quad\cdot\sum_{\beta_1'\neq\cdots\neq\beta_{k-i}'}\left|F^{-1}\left(\frac{\beta_1'}{m}\right)\cdots F^{-1}\left(\frac{\beta_{k-i}'}{m}\right)\right|\\&=\sum_{i+j=k}\frac{1}{i!}\sum_{\alpha_1'\neq\cdots\neq\alpha_i'}|y_{\alpha_1',n}\cdots y_{\alpha_i',n}|\frac{1}{j!}\sum_{\beta_1'\neq\cdots\neq\beta_j'}|y_{\beta_1',m}\cdots y_{\beta_j',m}|\end{aligned}$$

from which we obtain (12). □

Let us now introduce a sequence of continuous functions

$$g(d,n)=\begin{cases}\dfrac{1}{n}\log b_{j,n}\quad\text{if } d=\dfrac{j}{n}\quad(0\leqq j\leqq n),\\[2ex]\dfrac{1}{n}\log b_{j,n}+(nd-j)\left(\dfrac{1}{n}\log b_{j+1,n}-\dfrac{1}{n}\log b_{j,n}\right)\\[2ex]\qquad\qquad\qquad\text{if}\ \dfrac{j}{n}<d<\dfrac{j}{n+1}\qquad(0\leqq j<n).\end{cases} \tag{13}$$

These functions are bounded from below since every $|y_{j,n}| \geqq B$ so that

$$\begin{aligned}\inf_{0\leqq d\leqq 1} g(d,n) &= \min_{0\leqq j\leqq n} \frac{1}{n}\log b_{j,n}\\ &\geqq \min_{0\leqq j\leqq n} \frac{1}{n}\log \binom{n}{j} B^j\\ &\geqq \frac{1}{n}\min_{0\leqq j\leqq n}\left(\log\binom{n}{j}+j\log B\right)\\ &\geqq \frac{1}{n}\min_{0\leqq j\leqq n}\log\binom{n}{j}+\frac{1}{n}\min_{0\leqq j\leqq n} j\log B\end{aligned}$$

and because $\binom{n}{j}\geqq 1$, we then have

$$\inf_{0\leqq d\leqq 1} g(d,n) \geqq \min(\log B, 0). \tag{14}$$

LEMMA 3. *The sequence of functions* $\{g(d,n);\ n=1,2,3,\cdots,\ d\in[0,1]\}$ *converges uniformly on every compact set of* $(0,1)$ *to a concave continuous function* g.

Proof. First consider $d\in D_\infty=\{j/2^n;\ 0\leqq j\leqq 2^n,\ n\in\mathbb{N}\}$, then there exists an $N\in\mathbb{N}$ so that $d\in D_N=\{j/2^N,\ 0\leqq j\leqq 2^N\}$, say $d=K2^{-N}$. By Lemma 2 ($m=n=2^{M+N}$; $k=K2^{M+1}$) we have

$$b_{K2^{M+1},2^{M+N+1}} \geqq b^2_{K2^M,2^{M+N}}$$

so that by taking logarithms, we obtain

$$g(d, 2^{M+N+1}) \geqq g(d, 2^{M+N}).$$

The sequence $\{g(d,2^n),\ n=N, N+1,\cdots\}$ with $d\in D_N$ is therefore increasing. We always have $|y_{j,n}|\leqq A$ so that

$$\begin{aligned}\sup_{0\leqq d\leqq 1} g(d,n) &= \max_{0\leqq j\leqq n}\frac{1}{n}\log b_{j,n}\\ &\leqq \max_{0\leqq j\leqq n}\frac{1}{n}\log\binom{n}{j}A^j\\ &\leqq \frac{1}{n}\max_{0\leqq j\leqq n}\log\binom{n}{j}+\frac{1}{n}\max_{0\leqq j\leqq n} j\log A.\end{aligned}$$

Now $\binom{n}{j}\leqq 2^n$ from which

$$\sup_{0\leqq d\leqq 1} g(d,n)\leqq \log 2+\max(0,\log A)$$

so that the sequence $\{g(d,n);\ n=1,2,3,\cdots,\ d\in[0,1]\}$ is bounded from above. Together with (14) we may then conclude that there exists a function $g(d)$ on D_∞ such that

$$\lim_{n\to\infty} g(d,2^n)=g(d),\qquad d\in D_\infty.$$

Take d_1 and d_2 in D_∞, say $d_1=K_12^{-N}$ and $d_2=K_22^{-N}$, then by Lemma 2 ($n=m=2^{M+N}$, $k=(K_1+K_2)2^M$)

$$b_{(K_1+K_2)2^M,2^{M+N+1}}\geqq b_{K_12^M,2^{M+N}}b_{K_22^M,2^{M+N}}$$

and by taking logarithms

$$g\left(\frac{d_1+d_2}{2}, 2^{M+N+1}\right) \geqq \frac{1}{2}g(d_1, 2^{M+N}) + \frac{1}{2}g(d_2, 2^{M+N}).$$

Let M go to infinity to obtain the important inequality

$$(15) \qquad g\left(\frac{d_1+d_2}{2}\right) \geqq \frac{1}{2}g(d_1) + \frac{1}{2}g(d_2), \qquad d_1, d_2 \in D_\infty.$$

Now D_∞ is dense in $[0, 1]$, and by means of (15) and the boundedness of g we therefore can extend g to be a continuous concave function on $(0, 1)$ (see for example [8] or [13]), and since the functions $g(d, n)$ are all continuous we have

$$(16) \qquad \lim_{n\to\infty} g(d, 2^n) = g(d), \qquad 0 < d < 1$$

and this will hold uniformly on every compact subset of $(0, 1)$ because the convergence is monotonic ([14, Thm. 7.13]).

In order to show that the whole sequence $\{g(d, n), n = 1, 2, \cdots\}$ converges to g uniformly on every compact interval in $(0, 1)$ we need some extra results. Denote by d_n a number in $\{j/n; j = 0, 1, \cdots, n\}$ such that $|d - d_n| \leqq 1/n$. Since $g(d, n)$ is linear between any two numbers $1/n \log b_{j,n}$ and $1/n \log b_{j-1,n}$

$$|g(d_n, n) - g(d, n)| \leqq \frac{1}{n} \max_{0<j\leqq n} |\log b_{j,n} - \log b_{j-1,n}|$$

$$= \frac{1}{n} \max_{0<j\leqq n} \left|\log \frac{b_{j,n}}{b_{j-1,n}}\right|.$$

We will see later (Lemma 5) that the sequence $\{b_{j,n}/b_{j-1,n}; j = 1, 2, \cdots, n\}$ is decreasing, hence

$$\left|\log \frac{b_{j,n}}{b_{j-1,n}}\right| \leqq \max\left(\left|\log \frac{b_{n,n}}{b_{n-1,n}}\right|, |\log b_{1,n}|\right)$$

where

$$b_{1,n} = \sum_{j=1}^{n} |y_{j,n}| \sim n \int_{-A}^{-B} |x| \, dF(x),$$

$$\frac{b_{n-1,n}}{b_{n,n}} = \frac{\sum_{j=1}^{n} |y_{1,n} \cdots y_{j-1,n} y_{j+1,n} \cdots y_{n,n}|}{|y_{1,n} \cdots y_{n,n}|}$$

$$= \sum_{j=1}^{n} \frac{1}{|y_{j,n}|} \sim n \int_{-A}^{-B} \frac{1}{|x|} \, dF(x).$$

This means that

$$\sup_{0<d<1} |g(d_n, n) - g(d, n)| = O\left(\frac{\log n}{n}\right).$$

If K is a closed interval in $(0, 1)$ and $\varepsilon > 0$, then we can choose $N \in \mathbb{N}$ such that

$$\sup_{d\in K} |g(d, 2^k) - g(d)| < \varepsilon, \qquad k > N$$

(which is possible because of the uniform convergence on K) and

$$\sup_{d\in K} |g(d_{2^k})-g(d)|<\varepsilon, \qquad k>N$$

(which can be done since y is uniformly continuous on K). Then for $k>N$

$$|g(d_{2^k}, 2^k)-g(d)|\leqq |g(d_{2^k}, 2^k)-g(d_{2^k})|+|g(d_{2^k})-g(d)|$$

and if every $d_k\in K$, then

$$\begin{aligned}\sup_{d\in K} |g(d_{2^k}, 2^k)-g(d)| &\leqq \sup_{d\in K} |g(d, 2^k)-g(d)|+\sup_{d\in K} |g(d_{2^k})-g(d)|\\ &<2\varepsilon.\end{aligned}$$

If we decompose n into its binary decomposition

$$n=\sum_{k\in n^*} 2^k$$

where the definition of n^* is obvious, then

$$\begin{aligned}\sup_{d\in K} \frac{1}{n}\sum_{k\in n^*} 2^k|g(d_{2^k}, 2^k)-g(d)| &\leqq \frac{1}{n}\sup_{d\in K}\sum_{\substack{k\in n^*\\ k\leqq N}} 2^k|g(d_{2^k}, 2^k)-g(d)|\\ &\quad+\frac{1}{n}\sup_{d\in K}\sum_{\substack{k\in n^*\\ k>N}} 2^k|g(d_{2^k}, 2^k)-g(d)|\\ &\leqq \frac{2^{N+1}}{n}\max_{k\leqq N} 2^k|g(d_{2^k}, 2^k)-g(d)|+2\varepsilon\end{aligned}$$

which means that

$$\sup_{d\in K}\frac{1}{n}\sum_{k\in n^*} 2^k|g(d_{2^k}, 2^k)-g(d)|\to 0 \qquad (n\to\infty).$$

Now use Lemma 2 to find that

$$\begin{aligned}g(d, n)-g(d) &= g(d, n)-g(d_n, n)+g(d_n, n)-g(d)\\ &\geqq g(d, n)-g(d_n, n)+\frac{1}{n}\sum_{k\in n^*} 2^k\{g(d_{2^k}, 2^k)-g(d)\}\end{aligned}$$

and with the results above

$$\liminf_{n\to\infty}\inf_{d\in K}\{g(d, n)-g(d)\}\geqq 0.$$

In a similar way we may put $k=[\log_2 n]+1$ and decompose 2^k-n into its binary decomposition

$$2^k-n=\sum_{l\in n^{**}} 2^l;$$

then from Lemma 2

$$g(d_{2^k}, 2^k)\geqq\frac{n}{2^k}g(d_n, n)+\frac{2^k-n}{2^k}g(d_{2^k-n}, 2^k-n).$$

As in the previous estimation we find that

$$\begin{aligned} g(d,n)-g(d)&=g(d,n)-g(d_n,n)+g(d_n,n)-g(d)\\ &\leqq g(d,n)-g(d_n,n)+\frac{2^k}{n}\{g(d_{2^k},2^k)-g(d)\}\\ &\quad-\frac{2^k-n}{n}\{g(d_{2^k-n},2^k-n)-g(d)\}\\ &\leqq g(d,n)-g(d_n,n)+\frac{2^k}{n}\{g(d_{2^k},2^k)-g(d)\}\\ &\quad-\frac{1}{n}\sum_{l\in n^{**}}2^l\{g(d_{2^l},2^l)-g(d)\} \end{aligned}$$

from which

$$\limsup_{n\to\infty}\sup_{d\in K}\{g(d,n)-g(d)\}\leqq 0.$$

Hence $g(d,n)$ converges uniformly to $g(d)$ on every compact subset of $(0,1)$. □

Let us now relate the function g to the weak limit F.

LEMMA 4. *The limit g of Lemma 3 is differentiable on* $(0,1)$ *and if*

(17) $$h(x)=-\int_{-A}^{-B}\frac{y}{x-y}\,dF(y),\qquad x>0,$$

then

(18) $$g(h(x))+(1-h(x))\log x=\int_{-A}^{-B}\log(x-y)\,dF(y)$$

and

(19) $$g'(h(x))=\log x.$$

Proof. Since g is a concave function, it will be differentiable except for possibly a denumerable set at points $\Lambda\subset(0,1)$. The derivative of g will be a decreasing function and the left-hand and right-hand derivatives at points in Λ exist, call then $g'_-(d)$ and $g'_+(d)$, and satisfy $g'_-(d)>g'_+(d)$. Define

$$f(d)=e^{g'(d)},\qquad d\in(0,1)\backslash\Lambda$$

then f is a decreasing function with jumps at Λ. The inverse function f^{-1} then exists, and we call this function h. The function h is increasing and therefore differentiable except for a denumerable set of points Λ^*. If $y\notin\Lambda^*$ then $h(y)\in(0,1)\backslash\Lambda$, so that g is differentiable at $h(y)$. The function $g'(d)-\log x$ (with x fixed) clearly vanishes for $d=h(x)$ (by the definition of h and f), hence

(20) $$\sup_{0<d<1}\{g(d)+(1-d)\log x\}=g(h(x))+(1-h(x))\log x$$

whenever $x\notin\Lambda^*$ (this follows since $g(d)+(1-d)\log x$ can have at most one maximum).

On the other hand, it is clear that for every $j\leqq n$

$$b_{j,n}x^{n-j}\leqq q_n(x)$$

so that

$$\frac{1}{n}\log b_{j,n}+\frac{n-j}{n}\log x\leqq\frac{1}{n}\log q_n(x).$$

Now let $n\to\infty$ and $j/n\to d$, then from the uniform convergence of $g(d,n)$ to $g(d)$ we obtain

$$\sup_{0<d<1}\{g(d)+(1-d)\log x\}\leqq\int_{-A}^{-B}\log(x-y)\,dF(y).$$

We also find that

$$q_n(x)\leqq n\max_{0\leqq j\leqq n} b_{j,n}x^{n-j}$$

so that

$$\frac{1}{n}\log q_n(x)\leqq\frac{1}{n}\log n+\max_{0\leqq j\leqq n}\left\{\frac{1}{n}\log b_{j,n}+\frac{n-j}{n}\log x\right\}.$$

Let $n\to\infty$ and $j/n\to d$, then

$$\int_{-A}^{-B}\log(x-y)\,dF(y)\leqq\sup_{0<d<1}\{g(d)+(1-d)\log x\}.$$

Combination of these inequalities and (20) obviously leads to (18). Differentiate (18) with respect to x, then (remember that (19) holds)

$$1-h(x)=\int_{-A}^{-B}\frac{x}{x-y}\,dF(y)$$

so that h is differentiable for every $x>-B$ and (17) holds. The relations (18) and (19) then hold for all $x>-B$. □

Up until now we have proved Theorem 1(6) for the special sequence $\{q_n(x), n=1,2,3,\cdots\}$. For (7) of the theorem for this sequence we need the following lemma.

LEMMA 5 ([3, Thm. 2.82]). *Let $p_n(x)=\sum_{j=0}^{n}a_{j,n}x^j$ be a polynomial of degree n with only real zeros $(a_{-1,n}=a_{n+1,n}=0)$, then*

$$(j+1)a_{j+1,n}a_{j-1,n}\leqq ja_{j,n}^2,\qquad j=1,2,\cdots,n.$$

Applied to our case, we have $b_{j,n}=a_{n-j,n}$, so that

$$\frac{b_{j+1,n}}{b_{j,n}}\leqq\frac{n-j}{n-j+1}\frac{b_{j,n}}{b_{j-1,n}}\leqq\frac{b_{j,n}}{b_{j-1,n}}\qquad(j=1,\cdots,n).$$

From this inequality one easily finds that

$$\frac{b_{j+k,n}}{b_{j,n}}\leqq\left(\frac{b_{j,n}}{b_{j-1,n}}\right)^k,\qquad j=1,\cdots,n-k+1.$$

Hence

$$\frac{b_{j,n}}{b_{j-1,n}}\geqq\left(\frac{b_{j+k,n}}{b_{j,n}}\right)^{1/k}$$
$$=\exp\left\{\frac{n}{k}\left[\frac{1}{n}\log b_{j+k,n}-\frac{1}{n}\log b_{j,n}\right]\right\}.$$

Take $k=[\varepsilon n]$ $(\varepsilon>0)$ and let $n\to\infty$ $(j/n\to d)$, then

$$\liminf\frac{b_{j,n}}{b_{j-1,n}}\geqq\exp\left\{\frac{1}{\varepsilon}[g(d+\varepsilon)-g(d)]\right\}.$$

In a similar way, using $b_{j,n}/b_{j-k,n} \geqq (b_{j,n}/b_{j-1,n})^k$, $(j = k, \cdots, n)$ we obtain

$$\limsup \frac{b_{j,n}}{b_{j-1,n}} \leqq \exp\left\{\frac{1}{\varepsilon}[g(d) - g(d-\varepsilon)]\right\}.$$

Taking $\varepsilon \to 0$ we obtain (7). This proves Theorem 1 for the sequence of polynomials $\{q_n(x),\ n = 1, 2, \cdots\}$.

In a second step we will prove that the same limits (6) and (7) hold for any sequence of polynomials $\{p_n(x),\ n = 1, 2, \cdots\}$ that satisfy the conditions of the theorem. Set

$$p_n(x) = \sum_{j=0}^{n} a_{j,n} x^{n-j} = \prod_{j=1}^{n} (x - x_{j,n})$$

then we want to compare the coefficients $a_{j,n}$ with the coefficients $b_{j,m}$, with m close to n. Let $N(n;\ \alpha, \beta)$ be the number of zeros of p_n in the interval $(\alpha, \beta]$ and $N^0(n;\ \alpha, \beta)$ the number of zeros of q_n in $(\alpha, \beta]$, then

$$\begin{aligned} N(n;\ \alpha, \beta) &= n\{F(\beta) - F(\alpha)\} + c_n(\alpha, \beta), \\ N^0(n;\ \alpha, \beta) &= n\{F(\beta) - F(\alpha)\} + c_n^0(\alpha, \beta) \end{aligned} \tag{21}$$

where $(1/n)c_n(\alpha, \beta) \to 0$ and by construction $|c_n^0(\alpha, \beta)| \leqq 1$. Hence

$$\lim_{n\to\infty} \frac{N(n;\ \alpha, \beta)}{N^0(n;\ \alpha, \beta)} = 1. \tag{22}$$

Now let $\delta_k = -B\,e^{k\varepsilon}$ $(\varepsilon > 0$ arbitrary), then $\{(\delta_i, \delta_{i-1}],\ i = 1, \cdots, N = [(1/\varepsilon)\log(A/B)] + 1\}$ covers the interval $[-A, -B]$. From this sequence we delete all the intervals for which $F(\delta_{i-1}) - F(\delta_i) = 0$, so that we have a finite number of intervals $\{(\delta_i^*, \delta_{i-1}^*],\ i = 1, \cdots, M\}$. Define

$$m = \left[n\left\{1 - \max_{1\leqq i\leqq M} \left\{ \frac{|c_n(\delta_i^*, \delta_{i-1}^*)| + 1}{n\{F(\delta_{i-1}^*) - F(\delta_i^*)\}} \right\}\right\}\right]$$

where as usual $[\alpha]$ denotes the integer part of α. Clearly we find that

$$\max_{1\leqq i\leqq M} \frac{|c_n(\delta_i^*, \delta_{i-1}^*) + 1|}{n\{F(\delta_{i-1}^*) - F(\delta_i^*)\}} \to 0 \qquad (n \to \infty)$$

so that as $n \to \infty$ the ratio m/n tends to 1. Also, by (21),

$$\begin{aligned} N(n;\ &\delta_i^*, \delta_{i-1}^*) - N^0(m;\ \delta_i^*, \delta_{i-1}^*) \\ &= (n-m)\{F(\delta_{i-1}^*) - F(\delta_i^*)\} + c_n(\delta_i^*, \delta_{i-1}^*) - c_m^0(\delta_i^*, \delta_{i-1}^*) \\ &\geqq \left\{ \max_{1\leqq j\leqq M} \frac{|c_n(\delta_j^*, \delta_{j-1}^*)| + 1}{F(\delta_{j-1}^*) - F(\delta_j^*)} \right\} \{F(\delta_{i-1}^*) - F(\delta_i^*)\} + c_n(\delta_i^*, \delta_{i-1}^*) - 1 \\ &\geqq |c_n(\delta_i^*, \delta_{i-1}^*)| + c_n(\delta_i^*, \delta_{i-1}^*) \geqq 0 \end{aligned}$$

and if $(\delta_i, \delta_{i-1}]$ is such that $F(\delta_{i-1}) - F(\delta_i) = 0$ then obviously $N^0(m;\ \delta_i, \delta_{i-1}) = 0$ so that the number of zeros of p_n in each interval $(\delta_i, \delta_{i-1}]$ $(i = 1, \cdots, N)$ is greater than or equal to the number of zeros of q_m in that interval.

Define $\{t_{j,n},\ 1 \leqq j \leqq n\} = \{y_{j,m},\ 1 \leqq j \leqq m\} \cup \{t'_{j,n},\ 1 \leqq j \leqq n-m\}$ where $t'_{j,n}$ are points such that the number of points of $\{t_{j,n}\}$ in every interval $(\delta_i, \delta_{i-1}]$ is equal to the number of zeros of p_n in that interval. If we order the zeros $x_{1,n} \leqq x_{2,n} \leqq \cdots \leqq x_{n,n}$ and $t_{1,n} \leqq t_{2,n} \leqq \cdots \leqq t_{n,n}$ then $x_{j,n}$ and $t_{j,n}$ will belong to the same interval $(\delta_i, \delta_{i-1}]$ and

$$e^{-\varepsilon} = \frac{|\delta_{i-1}|}{|\delta_i|} \leqq \frac{|t_{j,n}|}{|x_{j,n}|} \leqq \frac{|\delta_i|}{|\delta_{i-1}|} = e^{\varepsilon}.$$

Let $\{c_{j,n},\ 0 \leqq j \leqq n\}$ be the coefficients of the polynomial with $\{t_{j,n},\ 1 \leqq j \leqq n\}$ as zeros,

$$\sum_{j=0}^{n} c_{j,n} x^{n-j} = \prod_{j=1}^{n} (x - t_{j,n})$$

then

$$\frac{c_{j,n}}{a_{j,n}} = \frac{\sum_{\alpha_1 \neq \cdots \neq \alpha_j} |t_{\alpha_1,n} \cdots t_{\alpha_j,n}|}{\sum_{\alpha_1 \neq \cdots \neq \alpha_j} |x_{\alpha_1,n} \cdots x_{\alpha_j,n}|} \leqq \max_{\alpha_1 \neq \cdots \neq \alpha_j} \frac{|t_{\alpha_1,n}| \cdots |t_{\alpha_j,n}|}{|x_{\alpha_1,n}| \cdots |x_{\alpha_j,n}|}$$
$$\leqq e^{j}$$

and similarly $c_{j,n}/a_{j,n} \geqq e^{-j\varepsilon}$, hence

$$\begin{aligned} -d\varepsilon &\leqq \liminf \left\{ \frac{1}{n} \log c_{j,n} - \frac{1}{n} \log a_{j,n} \right\} \\ &\leqq \limsup \left\{ \frac{1}{n} \log c_{j,n} - \frac{1}{n} \log a_{j,n} \right\} \leqq d\varepsilon \end{aligned} \tag{23}$$

where $n \to \infty$ and $j/n \to d$. Denote the polynomial with zeros $\{t'_{j,n}\}$ by

$$\sum_{j=0}^{n-m} c'_{j,n} x^{n-m-j} = \prod_{j=1}^{n-m} (x - t'_{j,n})$$

then, since $\sum_{j=0}^{n} c_{j,n} x^{n-j}$ is the product of the polynomials q_m and $\sum_{j=0}^{n-m} c'_{j,n} x^{n-m-j}$ their coefficients are related by

$$c_{j,n} = \sum_{i=0}^{n} c'_{i,n} b_{j-i,m}. \tag{24}$$

For the coefficients $\{c'_{j,n}, 0 \leqq j \leqq n-m\}$ we have the obvious inequality

$$\begin{aligned} c'_{j,n} &= \frac{1}{j!} \sum_{\alpha_1 \neq \cdots \neq \alpha_j} |t'_{\alpha_1,n} \cdots t'_{\alpha_j,n}| \\ &\leqq \frac{1}{j!} \left(\sum_{i=1}^{n-m} |t'_{i,n}| \right)^{j} \\ &\leqq \frac{\{(n-m)A\}^{j}}{j!} \end{aligned}$$

so that from (24) we have on one hand (for $j \leqq m$)

$$\frac{c_{j,n}}{b_{j,m}} \geqq c'_{0,n} = 1$$

and on the other hand

$$\begin{aligned} \frac{c_{j,n}}{b_{j,m}} &\leqq \sum_{i=0}^{n} \frac{b_{j-i,m}}{b_{j,m}} \frac{\{(n-m)A\}^{i}}{i!} \\ &\leqq \sum_{i=0}^{\infty} \left(\frac{b_{j-1,m}}{b_{j,m}} \right)^{i} \frac{\{(n-m)A\}^{i}}{i!} \\ &= \exp\left\{ \frac{b_{j-1,m}}{b_{j,m}} (n-m)A \right\} \end{aligned}$$

so that when $n \to \infty$ and $j/n \to d$

$$0 \leqq \limsup \left\{ \frac{1}{n} \log c_{j,n} - \frac{m}{n} \frac{1}{m} \log b_{j,m} \right\} \leqq \lim \left\{ \frac{n-m}{n} A \frac{b_{j-1,m}}{b_{j,m}} \right\}$$

from which

$$\lim \frac{1}{n} \log c_{j,n} = \lim \frac{1}{m} \log b_{j,m} = g(d).$$

This means that the limit of $(1/n) \log c_{j,n}$ is independent of ε and since (23) is valid for every $\varepsilon > 0$ we then obtain

$$\lim \frac{1}{n} \log a_{j,n} = g(d), \qquad 0 < d < 1$$

which proves (6) of Theorem 1. (7) now follows from Lemma 5 in exactly the same way as for the polynomials $\{q_n(x), n = 1, 2, \cdots\}$.

We now only have to remove the condition that $B > 0$. Define the functions f_ε, f_ε^{-1} and g_ε by (8) and (9), where the integration is over $[-A, -\varepsilon]$ $(\varepsilon > 0)$. It is clear (since F has no jump at 0) that $f_\varepsilon^{-1}(x)$ converges to $f^{-1}(x)$ for $x > 0$ and $f^{-1}(x)$ is a positive, continuous and decreasing function that maps $(0, \infty)$ into $(0, 1)$. Its inverse mapping f is then a positive, continuous and increasing function that maps $(0, 1)$ into $(0, \infty)$, and by (9) we conclude that g_ε converges to g, being a continuous function on $(0, 1)$. Define the polynomials

$$p_n^+(x) = \prod_{|x_{j,n}| \geqq \varepsilon} (x - x_{j,n}) = \sum_{j=0}^{n_1} a_{j,n}^+ x^{n_1 - j},$$

$$p_n^-(x) = \prod_{|x_{j,n}| < \varepsilon} (x - x_{j,n}) = \sum_{j=0}^{n-n_1} a_{j,n}^- x^{n - n_1 - j}.$$

The degree n_1 of $p_n^+(x)$ is $N(n; -A, -\varepsilon) = nF(-\varepsilon) + c_n(\varepsilon)$ where $c_n(\varepsilon)/n \to 0$ for every $\varepsilon > 0$. Since $p_n(x)$ is the product of p_n^+ and p_n^- we have the relation

$$a_{j,n} = \sum_{k=0}^{n} a_{j-k,n}^+ a_{k,n}^-$$

so that $(j \leqq n_1)$

$$\begin{aligned} 1 \leqq \frac{a_{j,n}}{a_{j,n}^+} &\leqq \sum_{k=0}^{n} \frac{a_{j-k,n}^+}{a_{j,n}^+} a_{k,n}^- \\ &\leqq \sum_{k=0}^{\infty} \left(\frac{a_{j-1,n}^+}{a_{j,n}^+} \right)^k \frac{1}{k!} \{(n - n_1)\}^k \\ &= \exp \left\{ \frac{a_{j-1,n}^+}{a_{j,n}^+} (n - n_1) \varepsilon \right\} \end{aligned}$$

where we have used Lemma 5 and a straightforward inequality for $a_{k,n}^-$. When we take logarithms and let $n \to \infty$, $j/n \to d \in (0, 1)$ this leads to

$$0 \leqq \limsup \left\{ \frac{1}{n} \log a_{j,n} - \frac{1}{n} \log a_{j,n}^+ \right\} \leqq \varepsilon \{1 - F(-\varepsilon)\} / f_\varepsilon(d).$$

Now if $n\to\infty$ and $j/n\to d\in(0,1)$

$$\begin{aligned}\frac{1}{n}\log a_{j,n}^{+}&=\frac{n_1}{n}\frac{1}{n_1}\log a_{j,n}^{+}\\&\to F(-\varepsilon)g_\varepsilon(d)\end{aligned}$$

so that

$$0\leqq \limsup \frac{1}{n}\log a_{j,n}-F(-\varepsilon)g_\varepsilon(d)\leqq \varepsilon\{1-F(-\varepsilon)\}\frac{1}{f_\varepsilon(d)}.$$

Letting ε tend to zero we obtain the result for the general case.

3. Applications. Let us apply Theorem 1 to some relevant cases.

Example 1 (Degenerate case). Consider the sequence of polynomials given by

$$p_n(x)=(x+a)^n \qquad (a>0).$$

Clearly we have

$$\{p_n(x)\}^{1/n}\to x+a=\exp\left\{\int_{-A}^{0}\log(x-y)\,dF(y)\right\},\qquad x>0 \tag{25}$$

so that the weak limit F is given by

$$\begin{aligned}F(t)&=1,\qquad t\geqq -a,\\&=0,\qquad t<-a.\end{aligned}$$

By (8) we easily find that

$$f^{-1}(x)=\frac{a}{x+a},\qquad x>0$$

from which, by inversion, we obtain

$$f(d)=a\frac{1-d}{d},\qquad 0<d<1. \tag{26}$$

Formula (9) becomes, using (25) and (26),

$$g(d)=d\log a-d\log d-(1-d)\log(1-d).$$

Applying Theorem 1 to this case we obtain

$$\begin{aligned}&\frac{1}{n}\log a_{j,n}\to d\log a-d\log d-(1-d)\log(1-d),\\&\frac{a_{j,n}}{a_{j-1,n}}\to a\frac{1-d}{d}\end{aligned} \tag{27}$$

where $n\to\infty$ and $j/n\to d\in(0,1)$. This result could also be obtained by a direct calculation. We have

$$p_n(x)=\sum_{j=0}^{n}\binom{n}{j}a^jx^{n-j}$$

from which $a_{j,n}=\binom{n}{j}a^j$ follows. Stirling's formula then leads to the result in (27).

Example 2 (Orthogonal polynomials on a compact set). Consider the sequence of monic polynomials $\{p_n(x),\ n=0,1,2,\cdots\}$ that satisfies

$$\int_E p_n(x)p_m(x)\,d\mu(x)=0,\qquad n\neq m \tag{28}$$

where E is a compact set on the negative real axis with positive capacity $C(E)>0$ [16], and μ is a positive measure on E. Let μ_E be the equilibrium measure on E (Frostman measure) [16]. This is the probability measure that minimizes the logarithmic energy

$$\int_E\int_E \log\frac{1}{|x-y|}\,d\mu_E(x)\,d\mu_E(y)=\inf_{\mu\in\Omega_E}\int_E\int_E \log\frac{1}{|x-y|}\,d\mu(x)\,d\mu(y)$$

where Ω_E is the set of all probability measures on E. Then from a result of Widom [19, Thm. 1] (see also Van Assche [18]) we can obtain the following lemma.

LEMMA 6. *Suppose $\mu=\mu_1+\mu_2$, where μ_1 is absolutely continuous with respect to $\mu_E(\mu_1\ll\mu_E)$ and μ_2 is singular with respect to $\mu_E(\mu\perp\mu_E)$. Denote the Radon-Nikodym derivative $d\mu_1/d\mu_E$ by w, then $\mu_E(\{w(x)>0\})=1$ implies the weak convergence of $F_n(t)$ to $F(t)=\mu_E((-\infty,t])$.*

(This follows because μ is an "admissible" measure, in Widom's terminology.)

If E is a compact set on the negative real axis, we then have

$$\frac{1}{n}\log p_n(x)\to\int_E \log(x-y)\,d\mu_E(y)\qquad(x>0)$$

so that Theorem 1 holds with

$$\begin{aligned}&f^{-1}(x)=-\int_E\frac{y}{x-y}\,d\mu_E(y),\qquad x>0,\\ (29)\qquad&\\ &g(d)=-(1-d)\log f(d)+\int_E \log[f(d)-y]\,d\mu_E(y),\qquad 0<d<1.\end{aligned}$$

Let us, for example, take $E=[-1,0]$, then

$$\mu_E(A)=\frac{1}{\pi}\int_A\frac{dt}{\sqrt{t(1+t)}},\qquad A\subset(0,1)$$

and

$$(30)\qquad \int_E \log(x-y)\,d\mu_E(y)=-2\log 2+\log\{2x+1+2\sqrt{x^2+x}\},\qquad x>0.$$

Simple calculations then yield

$$f^{-1}(x)=1-x\int_E\frac{1}{x-y}\,d\mu_E(y)=1-\frac{x}{\sqrt{x^2+x}},\qquad x>0$$

from which

$$(31)\qquad f(d)=\frac{(1-d)^2}{1-(1-d)^2},\qquad 0<d<1.$$

By means of (30) and (31) one then computes

$$g(d)=-2\log 2-d\log d+(2-d)\log(2-d)-2(1-d)\log(1-d),\qquad 0<d<1.$$

We therefore obtain the limits

$$\begin{aligned}&\frac{1}{n}\log a_{j,n}\to-2\log 2-d\log d+(2-d)\log(2-d)-2(1-d)\log(1-d),\\ (32)\qquad&\\ &\frac{a_{j,n}}{a_{j-1,n}}\to\frac{(1-d)^2}{1-(1-d)^2}\end{aligned}$$

when $n \to \infty$ and $j/n \to d \in (0, 1)$. These results could again have been obtained from a direct computation. We may, for example, take the sequence of polynomials $\{P_n^{(\alpha,\beta)}(1+2x),\ n = 0, 1, 2, \cdots\}$, where $P_n^{(\alpha,\beta)}$ is a Jacobi polynomial [15]. This sequence contains orthogonal polynomials with weight function

$$w(x;\,\alpha,\beta) = (-x)^\alpha(1+x)^\beta \qquad (\alpha > -1;\ \beta > -1)$$

and the monic polynomials are given by

$$\hat{P}_n^{(\alpha,\beta)}(1+2x) = \frac{\Gamma(\alpha+n+1)}{n!\,\Gamma(\alpha+\beta+2n+1)} \sum_{j=0}^{n} \binom{n}{j} \frac{\Gamma(\alpha+\beta+n+j+1)}{\Gamma(\alpha+j+1)} x^j$$

so that

$$a_{j,n} = \frac{\Gamma(\alpha+n+1)}{\Gamma(\alpha+\beta+2n+1)} \binom{n}{n-j} \frac{\Gamma(\alpha+\beta+2n-j+1)}{\Gamma(\alpha+\beta+n-j+1)}$$

and applying Stirling's formula leads to (32). In Figs. 1 and 2 we have plotted the functions g and f of (32).

Example 3 (Orthogonal polynomials with exponential weights). Let us take a sequence of monic polynomials $\{p_n(x),\ n = 0, 1, 2, \cdots\}$ for which

$$\int_{-\infty}^{0} p_n(x) p_m(x) w(x)\, dx = 0, \qquad n \neq m$$

where $w(x) > 0$ almost everywhere on $(-\infty, 0)$ and

$$\lim_{x \to -\infty} \frac{\log w(x)}{|x|^\gamma} = -1, \qquad \gamma > 0. \tag{33}$$

Rakhmanov [12] showed that for $\gamma > \frac{1}{2}$

$$\frac{1}{n} \log |k_n^{-n} p_n(k_n x)| \to \int_{-1}^{0} \log |x-y|\, dF^{(\gamma)}(y), \tag{34}$$

uniformly on compact subsets of $\mathbb{C}\backslash(-\infty, 0]$, where

$$k_n = \left(\frac{2n}{A(\gamma)}\right)^{1/\gamma}, \qquad A(\gamma) = \frac{\Gamma(2\gamma)}{2^{2\gamma-1}\{\Gamma(\gamma)\}^2} \tag{35}$$

and the distribution function $F^{(\gamma)}$ is given by

$$\begin{aligned} F^{(\gamma)}(t) &= \int_{-1}^{t} b_\gamma(s)\, ds, \qquad 0 \leqq t \leqq 1, \\ b_\gamma(s) &= \frac{\gamma}{\pi} |s|^{\gamma-1} \int_{|s|}^{1} u^{-\gamma-1/2} \frac{du}{\sqrt{1-u}}, \qquad -1 \leqq s \leqq 0. \end{aligned} \tag{36}$$

Mhaskar and Saff [10] have analysed the weight functions $w(x) = e^{-|x|^\gamma}$ and they obtained the same result for these functions, however, also for $0 < \gamma \leqq \frac{1}{2}$. The function $b_\gamma(s)$ can be written in terms of hypergeometric functions [10], [12], [17] and is sometimes called an Ullman weight (on $[-1, 0]$). Applying the theorem to this sequence yields the following: If

$$p_n(x) = \sum_{j=0}^{n} a_{j,n} x^{n-j},$$

then

$$k_n^{-n} p_n(k_n x) = \sum_{j=0}^{n} a_{j,n} k_n^{-j} x^{n-j}$$

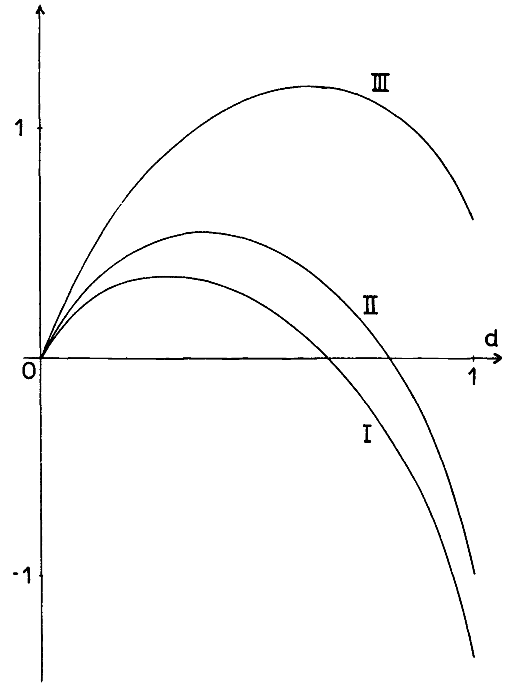


FIG. 1. *The function g for:* I—*Orthogonal polynomials on* $[-1, 0]$; II—*Laguerre polynomials on* $(-\infty, 0]$; III—*Iterations of* z^2+6z+1.

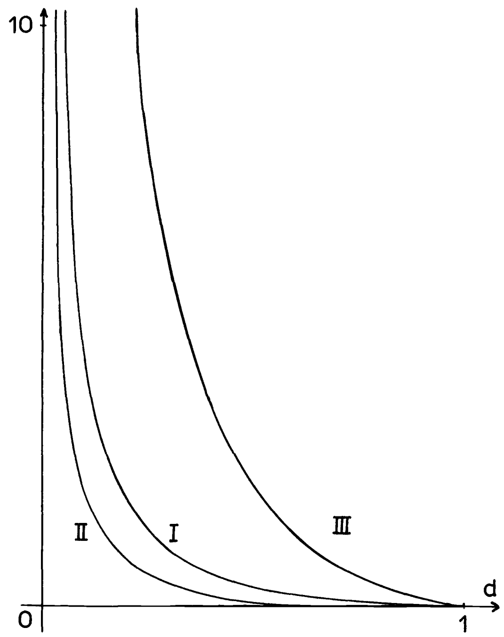


FIG. 2. *The function f for:* I—*Orthogonal polynomials on* $[-1, 0]$; II—*Laguerre polynomials on* $(-\infty, 0]$; III—*Iterations of* z^2+6z+1.

so that

$$
\tag{37}
\begin{aligned}
&\frac{1}{n}\log a_{j,n} - d\log k_n \to g(d),\\
&\frac{1}{k_n}\frac{a_{j,n}}{a_{j-1,n}} \to f(d),
\end{aligned}
\qquad n\to\infty, \qquad \frac{j}{n}\to d\in(0,1)
$$

where k_n is given by (35) and

$$
\tag{38}
\begin{aligned}
&f^{-1}(x) = -\int_{-1}^{0}\frac{y}{x-y}b_\gamma(y)\,dy, \qquad x>0,\\
&g(d) = -(1-d)\log f(d) + \int_{-1}^{0}\log\,[f(d)-y]b_\gamma(y)\,dy, \qquad 0<d<1.
\end{aligned}
$$

Let us take $\gamma=1$, then

$$b_1(s) = \frac{2}{\pi}\sqrt{\frac{1+s}{-s}}, \qquad -1\leqq s\leqq 0$$

and

$$
\begin{aligned}
\int_{-1}^{0} b_1(y)\log\,(x-y)\,dy = 2\log\,[x+\sqrt{x^2+x}] - \log x&\\
+2[-x+\sqrt{x^2+x}] - 2\log 2 - 1,& \qquad x>0.
\end{aligned}
$$

Straightforward calculus leads to

$$f^{-1}(x) = 1 - x\frac{2}{x+\sqrt{x^2+x}} = 2x+1-2\sqrt{x^2+x}$$

and by inversion

$$f(d) = \frac{(1-d)^2}{4d}.$$

The concave function g is then given by

$$g(d) = -2(1-d)\log\,(1-d) - d\log d - 2d\log 2 - d.$$

Moreover (35) becomes $k_n = 4n$, so that we find

$$
\tag{39}
\begin{aligned}
&\frac{1}{n}\log a_{j,n} - d\log n \to -2(1-d)\log\,(1-d) - d\log d - d,\\
&\frac{1}{4n}\frac{a_{j,n}}{a_{j-1,n}} \to \frac{(1-d)^2}{4d}
\end{aligned}
$$

where $n\to\infty$ and $j/n\to d\in(0,1)$. These results for $\gamma=1$ could also have been obtained by analyzing the sequence of polynomials $\{L_n^{(\alpha)}(-x),\ n=0,1,2,\cdots\}$, where $L_n^{(\alpha)}$ is a Laguerre polynomial [15] for which (33) is valid with $\gamma=1$. The Laguerre polynomial is given by

$$L_n^{(\alpha)}(x) = \sum_{j=0}^{n}\binom{n+\alpha}{n-j}\frac{(-x)^j}{j!}$$

and its asymptotic properties are given in [15], [9].

The coefficients of the monic polynomial are given by

$$a_{j,n} = n!\binom{n+\alpha}{j}\frac{1}{(n-j)!},$$

and by means of Stirling's formula one can then easily obtain (39). The limit functions of (39) are plotted in Figs. 1 and 2.

Example 4 (Iterations of polynomials). As a last example we consider a polynomial T of degree $k \geqq 2$:

$$T(z) = z^k + a_{k-1}z^{k-1} + \cdots + a_0$$

and its iterations $\{T^{(n)}(z),\ n = 0, 1, 2, \cdots\}$, where $T^{(0)}(z) = z$ and $T^{(n)}(z) = T^{(n-1)}(T(z))$. The Julia set J for this polynomial T is the set of complex numbers for which the sequence $\{T^{(n)}(z),\ n = 0, 1, 2, \cdots\}$ is not normal (in Montel's sense) [1], [2], [4]. It is known [4] that the set J is compact and has capacity one. We suppose that the set J is real and on the negative real axis. The orthogonal polynomials with respect to the equilibrium measure μ_J on the Julia set are denoted by $\{p_n(x; J),\ n = 0, 1, 2, \cdots\}$ and are related to the iterated polynomial sequence $\{T^{(n)}(z),\ n = 0, 1, 2, \cdots\}$ by ([1], [11])

$$\begin{aligned} p_1(x; J) &= x + \frac{1}{k}a_{k-1}, \\ p_{nk}(x; J) &= p_n(T(x); J), \qquad n = 0, 1, 2, \cdots. \\ p_{k^n}(x; J) &= T^{(n)}(x) + \frac{1}{k}a_{k-1}, \end{aligned}$$

If we write

$$T^{(n)}(x) = \sum_{j=0}^{k^n} t_{j,n}x^{k^n-j}$$

then we may apply (29) to obtain

$$\frac{1}{k^n}\log t_{j,n} \to g(d), \qquad \frac{t_{j,n}}{t_{j-1,n}} \to f(d)$$

where $n \to \infty$ and $j/k^n \to d \in (0, 1)$ and f and g are given by

$$f^{-1}(x) = -\int_J \frac{y}{x-y}\,d\mu_J(y), \qquad x > 0,$$

$$g(d) = -(1-d)\log f(d) + \int_J \log\,[f(d) - y]\,d\mu_J(y), \qquad 0 < d < 1.$$

As an example one may consider the polynomial

$$T(z) = z^2 + 2bz + c. \tag{40}$$

By the Möbius transformation $L(z) = z - b$, we find that

$$\tilde{T}(z) = L^{-1} \circ T \circ L(z) = z^2 - p \qquad (p = b^2 - b - c).$$

It is well known that for $p \geqq 2$ the Julia set $\tilde{J}$ for $\tilde{T}$ is real and contained in $[-q, q]$, $(q = \frac{1}{2} + \sqrt{\frac{1}{4} + p})$ ([4, Thm. 12.1]) and for $p > 2$ the Julia set is a Cantor set. The Julia set for the polynomial T is given by $L(J)$, so that J is contained in $[-q - b, q - b]$ whenever $b^2 - b - c \geqq 2$. If we choose the parameters b and c in (40) such that

$$b \geqq -\tfrac{1}{2}, \quad c \geqq 0, \quad b^2 - b - c \geqq 2$$

then J will be a Cantor set on the negative real axis, contained in $[-b-\frac{1}{2}-\sqrt{\frac{1}{4}+b^2-b-c}, -b+\frac{1}{2}+\sqrt{\frac{1}{4}+b^2-b-c}]$. In Figs. 1 and 2 we have plotted the functions f and g for $T(z)=z^2+6z+1$.

Theorem 1 is not valid for the iterations of the polynomial $T(z)=z^2-p$ $(p\geqq 2)$ since every function $T^{(n)}(x)$ is an even function and the coefficients of odd powers of x vanish. If $\{p_n(x; J),\ n=0,1,2,\cdots\}$ are the monic orthogonal polynomials with respect to the equilibrium measure on the Julia set J of z^2-p $(p\geqq 2)$, then $\{p_n^-(x);\ n=0,1,\cdots\}$ with

$$p_n^-(-x)=(-1)^n p_{2n}(\sqrt{x}; J)$$

will be orthogonal on a set J^+ in $[-q^2, 0]$ $(q=\frac{1}{2}+\sqrt{\frac{1}{4}+p})$. If we put

$$T^{(n)}(x)=\sum_{j=0}^{2^{n-1}}(-1)^j t_{j,n}x^{2^n-2j}$$

then

$$p_{2^{n-1}}^-(x)=T^{(n)}(i\sqrt{x})=\sum_{j=0}^{2^{n-1}} t_{j,n}x^{2^{n-1}-j}$$

and since Theorem 1 is valid for $\{p_n^-(x);\ n=1,2,\cdots\}$ we have

$$\frac{1}{2^{n-1}}\log t_{j,n}\to g(d),\qquad \frac{t_{j,n}}{t_{j-1,n}}\to f(d)$$

whenever $n\to\infty$ and $j2^{-n+1}\to d\in(0,1)$, with

$$f^{-1}(x)=-\int_{J^+}\frac{y}{x-y}\,d\mu_{J^+}(y),\qquad x>0,$$

$$g(d)=-(1-d)\log f(d)+\int_{J^+}\log[f(d)-y]\,d\mu_{J^+}(y),\qquad 0<d<1.$$

The case where $p=2$ gives essentially the Chebyshev polynomials of the first kind on $[-2,2]$ and, up to some constant, the formula (32) holds.

Acknowledgments. G. Fano is indebted to Dr. M. F. Barnsley for stimulating discussions and to Dr. S. Turrini for his help in the use of the algebraic SCHOONSCHIP language, which allowed numerical computations on the iterated maps.

REFERENCES

[1] M. F. Barnsley, J. S. Geronimo and A. N. Harrington, *Orthogonal polynomials associated with invariant measures on Julia sets*, Bull. Amer. Math. Soc., 7 (1982), pp. 381-384.

[2] D. Bessis and P. Moussa, *Orthogonality properties of iterated polynomial mappings*, Comm. Math. Phys., 88 (1983), pp. 503-529.

[3] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[4] H. Brolin, *Invariant sets under iteration of rational functions*, Ark. Mat., 6 (1965), pp. 103-144.

[5] G. Fano and G. Gallavotti, *Dense sums*, Ann. Inst. H. Poincaré Sect. A, 17 (1972), pp. 195-219.

[6] G. Fano and G. Loupias, *On the thermodynamic limit of the B.C.S. state*, Comm. Math. Phys., 20 (1971), pp. 143-166.

[7] M. E. Fisher, *The free energy of a macroscopic system*, Arch. Rational. Mech. Anal., 17 (1964), pp. 377-410.

[8] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press, London, 1952.

[9] M. Maejima and W. Van Assche, *Probabilistic proofs of asymptotic formulas for some classical polynomials*, Math. Proc. Cambridge Philos. Soc., 97 (1985), pp. 499-510.

[10] H. N. MHASKAR AND E. B. SAFF, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203-234.

[11] T. S. PITCHER AND J. R. KINNEY, *Some connections between ergodic theory and the iteration of polynomials*, Ark. Mat., 8 (1968), pp. 25-32.

[12] E. A. RAKHMANOV, *On asymptotic properties of polynomials orthogonal on the real axis*, Mat. Sb., 119 (161) (1982), pp. 163-203. (In Russian.); Math. USSR-Sb., 47 (1984), pp. 155-193. (In English.).

[13] A. W. ROBERTS AND D. E. VARBERG, *Convex Functions*, Academic Press, New York, 1973.

[14] W. RUDIN, *Principles of Mathematical Analysis*, McGraw-Hill, Tokyo, 1964.

[15] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Pub. 23, 4th ed., Providence, RI, 1975.

[16] M. TSUJI, *Potential Theory in Modern Function Theory*, Maruzen, Tokyo, 1959.

[17] J. L. ULLMAN, *Orthogonal polynomials associated with an infinite interval*, Michigan Math. J., 27 (1980), pp. 353-363.

[18] W. VAN ASSCHE, *Invariant zero behaviour for orthogonal polynomials on compact sets of the real line*, Bull. Soc. Math. Belg. Sér B, 38 (1986).

[19] H. WIDOM, *Polynomials associated with measures in the complex plane*, J. Math. & Mech., 16 (1967), pp. 997-1013.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

009

ON A SIMPLIFIED ASYMPTOTIC FORMULA FOR THE MATHIEU FUNCTION OF THE THIRD KIND*

D. NAYLOR†

Abstract. This paper considers the asymptotic form of solutions of the equation $y_{xx}=(u^2-2h^2\cosh 2x)y$ for fixed real values of x and h and large complex values of u. Attention is focused on that solution known as the Mathieu function of the third kind, $M_\nu^{(3)}(x)$, and for values of u in the half plane Re $(u)\geqq 0$. The basic asymptotic formulas require the determination of an elliptic integral but, when u is large, it is shown how this integral can be approximated by elementary functions.

Key words. Mathieu functions, asymptotic formulas

AMS(MOS) subject classifications. Primary 33A55, 41A60

1. Introduction. The solution of wave propagation problems involving infinite domains bounded internally by cylinders of elliptic cross section requires the consideration of the functions $M_\nu^{(j)}(x)\, j=1,2,3,4$, which satisfy the modified form of Mathieu's differential equation:

$$w_{xx}=(u^2-2h^2\cosh 2x)w, \qquad a\leqq x<\infty. \tag{1.1}$$

The constants h, a are supposed real and positive but the parameter u may take both real and complex values. The quantity ν (the characteristic exponent) used in the standard notation of the Mathieu functions is connected with the parameter u by a complicated relation which for large values of u can be approximated by the following equation [2, p. 125]:

$$u^2=\nu^2+O\left(\frac{h^2}{\nu^2}\right). \tag{1.2}$$

In this paper attention is focussed on the solution of (1) that is associated with outgoing waves. This solution is the function $M_\nu^{(3)}(x)$ which behaves [2, p. 170] for fixed ν and large x according to the formula

$$M_\nu^{(3)}(x)=H_\nu^{(1)}(2h\cosh x)[1+O(\operatorname{sech} x)] \tag{1.3}$$

as $x\to+\infty$, where $H_\nu^{(1)}$ denotes the Hankel function of the first kind. The object is to obtain an asymptotic formula describing the behaviour of $M_\nu^{(3)}(x)$ for large values of u in the half plane Re $(u)\geqq 0$. It is clear that, if x is confined to some bounded interval of values, solutions $w^{(1)}$ and $w^{(2)}$ of (1.1) exist possessing the asymptotic forms

$$w^{(1)}\sim e^{ux}, \qquad w^{(2)}\sim e^{-ux}$$

as $u\to\infty$. Since any solution of (1.1) is expressible as a linear combination of $w^{(1)}$ and $w^{(2)}$ it follows that

$$M_\nu^{(3)}(x)\sim c_1\,e^{ux}+c_2\,e^{-ux} \tag{1.4}$$

where the coefficients c_1, c_2 may depend on u but not on x. The precise form of the relation (1.4), which will be determined by following a procedure developed by Olver [4], is given by the following equation:

$$M_\nu^{(3)}(x)=\sqrt{\frac{2}{\pi u}}\,e^{i(u-\nu)(\pi/2)}[e^{ux-u\log(2u/he)}-ie^{-ux+u\log(2u/he)}]\left[1+O\left(\frac{1}{u}\right)\right]. \tag{1.5}$$

* Received by the editors June 23, 1986; accepted for publication (in revised form) December 11, 1986.

† Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada N6A 5B9.

Olver's method requires the consideration of basic solutions of (1.1) for complex values of the independent variable and matching combinations of these solutions in different domains. We therefore replace x by the complex variable z and consider the differential equation

$$w_{zz}=(u^2-2h^2\cosh 2z)w. \tag{1.6}$$

A standard discussion of the asymptotic form of solutions of this equation when both z and u are large may be affected by means of a Liouville transformation. The asymptotic expressions for the solutions of (1.6) obtained by this method involve a variable $\xi(z, u)$ defined by an integral of the form

$$\xi(z, u)=\int_c^z (u^2-2h^2\cosh 2t)^{1/2}\,dt \tag{1.7}$$

in which c is independent of z. Since (1.7) is an elliptic integral it is desirable to obtain an asymptotic expression for it in terms of simpler functions. This problem was considered, for real values of z, in [3] where it was shown how to approximate the value of such an integral for large values of u by means of elementary functions. In the next section of this paper it is shown how similar asymptotic expressions may be obtained for the value of the relevant elliptic integral when the variable z is complex. The actual construction of the formula (1.5) is carried out in § 3.

2. Asymptotic formulas for $\zeta(z, u)$. In this section asymptotic formulas are constructed for the function $\zeta(z, u)$ defined by the elliptic integral

$$\zeta(z, u)=\int_{z_0}^z (u^2-2h^2\cosh 2t)^{1/2}\,dt \tag{2.1}$$

where $z_0=x_0+iy_0$ is that solution of the equation $2h^2\cosh 2z_0=u^2$ that is given for large values of u by the approximate formula $z_0\sim\log(u/h)$ the principal value of the logarithm being taken. Because attention is confined to values of u such that $|\arg u|\leqq\pi/2$, the point z_0 is located in the strip $|\operatorname{Im} z_0|\leqq\pi/2$. Since the modified Mathieu functions are periodic in z we could consider the effect of the mapping (2.1) on the strip $0\leqq\operatorname{Im} z\leqq\pi$ but in applying Olver's technique it is helpful to consider the larger strip $|\operatorname{Im} z|\leqq\pi$. Branch cuts parallel to the real axis are introduced from z_0 to $z_0+\infty$ and from $-z_0$ to $-z_0-\infty$ and the branch of the radical chosen is that which is asymptotically equal to ihe^z as $z\to\infty$ on the lower side of the first such cut. With this choice the function ζ is also asymptotically equal to ihe^z as $z\to\infty$ in the stated manner.

In order to obtain approximate formulas for ζ in terms of elementary functions it will be necessary to consider separately three different parts of the strip $|\operatorname{Im} z|\leqq\pi$, as defined by the inequalities:

(i) $\operatorname{Re}(z)\geqq\frac{1}{2}x_0$,

(ii) $|\operatorname{Re}(z)|\leqq\frac{1}{2}x_0$, and

(iii) $\operatorname{Re}(z)\leqq-\frac{1}{2}x_0$.

Case (i). To extract the dominant contribution to the integral in (2.1) when u is large, the integrand is decomposed as the sum of four parts as specified by the following equation:

$$\begin{aligned}(u^2-2h^2\cosh 2t)^{1/2}=&-\frac{2h^2\sinh 2t}{(u^2-2h^2\cosh 2t)^{1/2}}+\frac{u^2\tanh 2t}{(u^2-2h^2\cosh 2t)^{1/2}}\\&-\frac{2h^2e^{-2t}}{(u^2-2h^2\cosh 2t)^{1/2}}+\frac{u^2e^{-2t}}{\cosh 2t(u^2-2h^2\cosh 2t)^{1/2}}.\end{aligned}$$

Upon integrating the terms on the right-hand side of the preceding equation along a path in the complex t-plane connecting the points z_0 and z we find that

$$\begin{aligned}\zeta(z, u) = -2h^2 &\int_{z_0}^{z} \frac{\sinh 2t\,dt}{(u^2-2h^2\cosh 2t)^{1/2}} \\ &+ u^2 \int_{z_0}^{z} \frac{\sinh 2t\,dt}{\cosh 2t(u^2-2h^2\cosh 2t)^{1/2}} - 2h^2 I_1 + u^2 I_2\end{aligned} \tag{2.2}$$

where

$$I_1 = \int_{z_0}^{z} \frac{e^{-2t}\,dt}{(u^2-2h^2\cosh 2t)^{1/2}}, \qquad I_2 = \int_{z_0}^{z} \frac{e^{-2t}\,dt}{\cosh 2t(u^2-2h^2\cosh 2t)^{1/2}}. \tag{2.3}$$

It will be shown shortly that $I_1 = O(u^{-2})$ and $I_2 = O(u^{-3})$. Since the first two integrals present on the right-hand side of (2.2) may be evaluated explicitly, we find the equation

$$\begin{aligned}\zeta(z, u) = (u^2-2h^2\cosh 2z)^{1/2} &- u \log[u + (u^2-2h^2\cosh 2z)^{1/2}] \\ &+ \frac{u}{2}\log(2h^2\cosh 2z) + O(u^{-1})\end{aligned} \tag{2.4}$$

which applies for sufficiently large u and $\operatorname{Re}(z) \geqq \frac{1}{2}\log|u/h|$.

To obtain the stated bounds on I_1 and I_2 the path of integration connecting z_0 and z is taken to consist of the straight line from $z_0 = x_0 + iy_0$ to $x_0 + iy$ together with the straight line from $x_0 + iy$ to $z = x + iy$. On writing $t = \tau + i\eta$ we find, since $u^2 = 2h^2\cosh 2z_0$, that

$$\begin{aligned}|u^2 - 2h^2\cosh 2t| &= 2h^2|\cosh 2z_0 - \cosh 2t| \\ &= 4h^2|\sinh(z_0+t)\sinh(z_0-t)| \\ &= 4h^2[\sinh^2(x_0+\tau) + \sin^2(y_0+\eta)]^{1/2} \\ &\quad\cdot[\sinh^2(x_0-\tau)+\sin^2(y_0-\eta)]^{1/2} \\ &\geqq 2\sqrt{2}\,h^2\sinh(x_0+\tau)[\sinh|x_0-\tau| + |\sin(y_0-\eta)|]\end{aligned} \tag{2.5}$$

after using the inequality $(a^2+b^2)^{1/2} \geqq (a+b)/\sqrt{2}$ which holds for any two positive numbers a and b. On the line joining z_0 and x_0+iy we have $t = x_0 + i\eta$ so that, by (2.5), the contribution of this part of the path of integration to the value of I_1 does not exceed the quantity

$$\frac{Ce^{-2x_0}}{(\sinh 2x_0)^{1/2}}\left|\int_{y_0}^{y}\frac{d\eta}{|\sin(y_0-\eta)|^{1/2}}\right| \tag{2.6}$$

where C is a constant. This expression is $O(e^{-3x_0}) = O(u^{-3})$, since $x_0 \sim \log|u/h|$, the integral in (2.6) being bounded independently of u because $|y| < \pi$ and $0 < y_0 < \pi/2$. On the part of the path of integration that connects z with (x_0+iy) we use the following simplified form of the inequality (2.5):

$$|u^2-2h^2\cosh 2t| \geqq 2\sqrt{2}\,h^2\sinh(x_0+\tau)\sinh|x_0-\tau|.$$

The contribution to I_1 of this part of the path is bounded by the quantity

$$C\left|\int_{x}^{x_0}\frac{e^{-2\tau}\,d\tau}{[\sinh(x_0+\tau)\sinh|x_0-\tau|]^{1/2}}\right|. \tag{2.7}$$

If $x_0 \leqq x < \infty$ this expression is less than the integral

$$\frac{C e^{-2x_0}}{(\sinh 2x_0)^{1/2}} \int_{x_0}^{\infty} \frac{d\tau}{[\sinh(\tau - x_0)]^{1/2}}.$$

This expression is $O(e^{-3x_0})$ which is again $O(u^{-3})$. If $x_0/2 \leqq x \leqq x_0$ we note that $\sinh(x_0 + \tau) \geqq \sinh(3x_0/2)$ so that the expression (2.7) does not exceed the quantity

$$(2.8)\quad \begin{aligned} &\frac{C}{[\sinh(3x_0/2)]^{1/2}} \int_x^{x_0} \frac{e^{-2\tau}\, d\tau}{[\sinh(x_0 - \tau)]^{1/2}} = \frac{C e^{-2x_0}}{[\sinh(3x_0/2)]^{1/2}} \int_0^{x_0 - x} \frac{e^{2\xi}\, d\xi}{(\sinh \xi)^{1/2}} \\ &\leqq \frac{2Ce^{-x_0-x}}{[\sinh(3x_0/2)]^{1/2}} \int_0^{x_0-x} \frac{\cosh \xi\, d\xi}{(\sinh \xi)^{1/2}} = \frac{4C e^{-x_0-x}[\sinh(x_0 - x)]^{1/2}}{[\sinh(3x_0/2)]^{1/2}} \end{aligned}$$

after setting $\xi = x_0 - \tau$ and using the fact that $e^{2\xi} \leqq 2e^{\xi} \cosh \xi \leqq 2e^{x_0 - x} \cosh \xi$. Since $x \geqq x_0/2$ the final expression in (2.8) is $O(e^{-2x_0})$ which is $O(u^{-2})$. On combining the above bounds we can see that $I_1 = O(u^{-2})$, as stated.

The treatment of I_2 is similar, except that since $|\cosh 2(\tau + i\eta)| \geqq \sinh 2\tau \geqq \sinh x_0$ on the entire path of integration, the extra factor $\cosh 2t$ present in the integrand of I_2 gives rise to an additional factor of $\operatorname{cosech} x_0 = O(e^{-x_0}) = O(u^{-1})$ in the final result so that $I_2 = O(u^{-3})$. This establishes the validity of (2.4).

If e^z is large compared with u^2, the formula (2.4) can be simplified by expanding the terms on the right-hand side of this equation in powers of $u^2 \operatorname{sech}^2 z$ and this leads to the formula

$$(2.9)\quad \zeta(z, u) = ihe^z - \frac{iu\pi}{2} + O(u^2 e^{-z}) + O(u^{-1}).$$

Case (ii). The bounds on I_1 and I_2 stated in the preceding section do not apply when $\operatorname{Re}(z)$ is large and negative or when $|\operatorname{Re}(z)| < \frac{1}{2}x_0$. To obtain an approximate formula for $\zeta(z, u)$ in the latter case we introduce the quantity $z_1 = \frac{1}{2}\log(u/h)$, whose real part is intermediate between 0 and x_0, and write

$$\zeta(z, u) = \int_{z_0}^{z_1} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt + \int_{z_1}^{z} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt.$$

The first integral appearing on the right-hand side of the preceding equation equals $\zeta(z_1, u)$ and can be estimated for large values of u by applying the formula (2.4) just proved. On setting $z = z_1$ in (2.4) and simplifying the resulting expression, we find that

$$(2.10)\quad \zeta(z, u) = u - \frac{h}{4} - \frac{u}{2} \log\left(\frac{4u}{h}\right) + \int_{z_1}^{z} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt + O(u^{-1}).$$

The integral remaining in this equation may be estimated by expanding the integrand in powers of $u^{-2} \cosh 2t$ and integrating term by term, since $|u^{-2} \cosh 2t| \leqq |u^{-2} \cosh 2\tau| \leqq |u^{-2} \cosh 2x_1| = O(u^{-1})$ therein. This leads to the formula

$$(2.11)\quad \int_{z_1}^{z} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt = uz - \frac{h^2}{2u} \sinh 2z - \frac{u}{2} \log\left(\frac{u}{h}\right) + \frac{h}{4} + O(u^{-1})$$

which holds whenever u is large and $|\operatorname{Re}(z)| \leqq \operatorname{Re}(z_1)$, where $z_1 = \frac{1}{2}\log(u/h)$. Upon inserting the result (2.11) into (2.10), we obtain the formula

$$(2.12)\quad \zeta(z, u) = uz - \frac{h^2}{2u} \sinh 2z - u \log(2u/he) + O(u^{-1})$$

for $|\mathrm{Re}(z)| \leqq \mathrm{Re}(z_1)$. This formula applies in particular for $z = a$ when it reduces to the following formula, which we record for future use,

$$\zeta(a, u) = -u \log\left(\frac{2u}{he^{a+1}}\right) + O(u^{-1}). \tag{2.13}$$

Case (iii). The formula (2.4) ceases to apply in the domain $\mathrm{Re}(z) < -\frac{1}{2}x_0$ where however $\zeta(z, u)$ may be calculated from the equation:

$$\zeta(z, u) = -g_0(u) - \zeta(-z, u) \tag{2.14}$$

where

$$g_0(u) = 2u \log\left(\frac{2u}{he}\right) + O(u^{-1}). \tag{2.15}$$

If $\mathrm{Re}(z) < -\frac{1}{2}x_0$ then $\mathrm{Re}\,(-z) > \frac{1}{2}x_0$ and the second term on the right-hand side of (2.14) may be estimated from the formula (2.4) already established. Therefore, for such values of z, equation (2.14) can be employed to determine $\zeta(z, u)$. To verify (2.14) we note from the definition (2.1) that

$$\zeta(z, u) + \zeta(-z, u) = \int_{z_0}^{z} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt + \int_{z_0}^{-z} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt.$$

On reversing the sign of the integration variable in the second integral it becomes one along a path connecting z to $-z_0$ and on combining the two integrals we find the formula

$$\begin{aligned}\zeta(z, u) + \zeta(-z, u) &= \int_{z_0}^{-z_0} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt \\ &= -2\int_{0}^{z_0} (u^2 - 2h^2 \cosh 2t)^{1/2}\, dt = 2\zeta(0, u).\end{aligned}$$

On calculating $\zeta(0, u)$ from (2.12) by inserting the value $z = 0$ therein we deduce the relation (2.14) stated above.

3. Determination of the asymptotic form of $M_\nu^{(3)}(x)$. The determination of the coefficients c_1, c_2 in the relation (1.4) is regarded as a connection formula problem which will be solved by applying the method used by Olver [4]. This method requires a consideration of the mapping of the z-plane onto the ζ-plane defined by the equation (2.1). Properties of this mapping have been investigated by Sharples [6]. For real values of u, Sharples reduces (2.1) to a Schwarz-Christoffel transformation by means of a change of the variable of integration, but this approach is not effective if u is complex. Alternatively, when u is large, a discussion of the mapping may be carried out with the aid of equations (2.4) and (2.12) and the intermediate variable $\tau = u^{-1}w^{1/2}$ where $w = 2h^2 \cosh 2z$. In terms of this variable, equation (2.4) reduces to

$$u^{-1}\zeta = (1 - \tau^2)^{1/2} - \log\left[\frac{1 + (1 - \tau^2)^{1/2}}{\tau}\right] + O(u^{-2}). \tag{3.1}$$

This decomposition of the mapping is applicable to large values of u, both real and complex, but like Sharples, it is helpful to consider real values of u first. Properties of the mapping of the τ-plane onto the plane of the function of τ appearing on the right-hand side of (3.1) are considered on p. 420 of [5]. Cuts are placed in the z-plane along the segments (x_0, ∞) and $(-\infty, -x_0)$ of the real axis, where x_0 denotes the positive root of the equation $2h^2 \cosh 2x_0 = u^2$. In Figs. 1 and 2 corresponding points are

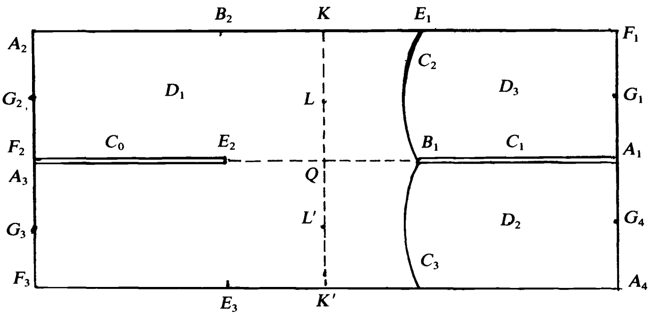


FIG. 1. *z-plane*, $\theta = 0$.

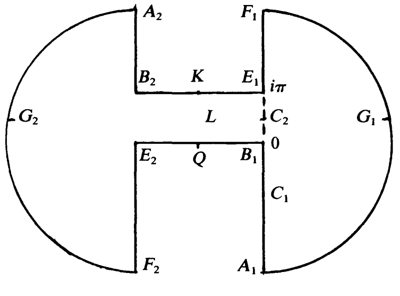


FIG. 2(a). $u^{-1}\zeta$-*plane*, $\theta = 0$.

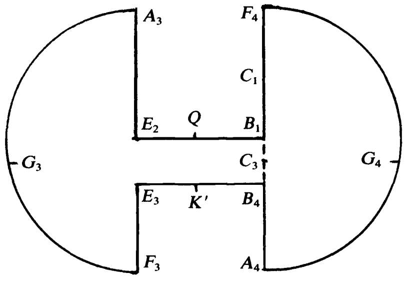


FIG. 2(b). $u^{-1}\zeta$-*plane*, $\theta = 0$.

indicated by the same letters. The distant points G_1, F_1 of the z-plane have ordinates $\pi/2$, π, respectively. The right-hand side of Fig. 2(a) illustrates the $u^{-1}\zeta$-map of the strip $\operatorname{Re}(z) > 0$, $0 < \operatorname{Im}(z) < \pi$. The point B_1 denotes $\zeta = 0$ and the point E_1 is situated at $\zeta = iu\pi$. The left-hand side of Fig. 2(a) is constructed from the right-hand side of the same figure by appealing to formula (2.14) together with the reflection principle, which, since $u^{-1}\zeta$ is real when z is real and between the points $\pm x_0$, implies that $u^{-1}\zeta(x - iy)$ is the conjugate of $u^{-1}\zeta(x + iy)$. On setting $z = -x + iy$ in (2.14) we find, for positive values of u, the necessary formula

$$\begin{aligned} u^{-1}\zeta(-x+iy) &= -u^{-1}\zeta(x-iy) - u^{-1}g_0(u) \\ &= -u^{-1}\overline{\zeta(x+iy)} - u^{-1}g_0(u). \end{aligned}$$

This equation relates the value of ζ at the point $x + iy$ with its value at the image point $-x + iy$ so that, for instance, $\zeta(A_2) = -\bar{\zeta}(F_1) - g_0(u)$, and similarly for the values of ζ at other points on the left-hand side of Fig. 1. In this way the left-hand side of Fig. 2(a) is obtained. The points K, L, Q, L', K' lie outside the domain of validity of (2.4) and their ζ-values are given instead by (2.12) which after division by u reduces to the formula

$$u^{-1}\zeta = z - \log\left(\frac{2u}{he}\right) + O(u^{-1}). \tag{3.2}$$

The strip $-\pi < y < \pi$ is divided into domains D_1, D_2, D_3 by the curves C_1, C_2, C_3 which emanate from the point B_1 and along each of which $\operatorname{Re}\zeta = 0$. The curve C_1 coincides with the segment (x_0, ∞) of the real axis and corresponds to the part B_1A_1 of the imaginary axis of the $u^{-1}\zeta$-plane, whilst C_2 corresponds to the segment B_1E_1 of that axis. Figure 2(a) gives the $u^{-1}\zeta$-map of the strip $0 < \operatorname{Im}(z) < \pi$, for real values of u. For such values of u the ζ-map of the adjoining strip $-\pi < \operatorname{Im}(z) < 0$, Fig. 2(b), is readily constructed from Fig. 2(a) by means of the reflection principle.

If u is complex we set $u = Re^{i\theta}$ where $0 \leqq \theta \leqq \pi/2$ and regard the points A_i, G_i, F_i, for $i = 1, 4$ as being fixed in the $u^{-1}\zeta$-plane, that is they correspond to values of τ independent of u as shown in Figs. 3(a) and 3(b).

Since $u^2\tau^2 = 2h^2 \cosh 2z \sim h^2e^{2z}$ the corresponding (distant) points A_1, G_1, F_1 of the z-plane will now have ordinates θ, $\theta + \pi/2$, $\theta + \pi$, respectively. The remote points A_2, G_2, F_2 on the left-hand side of the z-plane will be displaced a distance θ in the downwards direction. The points Q, L, L', K, K' are taken to be fixed in the z-plane, and their associated ζ-values are given by (3.2), which for complex values of u, can be written in the form

$$u^{-1}\zeta = z - \log\left(\frac{2R}{he}\right) - i\theta + O(R^{-1}).$$

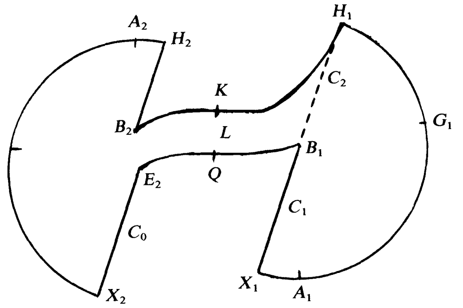


FIG. 3(a). $u^{-1}\zeta$-plane, $0 < \theta < \pi/2$.

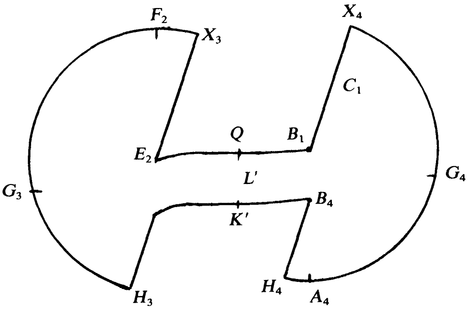


FIG. 3(b). $u^{-1}\zeta$-plane, $0 < \theta < \pi/2$.

This equation shows that the points Q, L, L', K, K' of the $u^{-1}\zeta$-plane move downwards through a distance equal to θ. The turning point B_1 is given for large values of u by the equation $z_0 \sim \log(u/h)$ and the curves C_1, C_2, C_3 through B_1 correspond to rays for which $\arg(u^{-1}\zeta) = \pm\pi/2 - \theta$ in the $u^{-1}\zeta$-plane. The images of these curves in the z-plane are depicted in Fig. 4. For complex values of u the cut emanating from B_1 is taken to coincide with the curve C_1 and that emanating from E_1 is taken to be the curve C_0 congruent to C_1 and orientated as shown in Fig. 4. In constructing the curves C_1, C_2, C_3 appearing in Figs. 1, 4 and 5 it is helpful to set $\tau = \operatorname{sech}\sigma$, $\sigma = \alpha + i\beta$, $u = \mathrm{R}e^{i\theta}$ in the formula (3.1) which then yields the equation

$$(3.3) \quad R^{-1}\operatorname{Re}\zeta = \left(\frac{\sinh 2\alpha}{\cosh 2\alpha + \cos 2\beta} - \alpha\right)\cos\theta - \left(\frac{\sin 2\beta}{\cosh 2\alpha + \cos 2\beta} - \beta\right)\sin\theta.$$

Since $2h^2\cosh 2z = u^2\tau^2 = u^2\operatorname{sech}^2\sigma$ we also find, for large positive values of x, the relations

$$(3.4) \qquad 1 + \cosh 2\alpha\cos 2\beta = e^{2(x_0 - x)}\cos 2(\theta - y),$$

$$(3.5) \qquad \sinh 2\alpha \sin 2\beta = e^{2(x_0 - x)}\sin 2(\theta - y),$$

$$(3.6) \qquad \cosh 2\alpha\cos 2\beta = e^{2(x_0 - x)}$$

where $x_0 \sim \log(R/h)$. The curves C_1, C_2, C_3 leave z_0 in directions making angles equal to $-\frac{2}{3}\theta$, $\frac{2}{3}(\pi - \theta)$ and $\frac{2}{3}(2\pi - \theta)$ with the positive real axis, respectively. When $\theta = 0$ the equation (3.3) shows that the curves along which $\operatorname{Re}\zeta$ vanishes are given by either $\alpha = 0$ or $\cosh 2\alpha + \cos 2\beta = \alpha^{-1}\sinh 2\alpha$. The equation $\alpha = 0$ gives the curve C_1 which coincides with the part of the x-axis to the right of the branch point B_1(Fig. 1). The other condition, when used in conjunction with (3.4) and (3.5) to eliminate the variable β, leads to the following equations:

$$\cos 2y = \cosh 2\alpha - \alpha\sinh 2\alpha, \qquad e^{2(x_0 - x)} = \alpha^{-1}\sinh 2\alpha.$$

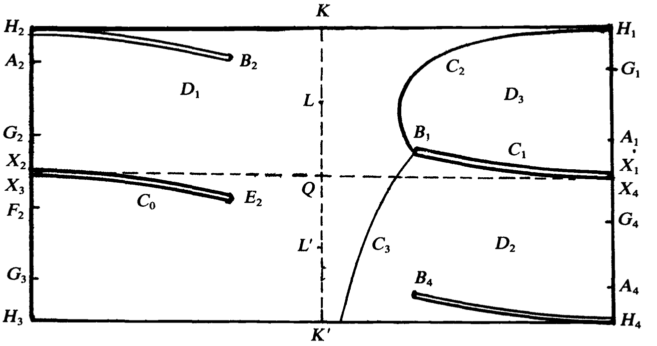


FIG. 4. *z-plane*, $0 < \theta < \pi/2$.

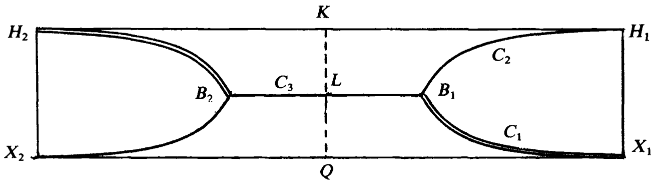


FIG. 5. *z-plane*, $\theta = \pi/2$.

These are the parametric equations of the two remaining curves C_2 and C_3 (for the value $\theta=0$), which are shown along with C_1 in Fig. 1. The corresponding curves associated with the value $\theta=\pi/2$ are shown in Fig. 5. In this case it follows from (3.3) that $\operatorname{Re}\zeta$ vanishes when either $\beta=0$ or $\cosh 2\alpha+\cos 2\beta=\beta^{-1}\sin 2\beta$. The curve C_3 now comprises the line joining the points $\pm x_0+i\pi/2$ and the condition $\beta=0$ gives the part of this line for which (3.3) holds, viz: $x\geqq\frac{1}{2}x_0$. The alternative condition $\cosh 2\alpha+\cos 2\beta=\beta^{-1}\sin 2\beta$, when combined with (3.4) and (3.6) to eliminate the variable α, yields the parametric equations of the curves C_2 and C_1 in the following form:

$$\cos 2y=-\cos 2\beta-\beta\sin 2\beta,\qquad e^{2(x_0-x)}=\beta^{-1}\sin 2\beta.$$

It should be noted that, for large values of R, the curves C_1, C_2, C_3 originating at z_0 can only extend into the left-hand side of the fundamental strip of the z-plane if the angle θ is sufficiently close to $\pm\pi/2$, so that $\cos\theta=O[(\log R)^{-1}]$. To verify this result we note that to reach the left-hand region such a curve must traverse the domain (ii) of § 2 in which the formula (2.12) applies. Upon setting $z=x+iy$, $u=Re^{i\theta}$ in this formula we find that

$$(3.7)\qquad R^{-1}\operatorname{Re}\zeta=\left[x-\log\left(\frac{2R}{he}\right)\right]\cos\theta-(y-\theta)\sin\theta+O(R^{-1}).$$

This formula applies for $|x|\leqq\frac{1}{2}\log(R/h)$. Since y, θ are bounded it follows that a $\operatorname{Re}\zeta=0$ curve cannot enter or traverse this domain unless $\cos\theta=O[(\log R)^{-1}]$ as stipulated.

If $\theta\neq 0$, $\pi/2$ explicit parametric equations for C_1, C_2, C_3 in terms of one or other of the variables α, β are not obtainable, for then the condition $\operatorname{Re}\zeta=0$ retains the more intricate form

$$(\alpha\cos\theta-\beta\sin\theta)(\cosh 2\alpha+\cos 2\beta)=\sinh 2\alpha\cos\theta-\sin 2\beta\sin\theta.$$

In this case the curves may best be sketched by following the variation of the slope dy/dx as the point z, starting at z_0, moves along such a curve. For this purpose appeal is made to the differential equation of the curves. If we write $\tan\phi=dy/dx$ and use the equation (2.1), it can be shown that this differential equation can be written in the following form:

$$\cot 2(\theta+\phi)=[\cos 2(\theta-y)-e^{2(x_0-x)}]\operatorname{cosec} 2(\theta-y).$$

The connection formula, in the form stated by Olver, will now be outlined. This formula relates three solutions w_1, w_2, w_3 of the differential equation (1.5). Let D be the domain $-\pi<\operatorname{Im} z<\pi$, then the solution w_j is defined in D cut along the corresponding curve C_j and along the branch line C_0 through $-z_0$ and possesses the property such that

$$(3.8)\qquad w_j(z)=\phi_j(z)e^{-\zeta_j(z)}[1+\varepsilon_j(z)]$$

where $\zeta_j(z)$ denotes the branch of $\zeta(z)$ that is continuous in D cut along C_0 and C_j and is such that $\operatorname{Re}\zeta_j>0$ in D_j whilst $\operatorname{Re}\zeta_j\leqq 0$ elsewhere. The function $\phi_3(z)$ may be defined to be that branch of $(u^2-2h^2\cosh 2z)^{-1/4}$ that is continuous in D cut along C_1 and C_0, except at $\pm z_0$, and which is asymptotically equal to $h^{-1/2}\exp(-x/2+i\pi/4)$

as $x \to +\infty$. The remaining functions ϕ_1 and ϕ_2 are defined by the equations $\phi_j = e^{i\pi/6}\phi_{j-1}$ for $z \in D_j \cup D_{j-1}$. The quantity $\varepsilon_j(z)$ appearing in (3.8) is subject to the bound

$$(3.9) \qquad |e_j(z)| \leqq \exp\left\{\left|\int_{a_j}^{z} |R(z)| \cdot |dz|\right|\right\} - 1$$

where

$$(3.10) \quad R(z) = 2h^2 \cosh 2z(u^2 - 2h^2 \cosh 2z)^{-3/2} + 5h^4 \sinh^2 2z(u^2 - 2h^2 \cosh 2z)^{-5/2}.$$

The integral appearing in (3.9) is taken along a path connecting z and an arbitrary reference point a_j, the path being one along which Re $\zeta_j(z)$ is nondecreasing and tends to $+\infty$ as $z \to a_j$. It follows that

$$(3.11) \qquad |\varepsilon_j(a_i)| \leqq \exp\left|\int_{a_i}^{a_j} |R(z)| \cdot |dz|\right| - 1$$

where the integral is taken along a path connecting a_i and a_j along which Re ζ_j is monotone. It is assumed that such a path exists for each pair of reference points. The connection formula is then

$$(3.12) \qquad -[1+\varepsilon_2(a_3)]w_1 + [1+\varepsilon_1(a_3)]e^{i\pi/3}w_2 + [1+\varepsilon_1(a_2)]e^{-i\pi/3}w_3 = 0.$$

To fix the solutions w_1, w_2, w_3 we let $a_1 = -\infty + i\pi/2$, $a_2 = \infty - i\pi/2$ and $a_3 = \infty + i\pi/2$. The resulting domains of definitions of w_1, w_2, w_3 are adequate for the intended applications. The function $\zeta_1(z)$ is the branch of $\zeta(z)$ that has positive real part in D_1 and Re $\zeta_1 \to \infty$ as Re $(z) \to -\infty$ in D_1 whilst Re $\zeta_1 \to -\infty$ as Re $(z) \to +\infty$ in D_2 and D_3. On referring to (2.9) we see that, for large z in D_2 and D_3,

$$\zeta_1 = \pm\left[ihe^x(\cos y + i \sin y) - \frac{iu\pi}{2} + O(u^2e^{-x}) + O(u^{-1})\right].$$

The stipulated asymptotic behaviour of ζ_1 requires that the positive sign be chosen for $z \in D_3$ and the negative sign chosen for $z \in D_2$. This leads to the relations

$$(3.13) \qquad \zeta_1 = ihe^z - \frac{iu\pi}{2} + O(u^2e^{-x}) + O(u^{-1})$$

for Re $(z) \to +\infty$ in D_3, and

$$(3.14) \qquad \zeta_1 = -\left[ihe^z - \frac{iu\pi}{2} + O(u^2e^{-x}) + O(u^{-1})\right]$$

for Re $(z) \to +\infty$ in D_2. The corresponding equations giving the behaviour of ζ_1 as Re $(z) \to -\infty$ (in D_1) are obtained by applying the relation (2.14) to (3.12) or (3.13), as the case may be. Thus we obtain the equations

$$(3.15) \qquad \zeta_1 = -2u \log\left(\frac{2u}{he}\right) + ihe^{-z} - \frac{iu\pi}{2} + O(u^2e^x) + O(u^{-1})$$

for Re $(z) \to -\infty$ in the part of D_1 above the cut C_0, and

$$(3.16) \qquad \zeta_1 = -2u \log\left(\frac{2u}{he}\right) - ihe^{-z} + \frac{iu\pi}{2} + O(u^2e^x) + O(u^{-1})$$

as Re $(z) \to -\infty$ in the part of D_1 below C_0. The asymptotic forms of the functions ζ_2 and ζ_3 can be obtained from the appropriate formulae (3.13)-(3.16) by using the relations (i) $\zeta_2 = \zeta_3 = -\zeta_1$ for $z \in D_1$ (ii) $\zeta_2 = -\zeta_3 = -\zeta_1$ for $z \in D_2$, and (iii) $\zeta_2 = -\zeta_3 = \zeta_1$

for $z \in D_3$. It will be shown that the quantities $\varepsilon_i(a_j)$ appearing in the formula (3.12) are such that $\varepsilon_2(a_3)=0$ whilst $\varepsilon_2(a_1)$ and $\varepsilon_3(a_1)$ are each $O(u^{-1})$. We consider $\varepsilon_3(a_1)$ first and apply the inequality (3.11) in which the path of integration connecting a_1 and a_3 is the straight line $\operatorname{Im} z=\pi/2$. On this line we set $z=x+i\pi/2$ and obtain

$$(3.17)\qquad \begin{aligned}(u^2-2h^2\cosh 2z)^{1/2}&=(R^2e^{2i\theta}+2h^2\cosh 2x)^{1/2}\\ &=(R^2\cos 2\theta+2h^2\cosh 2x+iR^2\sin 2\theta)^{1/2}.\end{aligned}$$

Since $\operatorname{Re}\zeta_3\to+\infty$ as $x\to+\infty$ it is necessary to select the branch of (3.17) whose real part is positive as $x\to+\infty$. It is then readily verified that $\operatorname{Re}(u^2-2h^2\cosh 2z)^{1/2}>0$ for all real values of x.

Since $d\zeta_3=(u^2-2h^2\cosh 2z)^{1/2}\,dx$, along the chosen path, it also follows that $(d/dx)(\operatorname{Re}\zeta_3)>0$ so that $\operatorname{Re}\zeta_3$ is nondecreasing along this path. Furthermore it is proved in the Appendix that

$$(3.18)\qquad |(u^2+2h^2\cosh 2x)^{1/2}|\geqq(R^2+2h^2\cosh 2x)^{1/2}\sin\delta\geqq(R^2+h^2e^{|2x|})^{1/2}\sin\delta$$

for u in the sector $0\leqq\theta\leqq\pi/2-\delta$. On using this inequality in the definition (3.10) of the function $R(x)$ it is seen that this function is $O[e^{|2x|}(R^2+h^2e^{|2x|})^{-3/2}]$ on the line joining a_1 and a_3 and that the integral present in (3.11) is $O(u^{-1})$ so that ε_{13} is itself $O(u^{-1})$.

To prove that $\varepsilon_2(a_3)=0$ we connect a_2 and a_3 by means of the three straight lines from a_2 to $\sigma-i\pi/2$, then to $\sigma+i\pi/2$ and finally to a_3, where σ is an arbitrarily large positive number. It follows from (2.9) that, on all three parts of the chosen path, $\zeta_3\sim-ihe^z+iu\pi/2$, the signs of (2.9) being reversed since $\operatorname{Re}\zeta_3\to+\infty$ as $x\to+\infty$ in D_3. Therefore $\operatorname{Re}\zeta_3\sim he^x\sin y+\text{constant}$, which verifies that $\operatorname{Re}\zeta_3$ increases steadily from $-\infty$ at a_2 to $+\infty$ at a_3. In addition since $|(u^2-2h^2\cosh 2z)^{1/2}|\sim he^x$ we see that $R(x)=O(e^{-x})$ and that

$$\int_{a_2}^{a_3}|R|\cdot|dz|\leqq 2C\int_\sigma^\infty e^{-x}\,dx+Ce^{-\sigma}\int_{-\pi/2}^{\pi/2}dy=2Ce^{-\sigma}+\pi Ce^{-\sigma}.$$

It follows that $|\varepsilon_3(a_2)|\leqq\exp[(\pi+2)Ce^{-\sigma}]-1$ where σ is arbitrarily large, so that, on letting $\sigma\to+\infty$ we see that $\varepsilon_3(a_2)$ is actually zero.

To show that $\varepsilon_2(a_1)=O(u^{-1})$ we again apply the inequality (3.11) in which the path of integration connecting a_1 and a_2 consists of the three straight lines from a_1 to $i\pi/2$, then to $-i\pi/2$ and finally to $a_2=\infty-i\pi/2$. On the line joining the points $\pm i\pi/2$ the formula (2.12) applies, with all signs on the right-hand side reversed, since the dominant term is the logarithmic one and we require $\operatorname{Re}\zeta_1>0$ in D_1. Therefore, on this segment

$$\begin{aligned}\zeta_1&=-uz+u\log\left(\frac{2u}{he}\right)+O(1)\\ &=-iuy+u\log\left(\frac{2u}{he}\right)+O(1)\end{aligned}$$

so that $\operatorname{Re}\zeta_1=yR\sin\theta+\operatorname{Re}[u\log(2u/he)]$ which decreases when the segment is described in the stated direction, and on this segment,

$$(u^2-2h^2\cosh 2z)^{1/2}=(R^2e^{2i\theta}-2h^2\cos 2y)^{1/2}=-\mathrm{R}e^{i\theta}+O(R^{-1})$$

uniformly for $|y|\leqq\pi/2$. On the infinite parts of the path the formula (3.17) applies, except that it is now necessary to choose the value whose real part is negative as $x\to+\infty$, since then we require that $\operatorname{Re}\zeta_1\to-\infty$. It follows that $\operatorname{Re}(u^2-2h^2\cosh 2z)^{1/2}<0$, and that Re ζ_1 is decreasing, on the whole path. In addition the inequality (3.18) applies on the infinite parts so that $R(x)=O[e^{|2x|}(R^2+h^2e^{|2x|})^{-3/2}]$ on the entire path and therefore

$$\int_{a_1}^{a_2}|R|\cdot|dz|=O(u^{-1}).$$

Finally we see that $|\varepsilon_2(a_1)|\leqq \exp[0(u^{-1})]-1$ and this implies that $\varepsilon_2(a_1)$ is itself $O(u^{-1})$ as claimed. Since $\varepsilon_2(a_3)=0$ and $\varepsilon_1(a_3)$ and $\varepsilon_1(a_2)$ are $O(u^{-1})$, (3.12) reduces to

$$w_1(z)=[e^{i\pi/3}w_2(z)+e^{-i\pi/3}w_3(z)][1+O(u^{-1})]. \tag{3.19}$$

To relate $M_\nu^{(3)}(x)$ to the solutions w_1, w_2, w_3 we appeal to the property (1.3) and the Hankel asymptotic formula [1]

$$H_\nu^{(1)}(2h\cosh x)=\frac{1}{\sqrt{\pi h\cosh x}}\exp\left[2ih\cosh x-i\left(\nu+\frac{1}{2}\right)\frac{\pi}{2}\right][1+O(e^{-x})] \tag{3.20}$$

as $x\to+\infty$. It follows on combining (1.3) and (3.20) that

$$M_\nu^{(3)}(x)=\frac{1}{\sqrt{\pi h\cosh x}}\exp\left[2ih\cosh x-i\left(\nu+\frac{1}{2}\right)\frac{\pi}{2}\right][1+O(e^{-x})] \tag{3.21}$$

as $x\to+\infty$.

Since $\zeta_3\sim -ihe^z+iu\pi/2$ as Re $(z)\to+\infty$ in D_2 and D_3 the formula (3.8) applied to w_3 shows that

$$w_3(z)=\phi_3(z)e^{ihe^z-iu\pi/2}[1+\varepsilon_3(z)]. \tag{3.22}$$

The above formula for w_3 applies for all points z that can be connected to a_3 by a path along which Re ζ_3 is nondecreasing. If $z=\sigma$ is positive, and large compared with u^2, a suitable path can be composed of the straight line from σ to $(\sigma+i\pi/2)$ together with the straight line from $\sigma+i\pi/2$ to a_3. This path is the upper half of that introduced in the preceding proof that $\varepsilon_2(a_3)=0$ and on it $R=O(e^{-x})$ and Re ζ_3 is nondecreasing so that

$$\int_{\sigma}^{a_3}|R|\cdot|dz|=O(e^{-\sigma}).$$

Therefore $\varepsilon_3(\sigma)\to 0$ as $\sigma\to-\infty$. Hence, since $\phi_3\sim h^{-1/2}e^{-x/2+i\pi/4}$ as $x\to\infty$, it follows from (3.21) that

$$w_3(x)=h^{-1/2}e^{-x/2+ihe^x-iu\pi/2+i\pi/4}[1+\varepsilon_3(x)] \tag{3.23}$$

where $\varepsilon_3(x)\to 0$ as $x\to+\infty$.

Upon comparing (3.21) with (3.23) we see that

$$M_\nu^{(3)}(x)=\sqrt{\frac{2}{\pi}}\,e^{i(u-\nu)(\pi/2)-i\pi/2}w_3(x) \tag{3.24}$$

$$=\sqrt{\frac{2}{\pi}}\,e^{i(u-\nu)(\pi/2)}[e^{-i\pi/6}w_1(x)-e^{i\pi/6}w_2(x)][1+O(u^{-1})] \tag{3.25}$$

by (3.19).

We now obtain the desired formula for $M_\nu^{(3)}(x)$. Since this is to be applicable to sufficiently large values of u but bounded values of x, the latter variable will be located in the domain D_1. In this domain Re $(\zeta_1)>0$ so that ζ_1 is given by (2.12) with the signs of the terms on the right-hand side of this equation reversed. For the values of

x in question, the second term on the right-hand side of the same equation is $O(u^{-1})$ and can be absorbed into the last term so that $\zeta_1(x, u) = -ux + \frac{1}{2}g_0(u)$ and

$$w_1(x) = \phi_1(x)e^{ux-(1/2)g_0(u)}[1+\varepsilon_1(x)]$$

where $g_0(u)$ is defined by equation (2.15). Since $\zeta_2 = -\zeta_1$ in D_1 we also have the equation

$$w_2(x) = \phi_2(x)e^{-ux+(1/2)g_0(u)}[1+\varepsilon_2(x)].$$

It is clear from the reasoning followed in discussing $\varepsilon_1(a_2)$ that $\varepsilon_1(x)$ and $\varepsilon_2(x)$ are $O(u^{-1})$. On inserting the above expressions for w_1 and w_2 into (3.25) and using the facts that $\phi_1 e^{-i\pi/6} = \phi_2 e^{-i\pi/3} = \phi_3 \sim u^{-1/2}$ in D_1, we find the formula (1.5) already stated.

If $-\pi/2 < \theta < 0$ the desired formula for $M_\nu^{(3)}(x)$ is obtainable at once from (3.24) where $w_3(z)$ is given by (3.22), the connection formula (3.19) being now unnecessary. The formula (3.22) itself now applies for fixed real values of z and, for such values, $\varepsilon_3(z) = O(u^{-1})$. To establish this result it is first verified that points representing fixed real values of z can be connected to a_3 by a path satisfying the monotonic property. If x is positive and small compared with $\log(2u/he)$, a suitable path consists of the straight line from x to $x + i\pi/2$ followed by the line from $x + i\pi/2$ to a_3. In the domain D_1, Re ζ_3 is negative and equal to the expression on the right-hand side of (3.7) which, since $\sin\theta$ is negative, confirms that Re ζ_3 increases along the first part of the stated path. The remainder of the path is positioned on the line Im $z = \pi/2$ on which (3.17) holds. Since Re $\zeta_3 \to +\infty$ as $z \to a_3$ along this line it follows as before that Re ζ_3 is nondecreasing on this section of the path as well. Since $R(z) = O(u^{-1})$ on the first part of the path and $R(z) = O[e^{2x}(|u|^2 + h^2e^{2x})^{-3/2}]$ on the second part it follows, as in the earlier discussion of $\varepsilon_2(a_1)$, that $\varepsilon_3(x) = O(u^{-1})$ for positive values of x small compared with $\log(2u/he)$. On combining (3.24) and (3.22) we find since $\phi_3 \sim u^{-1/2}$ in D_1 that

$$M_\nu^{(3)}(x) = \sqrt{\frac{2}{\pi u}}\, e^{i(u-\nu)\pi/2 - i\pi/2 - \zeta_3}[1 + O(u^{-1})]$$

$$= -i\sqrt{\frac{2}{\pi u}}\, e^{i(u-\nu)\pi/2 - ux + u\log(2u/he)}[1 + O(u^{-1})]$$

for $-\pi/2 < \theta < 0$.

Appendix. It remains to establish the inequality (3.18). We write

$$(u^2 + 2h^2\cosh 2x)^{1/2} = A + iB. \tag{A.1}$$

On setting $u = t + is$, squaring both sides and equating real and imaginary parts, we find the equations

$$t^2 - s^2 + 2h^2\cosh 2x = A^2 - B^2, \qquad st = AB. \tag{A.2}$$

Upon eliminating s from the last two equations and rearranging the result we find that

$$A^2 = t^2\left[1 + \frac{2h^2\cosh 2x}{B^2 + t^2}\right]. \tag{A.3}$$

It follows from this equation that $|A| \geqq |t|$, and therefore from (A.2) that $|B| \leqq |s|$. On using the latter inequality in (A.3), we find that

$$A^2 \geqq t^2\left[1 + \frac{2h^2\cosh 2x}{s^2 + t^2}\right]$$

so that, since $u = t + is = \mathrm{Re}^{i\theta}$,

$$|A| \geqq (R^2 + 2h^2\cosh 2x)^{1/2}|\cos\theta|.$$

Since

$$|u^2+2h^2\cosh 2x|^{1/2}=(A^2+B^2)^{1/2}\geqq|A|\geqq(R^2+2h^2\cosh 2x)^{1/2}|\cos\theta|$$

the result (3.18) follows for all values of u in the sector

$$0\leqq\theta\leqq\frac{\pi}{2}-\delta.$$

REFERENCES

[1] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1965.

[2] J. MEIXNER AND F. W. SCHÄFKE, *Mathieuschefunktionen und Spharoidfunktionen*, Springer-Verlag, Berlin, 1954.

[3] D. NAYLOR, *On simplified asymptotic formulas for a class of Mathieu functions*, this Journal, 15 (1984), pp. 1205-1213.

[4] F. W. J. OLVER, *Error analysis of phase integral methods. I. General theory for simple turning points*, J. Res. Nat. Bur. Standards, 69B (1965), pp. 271-290.

[5] ———, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[6] A. SHARPLES, *Uniform asymptotic forms of modified Mathieu functions*, Quart. J. Mech. Appl. Math., 20 (1967), pp. 365-380.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

ASYMPTOTIC EXPANSION OF A MULTIPLE INTEGRAL*

J. P. McCLURE† AND R. WONG†

Abstract. An alternative derivation is given for the asymptotic expansion, as $s \to 0^+$, of the multiple integral

$$J(s) = \int_{[0,1]^n} g(x^\alpha / s) x^\beta f(x)\, dx,$$

where $g \in \mathscr{S}(\mathbb{R})$ and $f \in C^\infty(\mathbb{R}^n)$. The integral $J(s)$ is first expressed as a contour integral, in which the integrand is a meromorphic function in the complex plane. The asymptotic expansion is then obtained by moving the contour to the left, the terms of the expansion being the residues of the integrand.

Key words. asymptotic expansion, multiple integral, Mellin transform

AMS (MOS) subject classification. Primary 41A60

1. Let $g \in \mathscr{S}(\mathbb{R})$, the Schwarz space, $f \in C^\infty(\mathbb{R}^n)$ and $K_n = [0, 1]^n$. Recently, Brüning [4] has derived an asymptotic expansion, as $s \to 0^+$, for the multi-dimensional integral

$$(1) \qquad J(s) = \int_{K_n} g\left(\frac{x^\alpha}{s}\right) x^\beta \log^\gamma x f(x)\, dx,$$

where $x \in \mathbb{R}^n$, α and $\beta \in \mathbb{R}^n_+$, $\gamma \in \mathbb{Z}^n_+$, $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and

$$\log^\gamma x = \log^{\gamma_1} x_1 \cdots \log^{\gamma_n} x_n.$$

Here, $\mathbb{R}_+ = [0, \infty)$. To avoid triviality, we assume that the components $\alpha_1, \cdots, \alpha_n$ of α are all positive. According to Brüning, integrals of this type play an important role in the asymptotic expansion of the trace of the equivariant heat kernel [3]. A related integral has been treated by Barlet [1] in the analysis of complex spaces. The main result of Brüning is the following.

THEOREM. *As $s \to 0^+$, we have that*

$$(2) \qquad J(s) \sim \sum I_{jkl}(f) s^{(\beta_l + j + 1)/\alpha_l} \log^k s,$$

where the summation is over all $j \geqq 0$, $0 \leqq k \leqq |\gamma| + n - 1$ and $1 \leqq l \leqq n$. The I_{jkl} are distributions with support in the set $\{x \in K_n : x^\alpha = 0\}$.

Brüning uses an inductive argument, and bases his analysis on real-variable techniques. However, his proof is difficult to follow; in particular, we believe that the argument of §§ 5, 6 and 7, which involves passing back and forth between one- and two-variable expansions, is more subtle than is indicated. Moreover, the combination of the inductive method and the reduction to various special cases makes the calculation of the coefficients of the expansion (2) impossible in any practical sense.

* Received by the editors August 11, 1986; accepted for publication February 17, 1987.

† Department of Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grants A8069 and A7359.

In the present note, we give a quite different and more straightforward method for deriving the expansion (2). A discussion of this method, but in the case of one-dimensional integrals, is given in [2]. This method involves some elementary complex-variable techniques and has two major advantages over Brüning's argument. First, the method remains the same, whatever the number of dimensions, so that no induction is necessary on the dimension of the integral. Second, our method leads to explicit formulas for the coefficients in (2).

2. For simplicity of presentation, we shall consider only the case $n=2$, and assume that $\gamma=0$ in (1). The logarithmic factors can always be introduced subsequently by differentiating both sides of the resulting expansion for $J(s)$ with respect to the exponents of x; see the last statement in § 4. The formulas, of course, become more complicated as the dimension increases. However, the method itself remains unchanged. Thus we are concerned with only the integral

$$J(s)=\int_0^1\int_0^1 g(x^a y^b/s)x^\alpha y^\beta f(x,y)\,dx\,dy, \tag{3}$$

where a and b are positive, α and β are nonnegative, $g\in\mathcal{S}(\mathbb{R})$ and $f\in C^\infty(\mathbb{R}^2)$.

Our approach is based on some properties of Mellin transforms. We recall that the Mellin transform of a locally integrable function h on $(0,\infty)$ is defined by the integral

$$M[h;z]=\int_0^\infty t^{z-1}h(t)\,dt \tag{4}$$

when it exists. If (4) holds, then the inverse Mellin transform is given by

$$h(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} t^{-z}M[h;z]\,dz, \tag{5}$$

where the path of integration is a suitably chosen vertical line in the domain of analyticity of $M[h;z]$. For conditions and proofs of (4) and (5), see [5, p. 46, § 1.29].

Since $g\in\mathcal{S}$, it is straightforward to show that the Mellin transform $M[J;z]$ exists and is analytic in the strip $-d_{11}<\operatorname{Re} z<0$, where

$$d_{11}=\min\left\{\frac{1+\alpha}{a},\frac{1+\beta}{b}\right\}. \tag{6}$$

Furthermore, in this strip,

$$\begin{aligned} M[J;z]&=\int_0^\infty s^{z-1}J(s)\,ds\\ &=\int_0^1\int_0^1 x^\alpha y^\beta f(x,y)\left[\int_0^\infty s^{z-1}g(x^a y^b/s)\,ds\right]dx\,dy \qquad (7)\\ &=M[g;-z]\int_0^1\int_0^1 x^{\alpha+az}y^{\beta+bz}f(x,y)\,dx\,dy. \end{aligned}$$

Put

$$F(z)=\int_0^1\int_0^1 x^{\alpha+az}y^{\beta+bz}f(x,y)\,dx\,dy. \tag{8}$$

Then, by the inversion formula, we have

$$J(s)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} s^{-z}F(z)M[g;-z]\,dz, \tag{9}$$

where $-d_{11}<c<0$; cf. [5, Thm. 28]. Note that the Mellin transform $M[g;-z]$ is analytic in the half-plane $\operatorname{Re} z<0$ and is bounded in $\operatorname{Re} z\leqq c$ for any $c<0$. Also, the double integral $F(z)$ is analytic in the half-plane $-d_{11}<\operatorname{Re} z$. In the following section, we shall show that $F(z)$ can be analytically continued to a meromorphic function in the entire z-plane.

3. For $-d_{11}<\operatorname{Re} z<0$, integration by parts gives

$$\begin{aligned} F(z)&=\frac{1}{\alpha+az+1}\left[\int_0^1 y^{\beta+bz}f(1,y)\,dy-\int_0^1\int_0^1 x^{\alpha+az+1}y^{\beta+bz}f_{1,0}(x,y)\,dx\,dy\right]\\ &=\frac{1}{(\alpha+az+1)(\beta+bz+1)}\left[f(1,1)-\int_0^1 y^{\beta+bz+1}f_{0,1}(1,y)\,dy\right.\\ &\qquad -\int_0^1 x^{\alpha+az+1}f_{1,0}(x,1)\,dx\\ &\qquad \left.+\int_0^1\int_0^1 x^{\alpha+az+1}y^{\beta+bz+1}f_{1,1}(x,y)\,dx\,dy\right]. \end{aligned} \tag{10}$$

The double integral on the extreme right is analytic for $\operatorname{Re} z>-d_{22}$, where

$$d_{22}=\min\left\{\frac{\alpha+2}{a},\frac{\beta+2}{b}\right\}.$$

Thus, through (10), $F(z)$ is extended to a meromorphic function in the half-plane $\operatorname{Re} z>-d_{22}$ with at least one pole at $z=-d_{11}$ and possibly two poles, one at $z=-(\alpha+1)/a$ and one at $z=-(\beta+1)/b$, depending on whether $d_{22}>\max\{(\alpha+1)/a,(\beta+1)/b\}$. Equation (10) also shows that $F(z)=O((\operatorname{Im} z)^{-2})$ as $z\to\infty$ along vertical lines in $\operatorname{Re} z>-d_{22}$. Now observe that the last double integral in (10) is in exactly the same form as $F(z)$. Hence, the procedure in (10) can be repeated, and n-applications of this give

$$\begin{aligned} F(z)=\sum_{k=1}^{n}&\left[\prod_{j=1}^{k}(\alpha+az+j)(\beta+bz+j)\right]^{-1}\\ &\cdot\left\{f_{k-1,k-1}(1,1)-\int_0^1 y^{\beta+bz+k}f_{k-1,k}(1,y)\,dy-\int_0^1 x^{\alpha+az+k}f_{k,k-1}(x,1)\,dx\right\}\\ &+\left[\prod_{j=1}^{n}(\alpha+az+j)(\beta+bz+j)\right]^{-1}\int_0^1\int_0^1 x^{\alpha+az+n}y^{\beta+bz+n}f_{n,n}(x,y)\,dx\,dy. \end{aligned}$$

Further integration by parts shows that for any positive integer n,

$$\begin{aligned} F(z)= &\sum_{k,l=1}^{n} (-1)^{k+l} f_{k-1,l-1}(1,1)\left[\prod_{j=1}^{k} (\alpha+az+j) \prod_{i=1}^{l} (\beta+bz+i)\right]^{-1} \\ &+\sum_{k=1}^{n} (-1)^{k+n-1}\left[\prod_{j=1}^{k} (\alpha+az+j) \prod_{i=1}^{n} (\beta+bz+i)\right]^{-1} \\ &\cdot \int_0^1 y^{\beta+bz+n} f_{k-1,n}(1,y)\,dy \\ &+\sum_{l=1}^{n} (-1)^{l+n-1}\left[\prod_{j=1}^{n} (\alpha+az+j) \prod_{i=1}^{l} (\beta+bz+i)\right]^{-1} \\ &\cdot \int_0^1 x^{\alpha+az+n} f_{n,l-1}(x,1)\,dx + \left[\prod_{j=1}^{n} (\alpha+az+j) \prod_{i=1}^{n} (\beta+bz+i)\right]^{-1} \\ &\cdot \int_0^1 \int_0^1 x^{\alpha+az+n} y^{\beta+bz+n} f_{n,n}(x,y)\,dx\,dy. \end{aligned} \tag{11}$$

From (11) it is evident that $F(z)$ is meromorphic for $\operatorname{Re} z > -d_{n+1,n+1}$, where

$$d_{nn}=\min\left\{\frac{\alpha+n}{a}, \frac{\beta+n}{b}\right\}.$$

Since $d_{nn}\to\infty$ as $n\to\infty$, we have proved the following result.

LEMMA. *The function $F(z)$, defined by* (8), *can be analytically continued to a meromorphic function in the entire z-plane, with poles at $-(\alpha+n)/a$, $-(\beta+n)/b$ ($n=1,2,\cdots$). Furthermore, $F(z)=O((\operatorname{Im} z)^{-2})$, as $|z|\to\infty$, uniformly in any strip $-\infty<d\leqq \operatorname{Re} z\leqq c<0$.*

To facilitate the calculation of the residues of $F(z)$, we state the following variant of (11), which is obtained by varying the number of integrations by parts. For any positive integers n and m, we have

$$\begin{aligned} F(z)= &\sum_{k=1}^{n}\sum_{l=1}^{m} (-1)^{k+l} f_{k-1,l-1}(1,1)\left[\prod_{j=1}^{k} (az+\alpha+j) \prod_{i=1}^{l} (bz+\beta+i)\right]^{-1} \\ &+\sum_{k=1}^{n} (-1)^{k+m-1}\left[\prod_{j=1}^{k} (az+\alpha+j) \prod_{i=1}^{m} (bz+\beta+i)\right]^{-1} \\ &\cdot \int_0^1 y^{\beta+bz+m} f_{k-1,m}(1,y)\,dy \\ &+\sum_{l=1}^{m} (-1)^{l+n-1}\left[\prod_{j=1}^{n} (az+\alpha+j) \prod_{i=1}^{l} (bz+\beta+i)\right]^{-1} \\ &\cdot \int_0^1 x^{\alpha+az+n} f_{n,l-1}(x,1)\,dx \\ &+(-1)^{n+m}\left[\prod_{j=1}^{n} (az+\alpha+j) \prod_{i=1}^{m} (bz+\beta+i)\right]^{-1} \\ &\cdot \int_0^1 \int_0^1 x^{\alpha+az+n} y^{\beta+bz+m} f_{n,m}(x,y)\,dx\,dy. \end{aligned} \tag{12}$$

Note that each integral on the right-hand side of (12) is analytic in $\operatorname{Re} z > -d_{n+1\,m+1}$, where

$$d_{nm} = \min\left\{\frac{\alpha+n}{a}, \frac{\beta+m}{b}\right\},$$

and that the sequence d_{nm} is monotonically increasing with respect to each of n and m. Moreover, $d_{nm} \to \infty$ as $n, m \to \infty$.

Now let w be a pole of $F(z)$, and let n and m be the smallest positive integers such that $w > -d_{n+1\,m+1}$. To be more specific, we suppose that $w = -(\alpha+n)/a$, and that $-(\alpha+n)/a \neq -(\beta+m)/b$ for all positive integers m. Thus w is a simple pole. From (12) we have

$$\begin{aligned}\operatorname{Res}\left[F; -\frac{\alpha+n}{a}\right] = \frac{1}{a(n-1)!} & \\ \cdot\Bigg\{ & \sum_{l=1}^{m} (-1)^{l-1} f_{n-1,l-1}(1,1) \prod_{i=1}^{l}\left(\beta+i-b\frac{\alpha+n}{a}\right)^{-1} \\ & + (-1)^m \prod_{i=1}^{m}\left(\beta+i-b\frac{\alpha+n}{a}\right)^{-1} \\ & \qquad \cdot \int_0^1 y^{\beta+m-b(n+\alpha)/a} f_{n-1,m}(1,y)\,dy \\ & + \sum_{l=1}^{m} (-1)^{l} \prod_{i=1}^{l}\left(\beta+i-b\frac{\alpha+n}{a}\right)^{-1} \\ & \qquad \cdot \int_0^1 f_{n,l-1}(x,1)\,dx + (-1)^{m-1} \prod_{i=1}^{m}\left(\beta+i-b\frac{\alpha+n}{a}\right)^{-1} \\ & \qquad \cdot \int_0^1\int_0^1 y^{\beta+m-b(n+\alpha)/a} f_{n,m}(x,y)\,dx\,dy \Bigg\}.\end{aligned} \tag{13}$$

Performing the obvious integration with respect to x in the second sum and the last term on the right simplifies (13) to

$$\begin{aligned}\operatorname{Res}\left[F; -\frac{\alpha+n}{a}\right] = \frac{1}{a(n-1)!}\Bigg\{ & \sum_{l=1}^{m} (-1)^{l+1} \prod_{i=1}^{l}\left(\beta+i-b\frac{\alpha+n}{a}\right)^{-1} f_{n-1,l-1}(0,1) \\ & + (-1)^m \prod_{i=1}^{m}\left(\beta+i-b\frac{n+\alpha}{a}\right)^{-1} \\ & \qquad \cdot \int_0^1 y^{\beta+m-b(n+\alpha)/a} f_{n-1,m}(0,y)\,dy \Bigg\}.\end{aligned} \tag{14}$$

Similarly, if $w = -(\beta+m)/b$ is a simple pole, and if n, m are chosen as before, then we have

$$\begin{aligned}\operatorname{Res}\left[F; -\frac{\beta+m}{b}\right] = \frac{1}{b(m-1)!}\Bigg\{ & \sum_{k=1}^{n} (-1)^{k+1} \prod_{j=1}^{k}\left(\alpha+j-a\frac{\beta+m}{b}\right)^{-1} f_{k-1,m-1}(1,0) \\ & + (-1)^n \prod_{j=1}^{n}\left(\alpha+j-a\frac{\beta+m}{b}\right)^{-1} \\ & \qquad \cdot \int_0^1 x^{\alpha+n-a(\beta+m)/b} f_{n,m-1}(x,0)\,dx \Bigg\}.\end{aligned} \tag{15}$$

Finally, suppose that for some integers n and m, we have $(\alpha+n)/a=(\beta+m)/b$. Then $F(z)$ has a double pole at $-(\alpha+n)/a$. The principal part of $F(z)$ can be calculated from (12), and is given by

$$\begin{aligned}
&\frac{1}{ab(n-1)!(m-1)!}\biggl\{f_{n-1,m-1}(1,1)-\int_0^1 f_{n-1,m}(1,y)\,dy\\
&\qquad -\int_0^1 f_{n,m-1}(x,1)\,dx+\int_0^1\int_0^1 f_{n,m}(x,y)\,dx\,dy\biggr\}\left(z+\frac{\alpha+n}{a}\right)^{-2}\\
&+\frac{1}{ab(n-1)!(m-1)!}\biggl\{\biggl[f_{n-1,m-1}(1,1)-\int_0^1 f_{n-1,m}(1,y)\,dy\\
&\qquad -\int_0^1 f_{n,m-1}(x,1)\,dx+\int_0^1\int_0^1 f_{n,m}(x,y)\,dx\,dy\biggr](aH_{n-1}+bH_{m-1})\\
&\qquad -b\int_0^1 \log y f_{n-1,m}(1,y)\,dy\\
&\qquad -a\int_0^1 \log x f_{n,m-1}(x,1)\,dx\\
&\qquad +\int_0^1\int_0^1 (a\log x+b\log y)f_{n,m}(x,y)\,dx\,dy\\
&\qquad -a\sum_{k=1}^{n-1}(n-k-1)!f_{k-1,m-1}(1,1)\\
&\qquad -b\sum_{l=1}^{m-1}(m-l-1)!f_{n-1,l-1}(1,1)\\
&\qquad +a\sum_{k=1}^{n-1}(n-k-1)!\int_0^1 f_{k-1,m}(1,y)\,dy\\
&\qquad\qquad +b\sum_{l=1}^{m-1}(m-l-1)!\int_0^1 f_{n,l-1}(x,1)\,dx\biggr\}\left(z+\frac{\alpha+n}{a}\right)^{-1},
\end{aligned}$$

where $H_n=\sum_{k=1}^n(1/k)$. Upon simplification, the above expression reduces to

$$\begin{aligned}
\operatorname{Prin}\left[F;-\frac{\alpha+n}{a}\right]&=\frac{1}{ab(n-1)!(m-1)!}f_{n-1,m-1}(0,0)\left(z+\frac{\alpha+n}{a}\right)^{-2}\\
&\quad+\frac{1}{ab(n-1)!(m-1)!}\\
&\quad\cdot\biggl\{f_{n-1,m-1}(0,0)(aH_{n-1}+bH_{m-1})\\
&\qquad -b\int_0^1 \log y f_{n-1,m}(0,y)\,dy-a\int_0^1 \log x f_{n,m-1}(x,0)\,dx\\
&\qquad -a\sum_{k=1}^{n-1}(n-k-1)!f_{k-1,m-1}(1,0)\\
&\qquad\qquad -b\sum_{l=1}^{m-1}(m-l-1)!f_{n-1,l-1}(0,1)\biggr\}\left(z+\frac{\alpha+n}{a}\right)^{-1}.
\end{aligned}\tag{16}$$

Observe that the right-hand side of (14) can be regarded as the result of a distribution acting on the C^∞-function $f(x, y)$ and supported on the coordinate axes. The same remark applies to (15) and (16).

4. We are now ready to derive the asymptotic expansion of $J(s)$. First we return to the contour integral representation in (9). For any positive number $d > -c$ we can choose an arbitrarily small ε such that $F(z)$ has no pole in the strip $-d < \operatorname{Re} z \leqq -d+\varepsilon$. Since $M[g; -z]$ is analytic and bounded for $\operatorname{Re} z \leqq c < 0$, by the lemma above we can shift the contour in (9) to the left and obtain

$$\begin{aligned} J(s) = &\sum_{-d+\varepsilon<\operatorname{Re} z<c} \operatorname{Res}\{s^{-z}F(z)M[g; -z]\} \\ &+\frac{1}{2\pi i}\int_{-d+\varepsilon-i\infty}^{-d+\varepsilon+i\infty} s^{-z}F(z)M[g; -z]\,dz. \end{aligned} \tag{17}$$

The last integral is clearly bounded by a constant multiple of $s^{d-\varepsilon}$. Thus we have

$$J(s) = \sum_{-d<\operatorname{Re} z<c} \operatorname{Res}\{s^{-z}F(z)M[g; -z]\} + O(s^{d-\varepsilon}), \tag{18}$$

as $s \to 0^+$. Each simple pole $w = -(\alpha+n)/a$ of $F(z)$ contributes a term of the form

$$s^{(\alpha+n)/a} M\left[g; \frac{\alpha+n}{a}\right] \operatorname{Res}\left[F; -\frac{\alpha+n}{a}\right] \tag{19}$$

to the asymptotic expansion of $J(s)$. Similarly, each simple pole $w = -(\beta+m)/b$ contributes a term of the form

$$s^{(\beta+m)/b} M\left[g; \frac{\beta+m}{b}\right] \operatorname{Res}\left[F; -\frac{\beta+m}{b}\right]. \tag{20}$$

If $w = -(\alpha+n)/a = -(\beta+m)/b$ is a double pole of $F(z)$, then the residue of

$$s^{-z}F(z)M[g; -z] = \exp(-z \log s)F(z)M[g; -z]$$

at w can be calculated from (16), and to the expansion of $J(s)$ the point contributes the term

$$a_n(f)s^{(\alpha+n)/a}\log s + b_n(f)s^{(\alpha+n)/a}, \tag{21}$$

where

$$a_n(f) = -\frac{f_{n-1,m-1}(0, 0)}{ab(n-1)!(m-1)!} M\left[g; \frac{\alpha+n}{a}\right] \tag{22}$$

and

$$\begin{aligned} b_n(f) = &-\frac{f_{n-1,m-1}(0, 0)}{ab(n-1)!(m-1)!} M'\left[g; \frac{\alpha+n}{a}\right] + \frac{1}{ab(n-1)!(m-1)!} \\ &\cdot\left\{ f_{n-1,m-1}(0, 0)(aH_{n-1}+bH_{m-1}) \right. \\ &-b\int_0^1 \log y f_{n-1,m}(0, y)\,dy - a\int_0^1 \log x f_{n,m-1}(x, 0)\,dx \\ &-a\sum_{k=1}^{n-1}(n-k-1)!f_{k-1,m-1}(1, 0) \\ &\left. -b\sum_{l=1}^{m-1}(m-l-1)!f_{n-1,l-1}(0, 1)\right\} M\left[g; \frac{\alpha+n}{a}\right]. \end{aligned} \tag{23}$$

Inserting (19), (20) and (21) into (18), we obtain an asymptotic expansion for $J(s)$ in terms of powers and logarithms of s. In view of the statement following (16), the theorem in § 1 is proved for the case $n=2$ and $\gamma=0$.

As remarked before, the above method can easily be extended to integrals of higher dimensions. The formulas for the coefficients in the expansion, however, will become overwhelmingly complicated, and hence will not be given. The case $\gamma\neq 0$ in (1) can be included by employing a frequently used device in asymptotics, namely, we first replace each term in the sum in (17) by its corresponding value given in (19), (20) or (21), and then differentiate both sides of (17) with respect to α and β an appropriate number of times.

REFERENCES

[1] D. BARLET, *Développement asymptotique des fonctions obtenue par intégration sur les fibres*, Invent. Math., 68 (1982), pp. 129–174.

[2] N. BLEISTEIN AND R. A. HANDELSMAN, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.

[3] J. BRÜNING AND E. HEINTZE, *The Minakschisundaram-Pleijel expansion in the equivariant case*, Duke Math. J., 51 (1984), pp. 959–980.

[4] J. BRÜNING, *On the asymptotic expansion of some integrals*, Arch. Math., 42 (1984), pp. 253–259.

[5] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, 2nd ed., Oxford University Press, London, 1948.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

011

INCOMPLETE LAPLACE INTEGRALS: UNIFORM ASYMPTOTIC EXPANSION WITH APPLICATION TO THE INCOMPLETE BETA FUNCTION*

N. M. TEMME†

Abstract. The incomplete Laplace integral

$$\frac{1}{\Gamma(\lambda)}\int_\alpha^\infty t^{\lambda-1}\,e^{-zt}f(t)\,dt$$

is considered for large values of z. Both λ and α are uniformity parameters in $[0,\infty)$. The basic approximant is an incomplete gamma function, that is, the above integral with $f=1$. Also, a loop integral in the complex plane is considered with the same asymptotic features. The asymptotic expansions are furnished with error bounds for the remainders in the expansions. The results of the paper combine four kinds of asymptotic problems considered earlier. An application is given for the incomplete beta function. The present investigations are a continuation of earlier works of the author for the above integral with $\alpha=0$. The new results are significantly based on the previous case.

Key words. uniform asymptotic expansion of integrals, incomplete gamma function, incomplete beta function, incomplete Laplace integral, construction of error bounds

AMS(MOS) subject classification. 41 A60, 30 F15, 33 A15, 44 A10

1. Introduction. This paper is the third in a set dealing with uniform asymptotic expansions of Laplace type integrals. The previous papers are [11] and [12]. In the present paper, we consider the integral

$$F_\lambda(z,\alpha)=\frac{1}{\Gamma(\lambda)}\int_\alpha^\infty t^{\lambda-1}\,e^{-zt}f(t)\,dt, \tag{1.1}$$

where z is a large parameter and f is holomorphic in a domain Ω that contains the nonnegative reals; λ, α and z are real variables for which the integral is properly defined. Say, $\alpha\geqq 0$, $\lambda\geqq 0$ and $z>0$. An interpretation of $F_0(z,0)$ follows from

$$\lim_{\lambda\to 0}F_\lambda(z,\alpha)=\begin{cases}0 & \text{if } \alpha>0,\\ f(0) & \text{if } \alpha=0.\end{cases}$$

The second case follows from integration by parts.

We are interested in the asymptotic expansion of (1.1) for $z\to\infty$ which is uniformly valid with respect to both λ and α in $[0,\infty)$. The parameters λ and α may be coupled with the large parameter z, or they may range independently through the uniformity interval. For a description of the various asymptotic features, four different cases with their own asymptotic phenomena can be distinguished.

(i) α fixed, λ fixed. For this classical case Watson's lemma gives an expansion. When $\alpha=0$, $f(t)$ is expanded in powers of t, when $\alpha>0$, $t^{\lambda-1}f(t)$ is expanded in powers of $t-\alpha$. See [6, p. 113].

(ii) $\alpha\geqq 0$, λ fixed. An incomplete gamma function (i.e., (1.1) with $f=1$) is needed to describe the uniform transition of $\alpha=0$ to $\alpha>0$; [4], [8], [9] and [14] are appropriate references. The asymptotic feature is the possible coalescence of two critical points: $t=0$ (an algebraic singularity) and $t=\alpha$ (end-point of integration).

* Received by the editors November 11, 1985; accepted for publication (in revised form) May 13, 1986.

† Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

(iii) $\alpha = 0$, $\lambda \geqq 0$. The saddle point of $t^{\lambda} e^{-zt}$, which occurs at $t = \mu := \lambda / z$, may coalesce with $t = 0$, the end-point of integration. In that event (i.e., when $\mu \to 0$) the saddle point disappears, since e^{-zt} does not have a saddle point. No extra special function is needed to describe this feature; in fact the (complete) gamma function, which is incorporated in (1.1) for normalization, can handle this case. See [11] and [12].

(iv) $\alpha > 0$ (fixed), $\lambda \geqq 0$. When $\mu := \lambda / z$ is larger than α, the saddle point is inside the interval of integration; otherwise it is outside. This transition can be described by using an error function. A transformation gives an integral of the form

$$\int_{\eta}^{\infty} e^{-zu^2} g(u)\, du,$$

and here the transition occurs at $\eta = 0$. See [10]; similar cases are considered in [2] and [13].

These four cases are combined in our approach, where $\alpha \geqq 0$, $\lambda \geqq 0$. As in case (ii), the basic approximant is the incomplete gamma function. However, in case (ii) the full ranges of both parameters of this approximant are not completely exploited. As discussed in [7], it is expected that a two-variable approximant is needed to handle a three-variable case as in (1.1).

Apart from combining four existing methods, our results are interesting in view of applications. We consider the well-known incomplete beta function $I_x(p, q)$, and we give an expansion for large values of p, valid uniformly with respect to both x and q; $x \in [0, 1]$, $q \geqq 0$. Since $I_x(p, q) = 1 - I_{1-x}(q, p)$, the parameters p and q are interchangeable. So we solve an open problem mentioned in [10], where the incomplete beta function is considered as belonging to case (ii), as well as to case (iv). A transition from one case to another was not available at that moment.

The plan of the paper is as follows. In § 2 we give the formal expansion of (1.1). It appears that an essential part of the expansion is that of the complete integral (1.1) with $\alpha = 0$. Subsequent sections give representations of the remainders, conditions on f, the asymptotic nature of the expansions and the construction of error bounds. In § 8 we consider analogous results for loop integrals in the complex plane. A loop integral with essentially the same asymptotic features as (1.1) has the form

$$\int t^{-\lambda} e^{zt} \frac{f(t)}{t - \alpha}\, dt.$$

For $\alpha = 0$ it reduces, just as (1.1), to a form that we considered earlier. Section 9 gives two new expansions for the incomplete beta function

Terminology. We call a variable *fixed* when it is independent of z, λ and α. A sequence of functions $\{\psi_s\}$ is called an *asymptotic scale* when, for $s = 0, 1, \cdots$, $\psi_{s+1} = o(\psi_s)$ as $z \to \infty$. The formal series $\sum_{s=0}^{\infty} f_s(z)$ is said to be an *asymptotic expansion* of $F(z)$ *with respect to the scale* $\{\psi_s\}$, if for $n = 0, 1, 2, \cdots$

$$F(z) - \sum_{s=0}^{n} f_n(z) = o(\psi_n) \quad \text{as } z \to \infty. \tag{1.2}$$

In this case we write

$$F(z) \sim \sum_{s=0}^{\infty} f_s(z), \qquad \{\psi_s\} \quad \text{as } z \to \infty. \tag{1.3}$$

In uniform expansions it is required that the "o" symbols in (1.2) and in the definition of the scale hold uniformly (with respect to α, λ or $\mu = \lambda / z$ in certain domains, say). See [3].

2. Uniform expansions: construction of the formal series.

Before considering the general case (1.1), we repeat the procedure for $F_\lambda(z, 0)$, which will be denoted by

$$F_\lambda(z)=\frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}f(t)\,dt. \tag{2.1}$$

This function and its expansion play an essential role in the expansion of (1.1). The following integration by parts procedure takes into account the role of the *critical point*: the saddle point of $t^\lambda\,e^{-zt}$, i.e., the point $t=\mu$, where

$$\mu=\lambda/z. \tag{2.2}$$

We write

$$f(t)=f(\mu)+(t-\mu)g(t), \tag{2.3}$$

and we obtain

$$\begin{aligned}F_\lambda(z)&=f(\mu)z^{-\lambda}+\frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}(t-\mu)g(t)\,dt\\ &=f(\mu)z^{-\lambda}-\frac{1}{z\Gamma(\lambda)}\int_0^\infty g(t)\,d(t^\lambda\,e^{-zt})\\ &=f(\mu)z^{-\lambda}+\frac{1}{z\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}f_1(t)\,dt,\end{aligned}$$

where we assume that integrated terms at $t=0$, $t=\infty$ vanish, and

$$f_1(t)=t\frac{d}{dt}g(t)=t\frac{d}{dt}\frac{f(t)-f(\mu)}{t-\mu}.$$

Repeating this process, we obtain the formal expansion

$$F_\lambda(z)\sim z^{-\lambda}\sum_{s=0}^\infty f_s(\mu)z^{-s},\qquad z\to\infty, \tag{2.4}$$

where $f_0(t)=f(t)$ and

$$f_{s+1}(t)=t\frac{d}{dt}\frac{f_s(t)-f_s(\mu)}{t-\mu},\qquad s=0,1,\cdots. \tag{2.5}$$

For the general case (1.1) we again take (2.3) as the first step. Now we have an integrated term at $t=\alpha$. It is not difficult to see that we obtain the formal expansion

$$F_\lambda(z,\alpha)\sim z^{-\lambda}Q(\lambda,\alpha z)\sum_{s=0}^\infty f_s(\mu)z^{-s}+\frac{\alpha^\lambda\,e^{-\alpha z}}{z\Gamma(\lambda)}\sum_{s=0}^\infty B_s(\alpha)z^{-s},\qquad z\to\infty, \tag{2.6}$$

where $f_s(\mu)$ are the same as in (2.4), $B_s(\alpha)$ are defined by

$$B_s(\alpha)=\frac{f_s(\alpha)-f_s(\mu)}{\alpha-\mu},\qquad s=0,1,\cdots, \tag{2.7}$$

and $Q(a,x)$ is the incomplete gamma function ratio

$$Q(a,x)=\frac{1}{\Gamma(a)}\int_x^\infty t^{a-1}\,e^{-t}\,dt. \tag{2.8}$$

We observe that the first series in (2.6) does not depend on α; in fact we recognize the expansion given in (2.4). Furthermore, the integrated terms at $t=\alpha$, which generate the second series, all vanish when $\alpha\to0$.

These observations lead us to the representation, including the definition of a new function B_λ,

$$F_\lambda(z,\alpha)=Q(\lambda,\alpha z)F_\lambda(z)+\frac{\alpha^\lambda e^{-\alpha z}}{z\Gamma(\lambda)}B_\lambda(z,\alpha), \tag{2.9}$$

which should not be interpreted as an asymptotic relation, but as an exact identity. We consider the incomplete gamma ratio as a known function, of which the asymptotic features are well known (see [10]). For numerical aspects concerning this function see [5]. As mentioned above, the asymptotic expansion of $F_\lambda(z)$, i.e., (2.4), is also settled earlier. More details on this point will be given below. So we are left with the function $B_\lambda(z,\alpha)$, of which the asymptotic expansion formally follows from the second series in (2.6).

A somewhat different method used to obtain the expansion for $B_\lambda(z,\alpha)$ is based on a differential equation for this function. By differentiating (2.9) with respect to α, we easily obtain

$$\frac{\alpha}{z}B'_\lambda(z,\alpha)+(\mu-\alpha)B_\lambda(z,\alpha)=z^\lambda F_\lambda(z)-f(\alpha). \tag{2.10}$$

Substitution of (2.4) and of the formal series

$$B_\lambda(z,\alpha)\sim\sum_{s=0}^{\infty}\frac{B_s(\alpha)}{z^s} \tag{2.11}$$

into (2.10) shows that this equation is formally satisfied if

$$\begin{aligned}(\mu-\alpha)B_s(\alpha)&=f_s(\mu)-\alpha B'_{s-1}(\alpha), \qquad s=1,2,\cdots,\\ B_0(\alpha)&=[f(\alpha)-f(\mu)]/(\alpha-\mu).\end{aligned} \tag{2.12}$$

Here, and in (2.10), the prime denotes differentiation with respect to α. It easily follows that (2.12) generates the same coefficients $B_s(\alpha)$ as those defined in (2.7). Therefore, by using (2.9), we again arrive at (2.6).

The following integral

$$E_\lambda(z,\alpha)=\frac{1}{\Gamma(\lambda)}\int_0^\alpha t^{\lambda-1}e^{-zt}f(t)\,dt \tag{2.13}$$

is strongly related to (1.1). It has a similar representation as (2.9). When we use the following complementary relations

$$E_\lambda(z,\alpha)+F_\lambda(z,\alpha)=F_\lambda(z), \qquad Q(\lambda,\alpha z)+P(\lambda,\alpha z)=1, \tag{2.14}$$

we obtain

$$E_\lambda(z,\alpha)=P(\lambda,\alpha z)F_\lambda(z)-\frac{\alpha^\lambda e^{-\alpha z}}{z\Gamma(\lambda)}B_\lambda(z,\alpha). \tag{2.15}$$

Consequently, when we give expansions for $F_\lambda(a)$ and $B_\lambda(z,\alpha)$, the results can be used for both integrals (1.1) and (2.13). The function $P(a,x)$ again is an incomplete gamma function, with representation

$$P(a,x)=\frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}\,dt. \tag{2.16}$$

3. Representations for the remainders. We introduce remainders for (2.4) and (2.11) by writing

$$z^\lambda F_\lambda(z)=\sum_{s=0}^{n-1} f_s(\mu)z^{-s}+z^{-n}\bar{f}_n, \tag{3.1}$$

$$B_\lambda(z,\alpha)=\sum_{s=0}^{n-1} B_s(\alpha)z^{-s}+z^{-n}\bar{B}_n, \tag{3.2}$$

where $n=0,1,2,\cdots$. When $n=0$ the sums are empty and they have to be replaced by 0.

The procedure leading to (2.4) yields for $\bar{f}_n$ the representation

$$\bar{f}_n=\frac{z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}f_n(t)\,dt. \tag{3.3}$$

To obtain a representation for $\bar{B}_n$, we write (2.9) in the form

$$B_\lambda(z,\alpha)=z\,e^{\alpha z}\alpha^{-\lambda}\int_\alpha^\infty t^{\lambda-1}\,e^{-zt}[f(t)-z^\lambda F_\lambda(z)]\,dt. \tag{3.4}$$

Writing

$$B_\lambda(z,\alpha)=B_0(\alpha)+z^{-1}\bar{B}_1,\qquad z^\lambda F_\lambda(z)=f(\mu)+z^{-1}\bar{f}_1,$$

and using integration by parts in the form

$$t^{\lambda-1}\,e^{-zt}\,dt=\frac{-1}{z(t-\mu)}\,d(t^\lambda\,e^{-zt}),$$

we obtain

$$\bar{B}_1=z\alpha^{-\lambda}\,e^{\alpha z}\int_\alpha^\infty t^{\lambda-1}\,e^{-zt}[f_1(t)-\bar{f}_1]\,dt.$$

Repeating this, and using the recursions

$$\bar{B}_n=B_n(\alpha)+z^{-1}\bar{B}_{n+1},\qquad \bar{f}_n=f_n(\mu)+z^{-1}\bar{f}_{n+1},$$

we finally have

$$\bar{B}_n=z\alpha^{-\lambda}\,e^{\alpha z}\int_\alpha^\infty t^{\lambda-1}\,e^{-zt}[f_n(t)-\bar{f}_n]\,dt, \tag{3.5}$$

where $n=0,1,2,\cdots$. For $n=0$ this equals the starting point (3.4). An equivalent representation is

$$\bar{B}_n=-z\alpha^{-\lambda}\,e^{\alpha z}\int_0^\alpha t^{\lambda-1}\,e^{-zt}[f_n(t)-\bar{f}_n]\,dt, \tag{3.6}$$

which easily follows by writing $\int_\alpha^\infty=\int_0^\infty-\int_0^\alpha$ and using (3.3).

The availability of both forms (3.5) and (3.6) is important in the analysis to be given below. Namely, for bounding $\bar{B}_n$ we always have an integral in which the saddle point μ is not an interior point of the interval of integration.

The above representations for $\bar{f}_n$ and $\bar{B}_n$ are formally obtained. In the next section we give the conditions on f to justify the above results, and to discuss the asymptotic nature of the expansions.

4. Assumptions on f. We consider real values of α, λ and z, with $\alpha, \lambda \geqq 0$, $z > 0$. We accept that f depends on the uniformity parameter μ defined in (2.2). The reason is that in applications usually a transformation to the standard form is needed, which yields a function f depending on μ. In [12] a detailed discussion of such a transformation is given. By means of several examples, it is shown that the assumption f depends on μ may be relevant and quite acceptable. The parameter α plays a completely different role, and we do not suppose that f depends on it. The example in § 9 on the incomplete beta function shows more details on the transformation to the standard form (1.1), and on the role of α and μ.

The analysis is based on the assumption that f is holomorphic in a domain of the complex plane. Again, for applications in the theory of special functions, this condition is quite natural. Another point is that part of the analysis runs rather elegantly when using complex function theory. However, the construction of the expansions, the representations of the remainders in (3.3), (3.5), and the construction of error bounds can also be given for functions f belonging to continuity classes $C^k([0, \infty))$. When $k < \infty$ we cannot, of course, define the complete expansions (2.4), (2.11).

We assume that f is holomorphic in a simply connected domain Ω of the complex plane; Ω may depend on μ and Ω should contain $\mathbb{R}^+$. We suppose that the distance $d(t)$ from $t \in \mathbb{R}^+$ to the boundary $\partial\Omega$ of Ω is increasing according to the following requirement:

$$d(t) \geqq d_0(\delta + t)^\kappa, \qquad t \geqq 0. \tag{4.1}$$

where δ, d_0 and κ are fixed, $\delta, d_0 > 0$, and $\frac{1}{2} \leqq \kappa \leqq 1$.

It follows that, for large μ, the singularities of f are rather far from the saddle point $t = \mu$, this distance being $\mathcal{O}(\mu^\kappa)$. This condition is important for investigating the asymptotic nature of the expansion (2.4). The requirement that (4.1) holds for any positive t, and not only for $t = \mu$, is important for expansion (2.11).

A geometrical interpretation of (4.1) is as follows. Let D_t be the disc around $t \in \mathbb{R}^+$ with radius $d_0(\delta + t)^\kappa$. Then the above condition implies that Ω contains the subset

$$\Omega_0 = \bigcup_{t \geqq 0} D_t. \tag{4.2}$$

When $\kappa = \frac{1}{2}$ the boundary $\partial\Omega_0$ of Ω_0 is a parabola; that is,

$$\partial\Omega_0 = \{t = u + iv \mid v^2 = d_0^2(u + \delta + \tfrac{1}{4}d_0^2)\}. \tag{4.3}$$

When $\kappa = 1$ we have two possibilities depending on d_0:

(i) $0 < d_0 \leqq 1$, Ω_0 is a sector with vertex at $t = -\delta$ such that $|\arg(t + \delta)| \leqq \arcsin(d_0)$; when $d_0 = 1$ this sector is the half-plane $\operatorname{Re} t \geqq -\delta$.

(ii) $d_0 > 1$, $\Omega_0 = \mathbb{C}$.

It is clear that for $\frac{1}{2} < \kappa < 1$ the set Ω_0 is something "between" a parabola-shaped domain and a sector. Geometrically, values of κ larger than unity make no sense, although the analysis will accept such values.

We also need a growth condition on f in Ω_0. We assume that, when μ is fixed, f is of algebraic growth at infinity. That is,

$$M(\mu) = \sup_{t \in \Omega} (1 + |t|^{-p})|f(t)| \tag{4.4}$$

should exist for all finite values of μ in $[0, \infty)$. Condition (4.4) will not exclude functions in (1.1) that can be written as

$$f(t) = e^{\sigma t}\tilde{f}(t), \qquad \sigma \text{ fixed in } \mathbb{C}.$$

When in such a decomposition $\tilde{f}$ meets the above conditions, we absorb the exponential part of this splitting into $\exp(-zt)$ of (1.1), just by a shift in the large parameter z. Afterwards, we proceed with $\tilde{f}$.

5. Estimates for $f_s(t)$, $f_s(\mu)$ and $B_s(\alpha)$. The conditions on f yield estimates for the functions $f_s(t)$ defined in (2.5) and for the coefficients $f_s(\mu)$. With the help of these estimates, the asymptotic scales for the expansions (3.1), (3.2) are chosen.

Let Ω_r be the subset of Ω_0 defined by

$$\Omega_r = \bigcup_{t \geqq 0} \tilde{D}_t, \tag{5.1}$$

where $\tilde{D}_t$ is a disc around t with radius $r(\tilde{\delta}+t)^\kappa$, with $0<\tilde{\delta}<\delta, 0<r<d_0$, $\tilde{\delta}$ and r fixed. Then the derivatives of f at t can be written as

$$f^{(s)}(t) = \frac{s!}{2\pi i}\int_{C_r} \frac{f(\tau)}{(\tau-t)^{s+1}}\,d\tau, \tag{5.2}$$

where C_r is the boundary of $\tilde{D}_t$.

It follows that we can assign numbers K_s, not depending on t and μ, such that

$$|f^{(s)}(t)| \leqq K_s M(\mu)(1+|t|)^{p-s\kappa}, \qquad s=0,1,2,\cdots,$$

for all $t \in \Omega_r$, and all $\mu \in [0,\infty)$. That is,

$$f^{(s)}(t) = M(\mu)(1+|t|)^{p-s\kappa}\mathcal{O}(1), \qquad s=0,1,2,\cdots, \tag{5.3}$$

with $t \in \Omega_r$, uniformly with respect to μ in $[0,\infty)$.

The functions $f_s(t)$ defined in (2.5) are analytic in Ω. They can be expressed in terms of the derivatives of f, as follows from their definition. For t-values near μ, the functions cannot be estimated by repeated application of (2.5), owing to the factors $1/(t-\mu)$ and powers of it. Another approach is using the mean value theorem on

$$\begin{aligned} f_s(t) &= t\frac{d}{dt}\int_0^1 f'_{s-1}[\mu+\tau(t-\mu)]\,d\tau \\ &= t\int_0^1 \tau f''_{s-1}[\mu+\tau(t-\mu)]\,d\tau, \end{aligned}$$

which gives $f_s(t) = \frac{1}{2}tf''_{s-1}(\tau_s)$, where τ_s is a value between t and μ. By repeating this, we obtain

$$f_s(t) = t\sum_{j=0}^{s-1} p_j f^{(s+j+1)}(\tau_j), \qquad s \geqq 1, \tag{5.4}$$

where τ_j are between t and μ, and p_j is a homogeneous polynomial of degree j of j variables, all between t and μ. The coefficients of p_j do not depend on μ and t. Therefore, we have

$$\begin{aligned} p_j &= (1+\mu+t)^j\mathcal{O}(1), \\ f^{(s+j+1)}(\tau_j) &= M(\mu)(1+\mu+t)^{p-(s+j+1)\kappa}\mathcal{O}(1), \end{aligned}$$

with $\mu \geqq 0$, $t \geqq 0$. It follows that (5.4) can be written as

$$f_s(t) = tM(\mu)(1+\mu+t)^{p-1-s\bar{\kappa}}\mathcal{O}(1), \qquad \bar{\kappa} = 2\kappa-1, \quad s=1,2,\cdots, \tag{5.5}$$

for all $t \in \Omega_r$, uniformly with respect to $\mu \in [0,\infty)$.

The values $f_s(\mu)$, which are the coefficients of the expansion (2.4), can be written as

$$f_s(\mu)=\mu\sum_{j=0}^{s-1} q_j\mu^j f^{(s+j+1)}(\mu), \qquad s=1,2,\cdots,$$

where q_j are fixed numbers. An estimate as in (5.5) reads

$$(5.6)\quad f_s(\mu)=\mu M(\mu)(1+\mu)^{p-1-s\bar{\kappa}}\mathcal{O}(1), \qquad \bar{\kappa}=2\kappa-1, \quad s=1,2,\cdots, \quad \mu\geqq 0.$$

The coefficients $B_s(\alpha)$ defined in (2.7) can be estimated by means of

$$B_s(\alpha)=\int_0^1 f_s'[\mu+\tau(\alpha-\mu)]\,d\tau=f_s'(\tau_s),$$

where τ_s is between μ and α. Therefore,

$$B_s(\alpha)=\frac{1}{2\pi_i}\int_{C_r}\frac{f_s'(\tau)}{(\tau-\tau_s)^2}\,d\tau,$$

where C_r is a circle in Ω_r around τ_s with radius $\mathcal{O}(1+\tau_s)^\kappa=\mathcal{O}(1+\mu+\alpha)^\kappa$. By using (5.5), it follows that

$$(5.7)\qquad B_s(\alpha)=M(\mu)(1+\mu+\alpha)^{p-s\bar{\kappa}-\kappa}\mathcal{O}(1), \qquad s=0,1,\cdots,$$

with α, $\mu\geqq 0$.

Example 5.1. When $f(t)=1/(1+t)$, we have $p=\kappa=1$, and M slightly larger than 1. The first coefficients B_0, B_1 are

$$B_0(\alpha)=-1/[(\alpha+1)(\mu+1)], \qquad B_1(\alpha)=(1-\alpha\mu)/[(\alpha+1)^2(\mu+1)^3],$$

which confirms (5.7).

To conclude this section we consider limiting values of $\bar{B}_n$ at $\alpha=0$ and $\alpha=\infty$. For $\alpha\to 0$ we write $t=\alpha\tau$ in (3.6). We obtain

$$(5.8)\qquad \bar{B}_n=\frac{\bar{f}_n-f_n(0)}{\mu} \quad \text{at } \alpha=0.$$

This expression is regular at $\mu=0$. To see this, replace (5.8) by

$$\bar{B}_n=\frac{f_n(\mu)-f_n(0)}{\mu}+\frac{1}{\lambda}\bar{f}_{n+1}.$$

From (2.5) we see that $f_n(t)/t$ is regular at $t=0$, when $n\geqq 1$. Hence, using (3.3), we obtain

$$(5.9)\qquad \bar{B}_n=f_n'(0)+\int_0^\infty e^{-zt}f_{n+1}(t)t^{-1}\,dt \quad \text{at } \alpha=0, \mu=0, \text{and } n=0,1,\cdots.$$

For the limiting value of $\bar{B}_n$ at $\alpha=\infty$, consider (3.5) in the following form:

$$(5.10)\qquad \bar{B}_n=z\int_0^\infty (1+t)^{\lambda-1}\,e^{-\alpha zt}[f_n(\alpha(1+t))-\bar{f}_n]\,dt.$$

Using (5.5), and considering $\alpha\gg\mu$, we can easily estimate $\bar{B}_n$ as $\alpha\to\infty$. For instance, when $p-n\bar{\kappa}=1$ and $\lim_{t\to\infty}f_n(t)/t$ exists (and is L), we have

$$(5.11)\qquad \bar{B}_n=L \quad \text{at } \alpha=\infty, \quad \mu \text{ finite}.$$

6. The asymptotic nature of the expansions.

6. The asymptotic nature of the expansions. First we discuss the expansions (2.4), (3.1), although the complete integral (2.1) is considered in the previous paper [12]. However, there we mainly investigated an expansion somewhat different from (2.4) and, therefore, it is appropriate to consider (2.4), (3.1) in the present set-up once again.

6.1. The asymptotic nature of (3.1). We introduce the asymptotic scale $\{\psi_s z^{-s}\}$ by writing

$$\begin{aligned}\psi_0 &= M(\mu)(1+\mu)^p,\\ \psi_s &= \mu M(\mu)(1+\mu)^{p-1-s\bar{\kappa}}, \qquad s=1,2,\cdots,\end{aligned} \tag{6.1}$$

which is suggested by (5.6). Since we allow f to depend on μ, we have to use a scale that reflects the possible growth of f when μ ranges in the domain $[0,\infty)$. With the above scale we are able to control the behaviour of the remainder $\bar{f}_n$ defined in (3.3).

It is easily verified that $\{\psi_s z^{-s}\}$ is a uniform asymptotic scale with z as large parameter and μ as uniformity parameter in $[0,\infty)$. Moreover, when $\kappa>\frac{1}{2}$ (i.e., $\bar{\kappa}>0$) it also is an asymptotic scale for $\mu\to\infty$, uniformity with respect to $z\in[z_0,\infty)$, z_0 being a fixed positive number. Observe that for $\kappa<\frac{1}{2}$ the scale fails to be uniform with respect to μ on $[0,\infty)$, but it still is on compact subsets of $[0,\infty)$.

THEOREM 6.1. *For the expansions* (2.4), (3.1) *we can write*

$$z^\lambda F_\lambda(z) \sim \sum_{s=0}^{\infty} f_s(\mu) z^{-s}, \qquad \{\psi_s z^{-s}\} \quad \textit{as } z\to\infty, \tag{6.2}$$

uniformly with respect to $\mu=\lambda/z$ *in* $[0,\infty)$.

Proof. It is sufficient to show that $\bar{f}_n=\mathcal{O}(\psi_n)$, where $\bar{f}_n$ is defined in (3.3). The interval of integration in (3.3) is split up as follows:

$$[0,\infty)=\Delta_-\cup[t_-,t_+]\cup\Delta_+, \tag{6.3}$$

where

$$\Delta_-=[0,t_-],\quad \Delta_+=[t_+,\infty),\quad t_\pm=\mu\pm\varepsilon(\mu+1)^\kappa,$$

with ε fixed, and small enough such that $[t_-,t_+]$ lies inside Ω_r of (5.1). When t_- happens to be negative, we replace it by 0. For $t\in[t_-,t_+]$ we have $t=\mathcal{O}(\mu)$. Therefore, (5.5) yields

$$f_s(t)=\mu M(\mu)(1+\mu)^{p-1-s\bar{\kappa}}\mathcal{O}(1), \qquad s=1,2,\cdots.$$

Hence, (3.3) can be written as

$$\bar{f}_n=I_-+I_++\mathcal{O}(\psi_n) \quad \text{as } z\to\infty, \tag{6.4}$$

where $I_\pm$ are the contributions to (3.3) from $\Delta_\pm$. They are of order $\mathcal{O}(\psi_n)$, also. It is possible to show more: they are asymptotically equal to 0 with respect to the scale $\{\psi_s\}$. That is, $I_\pm=\mathcal{O}(\psi_m)$ for any m as $z\to\infty$, uniformly with respect to $\mu\in[0,\infty)$.

Again, the proof can be based on the estimates given in (5.5). In [12, § 3.4] a detailed analysis is given for proving that contributions from $\Delta_\pm$ for similar integrals are asymptotically negligible. This analysis will not be repeated here. □

6.2. Two lemmas for (3.2). The next step is to consider (3.2), and to estimate the remainder defined in (3.5), (3.6). The analysis boils down to the following two lemmas, the results of which are formulated in terms of strict inequalities. So we are able to use them once again in § 7 for deriving strict error bounds.

LEMMA 6.1. *Consider the function*

$$g(\alpha, \lambda, z) = \alpha^{-\lambda} e^{\alpha z} \int_0^\alpha t^\lambda e^{-zt}\, dt, \tag{6.5}$$

where $0 \leqq \alpha \leqq \mu$, $\mu = \lambda/z$. *Let* ζ *be defined by* $\frac{1}{2}\zeta^2 = \alpha - \mu + \mu \log(\mu/\alpha)$, $\zeta \geqq 0$. *Then for* $z > 0$

$$g(\alpha, \lambda, z) \leqq \min\left\{\frac{\alpha}{(\mu-\alpha)z}, \frac{\alpha\zeta}{\mu-\alpha}\sqrt{\frac{\pi}{2\zeta}}\right\}. \tag{6.6}$$

Proof. Write g in the form

$$g(\alpha, \lambda, z) = \int_0^\alpha e^{-z\phi(t)}\, dt, \qquad \phi(t) = t - \mu \log t - \alpha + \mu \log \alpha.$$

Integrating with respect to ϕ, we have for $0 \leqq \alpha < \mu$

$$g(\alpha, \lambda, z) = \int_0^\infty e^{-z\phi} \frac{t}{\mu - t}\, d\phi \leqq \frac{\alpha}{(\mu-\alpha)z},$$

which gives the first possibility in (6.6). To obtain the second one we write

$$\phi(t) = \tfrac{1}{2}w^2 + \zeta w, \qquad w \geqq -\zeta,$$

with the corresponding relations

$$t = 0 \leftrightarrow w = +\infty, \qquad t = \alpha \leftrightarrow w = 0, \qquad t = \mu \leftrightarrow w = -\zeta,$$

where ζ is defined above. Now we obtain

$$g(\alpha, \lambda, z) = \int_0^\infty e^{-z[w^2/2+\zeta w]} f(w)\, dw, \qquad f(w) = \frac{t(w+\zeta)}{\mu - t}.$$

We have $0 \leqq f(w) \leqq f(0)$, $w \geqq 0$. To verify the upper bound, we write

$$f^2(w) = 2t \frac{1 - x + x \log x}{(x-1)^2}, \qquad x = \frac{\mu}{t}.$$

The x-part of this is monotonically decreasing on the x-interval $[1, \infty)$; $x = \mu/\alpha$ corresponds to $w = 0$. Hence we obtain $f(w) \leqq \alpha\zeta/(\mu - \alpha)$. It follows that

$$g(\alpha, \lambda, z) \leqq \frac{\alpha\zeta}{\mu - \alpha} \int_0^\infty e^{-zw^2/2}\, dw,$$

which gives the second possibility in (6.6). This proves the lemma. □

LEMMA 6.2. *Consider the function*

$$G_q(\alpha, \lambda, z) = \alpha^{-\lambda} e^{\alpha z}(1+\alpha)^{-q} \int_\alpha^\infty t^{\lambda-1} e^{-zt}(1+t)^q\, dt, \tag{6.7}$$

where $\alpha \geqq 0$, $\mu \leqq \alpha$, $\mu = \lambda/z$, q *fixed*, $q \in \mathbb{R}$. *Let* ζ *be defined by*

$$\zeta = \begin{cases} +\infty, & \mu < 0, \\ +\left[2\left(\alpha - \mu + \mu \log \dfrac{\mu}{\alpha}\right)\right]^{1/2}, & 0 \leqq \mu \leqq \alpha. \end{cases}$$

Then for $z > 0$

$$G_0(\alpha, \lambda, z) \leqq \min\left\{\frac{1}{(\alpha-\mu)z}, \frac{\zeta}{\alpha-\mu}\sqrt{\frac{\pi}{2z}}\right\}. \tag{6.8}$$

Furthermore, when $0 \leqq \mu \leqq \alpha$,

$$G_q(\alpha, \lambda, z) = \mathcal{O}[G_0(\alpha, \lambda, z)] \quad \text{as } z \to \infty, \tag{6.9}$$

uniformly with respect to α, μ.

Proof. We first consider G_0, for which we obtain

$$G_0(\alpha, \lambda, z) = \int_\alpha^\infty e^{-z\phi(t)} \frac{dt}{t} = \int_0^\infty e^{-z\phi} \frac{d\phi}{t-\mu} \leqq \frac{1}{(\alpha-\mu)z},$$

where $\phi(t)$ is the same as in Lemma 6.1. Observe that this result also holds for negative values of μ. For the second possibility in (6.8), we proceed with $\mu \geqq 0$. We again write $\frac{1}{2}w^2 + \zeta w = \phi(t)$, $w \geqq -\zeta$, now with the correspondences

$$t = \infty \leftrightarrow w = \infty, \qquad t = \alpha \leftrightarrow w = 0, \qquad t = \mu \leftrightarrow w = -\zeta.$$

It follows that

$$G_0(\alpha, \lambda, z) = \int_0^\infty e^{-z[w^2/2+\zeta w]} f(w)\, dw,$$

with $f(w) = (w+\zeta)/(t-\mu)$. Using

$$f^2(w) = \frac{2}{\mu} \frac{x - \log x - 1}{(x-1)^2} \leqq f^2(0), \qquad x = \frac{t}{\mu} \geqq 1,$$

we obtain the second choice in (6.8).

When $q \leqq 0$ the proposition (6.9) is trivial. Writing

$$G_{q+1} = \frac{1+\mu}{1+\alpha} G_q + \frac{\alpha^{-\lambda}}{} \frac{e^{\alpha z}(1+\alpha)^{-q-1}}{-z} \int_\alpha^\infty (1+t)^q d(e^{-zt}t^\lambda),$$

we obtain by performing integration by parts

$$G_{q+1} = \frac{1+\mu}{1+\alpha} G_q + \frac{1}{z(1+\alpha)} - \frac{q}{z(1+\alpha)^2} G_{q-1} + \frac{q}{z(1+\alpha)} G_q.$$

Since q is fixed and $0 \leqq \mu \leqq \alpha$, the result (6.9) follows by recursion, say from negative q-values. □

Remark 6.1. The first alternatives in both (6.6) and (6.8) grow indefinitely when $\alpha \to \mu$, whereas the other ones remain finite. We have

$$\zeta/(\mu - \alpha) \to 1, \qquad \zeta/(\alpha - \mu) \to 1,$$

for (6.6), (6.8), respectively. Therefore, the second alternatives give a bound for g and G_0 valid for the whole range of parameters given in the lemmas. The first bound is given since it is sharp when z is large and α and μ are bounded away from each other. A more uniform description, which includes both alternatives in (6.6), (6.8), is possible, by using a bound in terms of an error function. That is, in fact, g and G_0 can be estimated by

$$f(0) \int_0^\infty e^{-z[w^2/2+\zeta w]}\, dw = f(0) \sqrt{\frac{\pi}{2z}}\, e^{z\zeta^2/2} \operatorname{erf} c(\zeta\sqrt{z/2}).$$

We take $\zeta = 0$ because it gives a very simple and manageable result.

6.3. The asymptotic nature of (3.2). We proceed with (3.2), and we estimate the remainder $\bar{B}_n$ defined in (3.5), (3.6). We use the asymptotic scale $\{\chi_s z^{-s}\}$ defined by

$$\chi_s = M(\mu)(1+\mu+\alpha)^{p-s\bar{\kappa}-\kappa}, \qquad s = 0, 1, 2, \cdots, \tag{6.10}$$

z is the large parameter, α and μ are uniformity parameters. The choice of scale is suggested by (5.7).

THEOREM 6.2. *For the expansions* (2.11) *and* (3.2) *we can write*

$$B_\lambda(z,\alpha)\sim\sum_{s=0}^{\infty}B_s(\alpha)z^{-s},\qquad \{\chi_s z^{-s}\}\quad \textit{as } z\to\infty, \tag{6.11}$$

uniformly with respect to μ, α *in* $[0,\infty)$.

Proof. All $\mathcal{O}$-symbols in the proof hold uniformly with respect to μ and α in $[0,\infty)$; the large parameter is not needed in some results.

It is sufficient to show that $\bar{B}_n$ of (3.5) or (3.6) is $\mathcal{O}(\chi_n)$. Write

$$\bar{B}_n=B_n(\alpha)+z^{-1}\bar{B}_{n+1},\qquad n=0,1,2,\cdots. \tag{6.12}$$

Since $B_n(\alpha)=\mathcal{O}(\chi_n)$, we proceed with $\bar{B}_{n+1}$. That is, we consider (3.5), (3.6) for $n\geqq 1$. We have two cases.

(i) $0\leqq\alpha\leqq\mu$. In (3.6) we use

$$\bar{f}_n=\mathcal{O}(\psi_n)=\mu M(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}\mathcal{O}(1),$$
$$f_n(t)=tM(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}\mathcal{O}(1),$$

where the first line follows from Theorem 6.1 and the second one from (5.5). The estimate for $f_n(t)$ gives in (3.6) a contribution

$$zM(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}g(\alpha,\lambda,z)\mathcal{O}(1), \tag{6.13}$$

where g is defined in (6.5). The above estimate for $\bar{f}_n$ gives in (3.6) a contribution

$$\begin{aligned}
&z\mu M(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}\alpha^{-\lambda}\,e^{\alpha z}\mathcal{O}(1)\int_0^\alpha t^{\lambda-1}\,e^{-zt}\,dt\\
&\quad=M(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}\alpha^{-\lambda}\,e^{\alpha z}\mathcal{O}(1)\int_0^\alpha e^{-zt}\,dt^\lambda\\
&\quad=zM(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}}[g(\alpha,\lambda,z)\mathcal{O}(1)+\mathcal{O}(z^{-1})].
\end{aligned}\tag{6.14}$$

From Lemma 6.1 it follows that

$$g(\alpha,\lambda,z)=\mathcal{O}(\sqrt{\alpha/z})=\mathcal{O}[(1+\mu+\alpha)^\kappa/\sqrt{z}].$$

Neglecting the term $\mathcal{O}(z^{-1})$ in the last line of (6.14), we conclude that both (6.13) and (6.14) are estimated by

$$\sqrt{z}M(\mu)(1+\mu+\alpha)^{p-1-n\bar{\kappa}+\kappa}\mathcal{O}(1).$$

Taking into account that in (6.12) $z^{-1}\bar{B}_{n+1}$ has to be considered, we obtain

$$\bar{B}_n=\mathcal{O}(\chi_n)+M(\mu)(1+\mu+\alpha)^{p-1-(n+1)\bar{\kappa}+\kappa}\mathcal{O}(z^{-1/2})=\mathcal{O}(\chi_n).$$

This finishes the first part of the proof.

(ii) $0\leqq\mu\leqq\alpha$. In this case the starting point for $\bar{B}_n$ is (3.5). For $\bar{f}_n$ we take the representation as in the previous case, for $f_n(t)$ we consider (5.5). We integrate by parts in the contribution from $\bar{f}_n$. Starting with (6.12) and using Lemma 6.2 twice, with $q=0$ and with $q=p-1-(n+1)\bar{\kappa}$, we obtain the required estimate $\bar{B}_n=\mathcal{O}(\chi_n)$. □

As remarked after the introduction of the scale functions ψ_s in (6.1), the large parameter z and the uniformity parameter μ are interchangeable, only if $\kappa>\frac{1}{2}$. It is important enough to formulate this property as a theorem.

THEOREM 6.3. *Let $\frac{1}{2}<\kappa\leqq 1$. Then in* (6.2) *$\mu$ may act as the large parameter and z as the uniformity parameter. In* (6.9), *α (or μ) may act as large parameter, μ (or α) and z as uniformity parameters. The uniformity domain for μ and α is $[0,\infty)$, for z it is $[z_0,\infty)$, z_0 being a fixed positive number.*

Proof. The proof follows easily from the properties of the asymptotic scales used in (6.2) and (6.11), and from the proofs of the earlier theorems. □

Remark 6.2. The above theorems are formulated and proved for real values of the parameters. By slight adaptation of the scales they hold for complex values of α and μ, as long as these values are restricted to Ω_r introduced in (5.1). Some care in the interpretation of (1.1) is needed when α assumes complex values around the origin, since t^λ is not single valued. However, the many-valuedness of $F_\lambda(z,\alpha)$ is completely described in (2.9) by the known functions $Q(\lambda, \alpha z)$ and α^λ. From (3.6) it easily follows that $B_\lambda(z,\alpha)=\bar{B}_0$ is regular at $\alpha=0$, and in fact in Ω_r. See also (5.8). Another problem is: How do we handle complex values of z? The holomorphic function f in (1.1) allows the contour of integration to be deformed. When doing so, we can extend the domains for the parameters z, α and μ considerably. We will not go into further details here for this complicated technical problem.

Remark 6.3. It is tempting to take $\{f_s(\mu)z^{-s}\}$ and $\{B_s(\alpha)z^{-s}\}$ as asymptotic scales in (6.2) and (6.1.1). However, the conditions on f do not imply that they have this property. When they do not the theorems may still be applicable; however, rather useless expansions may arise. It is instructive to consider what is happening in the case $f(t)=1+\exp(-t)$.

7. Error bounds for the remainders. The theorems of the previous section are based on the concept of generalized asymptotic expansions. The estimates for proving the asymptotic properties are given in terms of $\mathcal{O}$-symbols. So far, no information is available on the sharpness of these estimates, say in terms of exact error bounds. That is, it would be interesting to have available an estimate in the form

$$(7.1)\qquad |\bar{f}_n|\leqq K_n|f_n(\mu)|,\qquad \mu\geqq 0,\quad z\geqq z_0>0,$$

instead of $\bar{f}_n=\mathcal{O}(\psi_n)$, used in Theorem 6.1. When $f_n(\mu)$ happens to vanish, (7.1) can be modified. K_n in (7.1) may depend on z and μ.

The required form of the bound (7.1) reflects the expectation that $\bar{f}_n$ will not deviate too much from $f_n(\mu)$. For slowly varying functions f, say for $f(t)=1/(1+t)$, this surely will be true, especially when z is large. However, the scale functions ψ_n, χ_n are constructed in terms of the global estimate $M(\mu)$, introduced in (4.4). Consequently, the asymptotic scales used in the theorems may be too rough to describe what is really happening in the asymptotic expansions.

To show this by way of a simple example, we consider $f(t)=\exp[\mu/(1+t)]$. It is easily verified that is satisfies the conditions of § 4; $\Omega=\mathbb{C}\backslash\{-1\}$, Ω_0 is the half-plane $\operatorname{Re} t\geqq-\delta$, where $0<\delta<1$. In (4.4), $p=0$ and $M(\mu)=\exp[\mu/(1-\delta)]$, which is exponentially large, when μ is large. However, we expect that the remainder $\bar{f}_n$ in the expansion (3.1) is comparable with $f_n(\mu)$, which is only algebraic in μ. Therefore, the theorems of the previous section are applicable, but the chosen asymptotic scales are not able to control the remainders of the expansion in a realistic way. This is especially true for expansion (6.2); for (6.9), which is more global in character due to the second uniformity parameter, the chosen scale may be more suitable. We want to emphasize that in this example only the scales default, whereas the expansions themselves are appropriate and may be of interest. The above noticed imperfections (see also Remark 6.3) are inherent in the definition of generalized asymptotic expansions. We have chosen this framework in order to be able to describe precisely the propositions and what we want to prove. We consider this important in asymptotics, especially when one or more uniformity parameters are involved. On the other hand, we have the need of constructing sharp error bounds for the remainders in the expansions, so that we can interpret the expansions in a realistic way. An ideal procedure would be a combination of both approaches, but in this stage, for lack of a unified approach, we prefer to discuss them separately.

7.1. Error bounds for (3.1). As remarked earlier, the quantities $f_s(\mu)$ and $B_s(\alpha)$ may be grossly overestimated by the global upper bound $M(\mu)$ introduced in (4.4). A better approach, say for arriving at (7.1), seems to be to give a sharp estimate for $f_n(t)$ near $t=\mu$, whereas the estimate "far away" of this saddle point may be rather crude. To be more explicit, we need a *comparison function* $w(t,\mu)$, $w:\mathbb{R}^+\times\mathbb{R}^+\to[1,\infty)$, that satisfies the condition $w(\mu,\mu)=1$, and that may be large outside an interval around $t=\mu$. We suppose that we can assign quantities M_n, which may depend on μ and which are strictly larger than unity:

$$M_n \geqq 1+\varepsilon_n, \quad \varepsilon_n \text{ fixed and positive,} \tag{7.2}$$

such that for all $t\geqq 0$ we have

$$|f_n(t)|\leqq M_n|f_n(\mu)|w(t,\mu). \tag{7.3}$$

Furthermore, we suppose that it is easy to calculate or estimate the integral

$$|\bar{f}_n|\leqq\frac{M_n|f_n(\mu)|z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}w(t,\mu)\,dt, \tag{7.4}$$

obtained from (3.3) by bounding $f_n(t)$ in this way. When $f_n(\mu)$ happens to have zeros on $(0,\infty)$, (7.3) and (7.4) have to be modified, say by replacing $|f_n(\mu)|$ by $\delta_n+|f_n(\mu)|$, $\delta_n>0$.

Since f, and hence all f_n, have algebraic growth on $[0,\infty)$ (see (4.4)), it will be sufficient that $w(t,\mu)=\mathcal{O}[\exp(\sigma t)]$ for some positive σ. This suggests as a possible choice $w(t,\mu)=\cosh[\sigma(t-\mu)]$, which meets all requirements formulated thus far. Substituting this into (7.4), we obtain

$$\begin{aligned}\frac{z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}\cosh[\sigma(t-\mu)]\,dt&=\tfrac{1}{2}[(1-\sigma/z)^\lambda\,e^{-\sigma\mu}+(1+\sigma/z)^\lambda\,e^{\sigma\mu}]\\ &=\mathcal{O}[\cosh(\tfrac{1}{2}\lambda\sigma^2/z^2)],\end{aligned} \tag{7.5}$$

as $z\to\infty$. When $\sigma=\mathcal{O}(1)$ $(\mu\geqq 0)$, this contribution to the right-hand side of (7.4) is quite acceptable, so long as $\lambda=o(z^2)$. But for a uniformity domain $[0,\infty)$ it is unacceptable.

This brings us to a further requirement that

$$\frac{z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt}w(t,\mu)\,dt=\mathcal{O}(1), \tag{7.6}$$

as $z\to\infty$, uniformly with respect to $\mu\in[0,\infty)$.

For several reasons the following comparison function is very convenient:

$$w_\sigma(t,\mu)=[(t/\mu)^{-\mu}\,e^{t-\mu}]^\sigma, \qquad \sigma\geqq 0, \tag{7.7}$$

where σ may depend on μ, but not on t. For $\mu=0$ we define $w_\sigma(t,0)=\exp(\sigma t)$. This choice fits better in the dominant part $t^\lambda\,e^{-zt}$ of (7.4) than the cosh-function tried before.

For (7.3) we write

$$|f_n(t)|\leqq M_n|f_n(\mu)|w_{\sigma_n}(t,\mu), \tag{7.8}$$

and (7.4) becomes

$$(7.9)\qquad |\bar{f}_n| \leqq M_n Q_n |f_n(\mu)|, \qquad Q_n = (1-\sigma_n/z)^{-1/2}\Gamma^*(\lambda - \mu\sigma_n)/\Gamma^*(\lambda),$$

where

$$(7.10)\qquad \Gamma^*(z) = [z/(2\pi)]^{1/2}\, e^z z^{-z}\Gamma(z) = 1 + \mathcal{O}(z^{-1}),$$

as $z \to \infty$. Compared with (7.5), this result is much more acceptable, since now we have

$$(7.11)\qquad Q_n = 1 + \mathcal{O}(z^{-1}) \quad \text{as } z \to \infty,$$

uniformly with respect to λ or μ in $[0, \infty)$. Especially, large values of λ are in favor in (7.11). The only condition is that σ_n in (7.8) is fixed or a bounded function of μ on $[0, \infty)$. In this event $z - \sigma_n$ can be viewed as large parameter in Q_n, although $z > \sigma_n$ is enough.

For the construction of error bounds it is sufficient that $f_n(t)$ is continuous on $[0, \infty)$, i.e., that f belongs to the continuity class $C^{2n}([0, \infty))$ (see (5.4)). A different point is that, as remarked earlier, a slight modification is needed when $f_n(\mu) = 0$. A special case is $\mu = 0$, where $f_n, \bar{f}_n$ vanish for $n \geqq 1$. In that case we can define $\sigma_n = 0$. When we construct error bounds, the assumption (4.4) on the algebraic growth is only needed for $t \geqq 0$. Another assumption on f may be that σ_n of (7.8) is a bounded function of μ. A proper choice of M_n, for instance by making M_n a function of μ, will yield a wide class of admissible functions f. The construction of error bounds is not enough to investigate the nature of the asymptotic expansion. However, when $\sigma_n, M_n, f_n(\mu)/f(\mu)$ $(f(\mu) \neq 0)$ are bounded functions of μ on $[0, \infty)$ for each $n \geqq 0$, then we can use the Poincaré-type scale $\{z^{-s}\}$, and the uniformity with respect to μ in $[0, \infty)$ easily follows.

A possible approach to compute M_n and σ_n of (7.8) is to start with trial values of M_n satisfying (7.2). Then we compute

$$(7.12)\qquad \sigma_n = \sup_{t \geqq 0} \tilde{f}_n(t), \quad \mu \text{ fixed in } [0, \infty),$$

where

$$(7.13)\qquad \tilde{f}_n(t) = \frac{\log |f_n(t)/[M_n f_n(\mu)]|}{t - \mu - \mu \log(t/\mu)}, \qquad t \neq \mu, \quad f_n(\mu) \neq 0.$$

For two examples we have computed σ-values. A third example is considered in § 9 for the incomplete beta function.

Example 7.1. $f(t) = 1/(1+t)$. We have

$$f_2(\mu) = \frac{\mu(\mu-2)}{(1+\mu)^5}, \qquad f_2(t) = \frac{t(\mu t - \mu - 2)}{(1+\mu)^3(1+t)^3}.$$

Since $f_2(\mu)$ vanishes at $\mu = 2$, we replace it by

$$f_2^*(\mu) = \frac{\mu(1 + |\mu - 2|)}{(1+\mu)^5}.$$

We consider three choices of M_2 and we obtain σ_2 via (7.12) for several values of μ. We also show corresponding values of Q_2 of (7.9) for $z = 5$. For larger values of z, Q_2 is closer to unity. The results are shown in Table 7.1. It follows that the remainder $z^{-2}\bar{f}_2$ of (3.3) is rather close to the first neglected term $z^{-2}f_2(\mu)$ for the values of μ and z used in the table. Larger values of μ and z confirm this tendency even better.

TABLE 7.1

μ	$M_2=1.1$		$M_2=1.5$		$M_2=2.0$	
	σ_2	Q_2	σ_2	Q_2	σ_2	Q_2
0	0	1	0	1	0	1
1	−0.64	0.9397	−0.87	0.9210	−0.98	0.9120
5	0.02	1.0023	−0.01	0.9989	−0.04	0.9960
10	0.04	1.0038	0.00	1.0004	−0.01	0.9993
25	0.11	1.0108	0.03	1.0031	0.01	1.0014
50	0.08	1.0079	0.02	1.0024	0.01	1.0014
100	0.05	1.0047	0.01	1.0014	0.00	1.0001

Example 7.2. $f(t)=\exp[\mu/(1+t)]$. We use (7.3), (7.7) with $n=0$, which gives a bound (7.8) for $\bar{f}_0$ of (3.3), i.e., for $z^\lambda F_\lambda(z)$ of (2.1). The results are shown in Table 7.2, again with $z=5$.

TABLE 7.2

μ	$M_0=1.1$		$M_0=1.5$		$M_0=2.0$	
	σ_0	Q_0	σ_0	Q_0	σ_0	Q_0
0	0	1	0	1	0	1
1	0.38	1.0417	0.03	1.0027	−0.02	0.9976
5	0.66	1.0737	0.31	1.0329	0.26	1.0270
10	0.46	1.0501	0.26	1.0276	0.25	1.0259
25	0.21	1.0221	0.21	1.0216	0.21	1.0213
50	0.18	1.0187	0.18	1.0185	0.18	1.0184
100	0.16	1.0162	0.16	1.0161	0.16	1.0161

7.2. Error bounds for (3.2). For the construction of error bounds for the remainder of expansion (3.2), we use as comparison function $w_\sigma(t,\alpha)$, with w_σ defined in (7.7). The comparison function $w_\sigma(t,\mu)$ may yield unrealistic bounds, when α and μ are not of the same size. In Theorem 6.2 we used (6.10) in order to get rid of the factor z in (3.5), (3.6). However, this factor is neutralized by the expression $f_n(t)-\bar{f}_n$ in the integral (the minus-sign is important here). In the following error analysis this expression will not be replaced by $|f_n(t)|+|\bar{f}_n|$.

We write

$$g_n(t)=f_n(t)-\bar{f}_n,\qquad n=0,1,\cdots. \tag{7.14}$$

When $g_n(\alpha)\neq 0$, we estimate g_n as follows:

$$|g_n(t)|\leqq M_n|g_n(\alpha)|w_{\tau_n}(t,\alpha), \tag{7.15}$$

where M_n satisfies (7.2) and w is defined in (7.7). We consider two cases.

(i) $0\leqq\alpha\leqq\mu$. Starting point is (3.6); (7.15) has to be considered for $t\in[0,\alpha]$. We obtain

$$|\bar{B}_n|\leqq zM_n\alpha^{\alpha\tau_n-\lambda}\,e^{\alpha(z-\tau_n)}|g_n(\alpha)|\int_0^\alpha t^{\lambda-\alpha\tau_n-1}\,e^{-(z-\tau_n)t}\,dt.$$

Integration by parts gives

$$|\bar{B}_n|\leqq\frac{zM_n}{\lambda-\alpha\tau_n}|g_n(\alpha)|[1+(z-\tau_n)g(\alpha,\lambda-\alpha\tau_n,z-\tau_n)],$$

where g is defined in (6.5). In Lemma 6.1 it is supposed that in $g(\alpha, \lambda, z)$ the parameters satisfy $0 \leqq \alpha \leqq \lambda/z$. In the present g-function this relation holds as well. The only condition is $z > \tau_n$. For a better representation of the above bound for $\bar{B}_n$ we write

$$\frac{z}{\lambda - \alpha\tau_n}|g_n(\alpha)| = \frac{\lambda}{\lambda - \alpha\tau_n}\frac{|g_n(\alpha)|}{\mu}.$$

The first factor at the right equals $1/(1-\alpha\tau_n/\mu z)$ which is $1+\mathcal{O}(z^{-1})$ as $z \to \infty$, uniformly with respect to α and μ, $0 \leqq \alpha \leqq \mu$. The second factor $|g_n(\alpha)|/\mu$ is properly defined in the limit $\mu \to 0$, as follows from a similar argument as used for (5.8). By using Lemma 6.1 we obtain for $z > \tau_n$

$$|\bar{B}_n| \leqq \frac{\lambda M_n}{\lambda - \alpha\tau_n}\frac{|g_n(\alpha)|}{\mu}\left[1 + \frac{\alpha}{\bar{\mu} - \alpha}\min\{1, \bar{\zeta}\sqrt{\tfrac{1}{2}\pi(z-\tau_n)}\}\right], \tag{7.16}$$

where $\bar{\mu}$ and $\bar{\zeta}$ are defined by

$$\tfrac{1}{2}\bar{\zeta}^2 = \alpha - \bar{\mu} + \bar{\mu}\log(\bar{\mu}/\alpha), \quad \bar{\zeta} \geqq 0, \quad \bar{\mu} = (\lambda - \alpha\tau_n)/(z-\tau_n). \tag{7.17}$$

The conditions $0 \leqq \alpha \leqq \mu$ and $z > \tau_n$ imply $0 \leqq \alpha \leqq \bar{\mu}$.

(ii) $0 \leqq \mu \leqq \alpha$. We consider (7.15) for $t \geqq \alpha$. Representation (3.5) gives for $z > \tau_n$

$$|\bar{B}_n| \leqq zM_n|g_n(\alpha)|G_0(\alpha, \lambda - \alpha\tau_n, z - \tau_n).$$

Using Lemma 6.2, we obtain

$$|\bar{B}_n| \leqq \frac{M_n|g_n(\alpha)|}{(\alpha - \bar{\mu})(1-\tau_n/z)}\min\{1, \bar{\zeta}\sqrt{\tfrac{1}{2}\pi(z-\tau_n)}\}. \tag{7.18}$$

When $\lambda - \alpha\tau_n < 0$ we define $\bar{\zeta} = +\infty$, otherwise it is defined by (7.17), with $0 \leqq \bar{\mu} \leqq \alpha$.

Remark 7.1. The numbers $\bar{\zeta}$, M_n and τ_n in (7.16) and (7.18) need not be the same. When $\alpha \to \bar{\mu}$, $\bar{\zeta}$ has to be combined with the factor $1/(\bar{\mu} - \alpha)$, as explained in Remark 6.1.

Remark 7.2. We can write $g_n(t)$ of (7.14) in the form

$$g_n(t) = \tilde{g}_n(t) - \frac{1}{z}\bar{f}_{n+1}, \qquad \tilde{g}_n(t) = f_n(t) - f_n(\mu). \tag{7.19}$$

Bounds for $\bar{f}_{n+1}$ follow from (7.8). Contributions owing to $\tilde{g}_n(t)$ are as in (7.16), (7.18) with $g_n(\alpha)$ replaced by $\tilde{g}_n(\alpha)$. Observe that (see (2.7))

$$\tilde{g}_n(\alpha)/(\alpha - \mu) = B_n(\alpha),$$

which shows up in (7.18) when we use (7.19). There is something to recommend about the approach based on (7.19). The point is that g_n of (7.14) may be difficult to evaluate without the splitting in (7.19). Furthermore, M_n and τ_n of (7.15) may depend on z. However, this dependence will be very weak when z is large.

8. A loop integral with analogue asymptotic features. In previous papers [11], [12] we stated the analogy between the following integrals:

$$F_\lambda(z) = \frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}f(t)\,dt, \tag{8.1}$$

$$G_\lambda(z) = \frac{\Gamma(\lambda+1)}{2\pi i}\int_{-\infty}^{(0+)} t^{-\lambda-1}e^{zt}f(t)\,dt, \tag{8.2}$$

where (8.1) is the complete integral given in (2.1). The contour in (8.2) starts and ends at $t=-\infty$ (respectively, with $\arg t=-\pi$, $\arg t=\pi$), and encircles the origin in positive direction. The analogy, from an asymptotic point of view, is that of their asymptotic expansions:

$$z^{\lambda}F_{\lambda}(z)\sim\sum_{s=0}^{\infty}f_s(\mu)z^{-s}, \tag{8.3}$$

$$z^{-\lambda}G_{\lambda}(z)\sim\sum_{s=0}^{\infty}(-1)^s f_s(\mu)z^{-s}, \tag{8.4}$$

as $z\to\infty$; (8.3) is considered in the present paper, for instance in (2.4). The construction of (8.4) is based on the same integration by parts procedure, by using Hankel's integral for the reciprocal gamma function. That is, (8.2) reduces to z^{λ}, when f is the identity. Conditions on f, especially the domain of holomorphy, have to be modified, before we can state that (8.2) has (8.4) as a uniform expansion with z as the large parameter and $\mu=\lambda/z$ as uniformity parameter.

It seems that the following integral

$$G_{\lambda}(z,\alpha)=\frac{\Gamma(\lambda+1)}{2\pi i}\int_{-\infty}^{(0^+)}t^{-\lambda}e^{zt}g(t)\frac{dt}{t-\alpha} \tag{8.5}$$

is the relative of $F_{\lambda}(z,\alpha)$ defined in (1.1). That is, (8.5) has four asymptotic phenomena that are in some sense equivalent to the four discussed in § 1 for (1.1). However, the asymptotic expansions show an interesting difference, although the characteristics of both are exactly the same.

8.1. Uniform approximation for loop integrals. We suppose that the contour in (8.5) cuts the positive real axis at the point t_0. We first give (8.5) for $g=1$. Multiplying by $\exp(-\alpha z)$ and differentiating with respect to z, we obtain (8.2) with $f=1$. Integrating with respect to z, and taking into account some limiting values, we obtain two forms for (8.5) with $g=1$:

$$G_{\lambda}(z,\alpha)=\lambda\alpha^{-\lambda}e^{\alpha z}\begin{cases}\gamma(\lambda,\alpha z), & t_0>\alpha,\\ (-1)\Gamma(\lambda,\alpha z), & 0<t_0<\alpha,\end{cases} \tag{8.6}$$

where $\gamma(a,x)$, $\Gamma(a,x)$ are incomplete gamma functions. The transition from one form to the other in (8.6) also follows from (8.5), by shifting the contour across the pole at $t=\alpha$, and using $\gamma(a,x)+\Gamma(a,x)=\Gamma(a)$.

Suppose now that $0<t_0<\alpha$. Writing $g(t)=g(\alpha)+[g(t)-g(\alpha)]$, we obtain for (8.5)

$$G_{\lambda}(z,\alpha)=G_{\lambda}(z)-\lambda g(\alpha)\alpha^{-\lambda}e^{\alpha z}\Gamma(\lambda,\alpha z) \tag{8.7}$$

where

$$G_{\lambda}(z)=\frac{\Gamma(\lambda+1)}{2\pi i}\int_{-\infty}^{(0^+)}t^{-\lambda-1}e^{zt}h(t)\,dt,\qquad h(t)=t\frac{g(t)-g(\alpha)}{t-\alpha}. \tag{8.8}$$

Therefore, assuming appropriate conditions on g, and hence on h, the asymptotic expansion of $G_{\lambda}(z)$ is given in (8.4) with $f_s(\mu)$ replaced by $h_s(\mu)$. The latter are generated as $f_s(\mu)$ in (2.5), with f_0 replaced by h and f_s by h_s.

We conclude that, apart from normalization, (8.7) has with (8.4) a similar expansion as $F_\lambda(z,\alpha)$ in (2.6). An interesting difference is that now the incomplete gamma function does not multiply a full asymptotic expansion but just one term involving $g(\alpha)$. This gives a simpler asymptotic problem. For instance, the construction of error bounds only applies to $G_\lambda(z)$, i.e., for the complete integral (8.2), where f is replaced by h of (8.8).

A representation for the remainders in the expansion of $G_\lambda(z)$ in (8.7) follows by writing

$$z^{-\lambda}G_\lambda(z)=\sum_{s=0}^{n-1}(-1)^s h_s(\mu)z^{-s}+(-1)^n z^{-n}\bar{h}_n, \tag{8.9}$$

$$\bar{h}_n=\frac{z^{-\lambda}\Gamma(\lambda+1)}{2\pi i}\int_{-\infty}^{(0^+)} t^{-\lambda-1}\,e^{zt}h_n(t)\,dt, \tag{8.10}$$

where $h_n(t)$ is generated by the recursion (2.5), with starting function $h_0=h$ defined in (8.8).

8.2. Error bounds for loop integrals. For the construction of error bounds, we select a special contour in (8.10). Writing

$$t^{-\lambda}\,e^{zt}=e^{z[\rho e^{i\theta}-\mu\log\rho-i\mu\theta]},\qquad t=\rho\,e^{i\theta},$$

we see that the imaginary part of the phase function will vanish, when we take

$$\rho=\rho(\theta)=\mu\theta/\sin\theta,\qquad -\pi<\theta<\pi.$$

This defines the path of steepest descent through the saddle point $t=\mu$. Integrating (8.10) with respect to the parameter θ along this contour, we use

$$\frac{1}{t}\frac{dt}{d\theta}=\frac{1}{\rho}\frac{d\rho}{d\theta}+i,$$

giving

$$\bar{h}_n=\frac{\Gamma(\lambda+1)\lambda^{-\lambda}\,e^{\lambda}}{2\pi}\int_{-\pi}^{\pi} e^{-\lambda\phi(\theta)}\tilde{h}_n(t)\,d\theta, \tag{8.11}$$

where

$$t=\rho\,e^{i\theta}=\mu[\theta\operatorname{cotg}\theta+i\theta],$$

$$\tilde{h}_n(t)=\frac{1}{it}\frac{dt}{d\theta}h_n(t), \tag{8.12}$$

$$\phi(\theta)=-\theta\operatorname{cotg}\theta+\log\frac{\theta}{\sin\theta}+1.$$

A bound for $\tilde{h}_n$ is obtained by writing as in (7.8)

$$|\tilde{h}_n(t)|\leqq M_n|h_n(\mu)|\,e^{\mu\delta_n\phi(\theta)}, \tag{8.13}$$

where M_n satisfies (7.2), $\delta_n<z$, and where it is assumed that $h_n(\mu)\neq 0$. We obtain as in (7.9)

$$|\bar{h}_n|\leqq M_n\tilde{Q}_n|h_n(\mu)|, \tag{8.14}$$

where $\tilde{Q}_n=1/[(1-\delta_n/z)Q_n]$. We have used that (8.10) reduces to unity when h_n equals unity.

Observe that to compute error bounds for the expansion (8.9), exactly the same comparison function is used as in (7.8); the forms are different owing to the mapping $t \to \theta$.

Remark 8.1. Representation (8.7) is obtained under the condition $0 < t_0 < \alpha$, where t_0 is the point where the contour of (8.5) cuts the real positive axis. When (8.5) is presented with $t_0 > \alpha$, $G_\lambda(z, \alpha)$ has representation (8.7), with $\Gamma(\lambda, \alpha z)$ replaced by $-\gamma(\lambda, \alpha z)$. This complementary relation is of the same kind as for the real integrals described in (2.14), (2.15).

9. The incomplete beta function. We use the incomplete beta function in the notation

$$I_x(p, q) = \frac{1}{B(p, q)} \int_0^x \tau^{p-1}(1-\tau)^{q-1}\, d\tau, \tag{9.1}$$

where $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ is the complete beta function. The asymptotic problem is to give an expansion of $I_x(p, q)$ with p as large parameter and $x \in [0, 1]$ and $\mu = q/p \in [0, \infty)$ as uniformity parameters. We can use

$$I_x(p, q) = 1 - I_{1-x}(q, p) \tag{9.2}$$

to interchange the role of p and q. For information on $I_x(p, q)$ we refer to [1, p. 944]. In [10] we considered the asymptotic problem for $I_x(p, q)$ for more restricted ranges of the parameters. We believe that the expansions of this section are new in the sense that the uniformity domain of μ or q is the complete interval $[0, \infty)$. Earlier results prescribed q to belong to a compact subset of $(0, \infty)$ (case (ii) of § 1), or p/q to a compact subset of $(0, \infty)$ (case (iv)). Extension to complex values of the parameters is possible, but will not be considered here.

To describe the asymptotic features of (9.1) in more detail, we compute the saddle point of $\tau^p(1-\tau)^q$. It occurs at

$$\tau_0 = \frac{p}{p+q}. \tag{9.3}$$

When $p+q$ is large, the value $I_x(p, q)$ is very small when $x < \tau_0$, and it is close to unity when $x > \tau_0$. When τ_0 is restricted to a compact subset of $(0, 1)$, this transition can properly be described by an error function (normal distribution function); when $\tau_0 \to 1$ the basic approximant is an incomplete gamma function. We will show that this function can handle the complete uniformity domain for q, i.e., $[0, \infty)$. It is essential to transform (9.1) to the standard form (1.1), by means of a rather complicated transformation.

9.1. Transformation to standard form. A first transformation $\tau \to e^{-\tau}$ gives

$$I_x(p, q) = \frac{1}{B(p, q)} \int_{-\log x}^{\infty} (1-e^{-\tau})^{q-1} e^{-p\tau}\, d\tau. \tag{9.4}$$

Comparing this with (1.1), we observe that it has the standard form when $\lambda = q$, $z = p$ and $f(t) = [(1-e^{-\tau})/\tau]^{q-1}$. However, for several reasons this choice of f will not give a uniform expansion for the q-interval $[0, \infty)$. One reason is that large values of q will have much influence on coefficients $f_s(\mu)$, $B_s(\alpha)$. Observe that for f dependence only on μ is assumed in § 4, and not on λ.

A better way for transforming (9.4) in (1.1) is to use the mapping $\tau \to t(\tau)$ defined by

$$\tau - \mu \log(1-e^{-\tau}) = t - \mu \log t + A(\mu), \tag{9.5}$$

where $\mu = q/p$. The left-hand side has a saddle point at

$$\tau_1 = \log(\mu+1) = -\log \tau_0, \tag{9.6}$$

the right-hand side at $t=\mu$. To make the mapping properly defined we require the correspondences

$$\tau = 0 \leftrightarrow t = 0, \qquad \tau = \tau_1 \leftrightarrow t = \mu, \qquad \tau = +\infty \leftrightarrow t = +\infty. \tag{9.7}$$

The middle one gives

$$A(\mu) = (1+\mu)\log(1+\mu) - \mu. \tag{9.8}$$

The point $\tau = -\log x$, the lower end point of integration in (9.4), is mapped to $t = \alpha$, which is defined by the implicit relation

$$-\log x - \mu \log(1-x) = \alpha - \mu \log \alpha + A(\mu), \tag{9.9}$$

with corresponding points

$$x = 0 \leftrightarrow \alpha = +\infty, \qquad x = \tau_0 \leftrightarrow \alpha = \mu, \qquad x = 1 \leftrightarrow \alpha = 0. \tag{9.10}$$

Observe that the middle one satisfies (9.9) due to the choice (9.8). In fact, the mappings (9.5) and (9.9) are the same, up to parametrization.

The transformed version of (9.4) is

$$I_x(p,q) = \frac{e^{-pA(\mu)}}{B(p,q)} \int_\alpha^\infty t^{q-1} e^{-pt} f(t)\, dt, \tag{9.11}$$

where

$$f(t) = \frac{t}{1-e^{-\tau}} \frac{d\tau}{dt} = \frac{t-\mu}{1-(1+\mu)\, e^{-\tau}}. \tag{9.12}$$

The regularity of the transformation (9.5), and that of f, is extensively discussed in [12, § 4]. From that analysis it follows that f satisfies the conditions of § 4. In (4.1) we have to take $\kappa = \frac{1}{2}$, and d_0 and δ both somewhat less than $\sqrt{2\pi}$. Ω_0 is a parabola-shaped domain, and for the number p in (4.4) we take $p = 1$ (which follows easily from (9.12)). The function f is positive on $[0, \infty)$; $f(0) = 1$, $f(\mu) = \sqrt{1+\mu}$. We verified numerically that

$$\sup_{\substack{t \geqq 0 \\ \mu \geqq 0}} \frac{f(t)}{1+t} = 1.$$

Therefore, $M(\mu)$ of (4.4) will not deviate very much from unity, especially when δ and d_0 are small.

9.2. Uniform expansion of incomplete beta function. In the notation of (1.1), (2.1), (2.9), we can write

$$\begin{aligned} I_x(p,q) &= \frac{e^{-pA(\mu)}\Gamma(p+q)}{\Gamma(p)} F_q(p,\alpha), \\ F_q(p,\alpha) &= Q(q,\alpha p) F_q(p) + \frac{\alpha^q\, e^{-\alpha p}}{p\Gamma(q)} B_q(p,\alpha). \end{aligned} \tag{9.13}$$

However, the "complete" integral $F_q(p)$ can be written in this case in terms of known functions. Since $I_1(p,q) = 1$, we have

$$F_q(p) = e^{pA(\mu)}\Gamma(p)/\Gamma(p+q) = \left(\frac{p+q}{p}\right)^{p+q} e^{-q}\Gamma(p)/\Gamma(p+q). \tag{9.14}$$

Although it is possible to give for $F_q(p)$ an expansion as in (2.4) (see [12, § 4] for a related expansion), it is more attractive now to write (9.13) in the form

$$I_x(p,q)=Q(q,\alpha p)+\frac{\alpha^q\,e^{-\alpha p+q}}{pB(p,q)}\left(\frac{p}{p+q}\right)^{p+q}B_q(p,\alpha). \tag{9.15}$$

By using (9.9), and (2.11) we can write

$$I_x(p,q)=Q(q,\alpha p)+\frac{x^p(1-x)^q}{pB(p,q)}B_q(p,\alpha),\qquad B_q(p,\alpha)\sim\sum_{s=0}^{\infty}\frac{B_s(\alpha)}{p^s}\quad \text{as } p\to\infty. \tag{9.16}$$

It follows that we have only to consider the asymptotic expansion for $B_q(p,\alpha)$, which, however, is not the simpler one of (2.4), (2.11). The first coefficient is

$$B_0(\alpha)=\frac{f(\alpha)-f(\mu)}{\alpha-\mu}, \tag{9.17}$$

with

$$f(\alpha)=\frac{\alpha-\mu}{1-(1+\mu)x},\qquad f(\mu)=\sqrt{1+\mu}.$$

Special values are

$$B_0(0)=\frac{\sqrt{1+\mu}-1}{\mu},\qquad B_0(\mu)=\lim_{\alpha\to\mu}f'(\alpha)=\frac{\mu-1+\sqrt{1+\mu}}{3\mu},\qquad B_0(\infty)=1, \tag{9.18}$$

and they satisfy $0\leqq B_0(0)\leqq B_0(\mu)\leqq B_0(\infty)$, for $\mu\geqq 0$.

In fact, all $B_s(\alpha)$ can be expressed in terms of α, x and μ, and those three are related by (9.9). When $\alpha\neq\mu$, an explicit representation of $B_s(\alpha)$ in terms of say x and $\mu=q/p$ is not possible, since (9.9) cannot be solved explicitly for α. When $\mu=0$, (9.5) reduces to the identity mapping and f of (9.12) becomes $t/[1-\exp(-t)]$. The latter has singularities at $t=\pm 2\pi i, \pm 4\pi i,\cdots$. When $\mu>0$, singular points of f originate from these points, of which $\pm 2\pi i$ are most important. The singular points of f starting from $\pm 2\pi i$ $(\mu=0)$ are located in the half-plane $\operatorname{Re} t>0$. For large values of μ they are approximately near $\mu+2\sqrt{\pi\mu}\exp(\pm i\pi/4)$. The value $\kappa=\frac{1}{2}$ in (4.1) comes from $\sqrt{\mu}$ in this asymptotic value. The coefficients $B_s(\alpha)$ have the same domain of regularity as $f(\alpha)$. Since $M(\mu)=\mathcal{O}(1)$, $p=1$, $\kappa=\frac{1}{2}$, $\bar\kappa=0$, we have for (5.7)

$$B_s(\alpha)=\mathcal{O}(1+\mu+\alpha)^{1/2},\quad s=0,1,2,\cdots,\quad \mu\geqq 0,\quad \alpha\geqq 0. \tag{9.19}$$

This estimate gives a good impression of the asymptotic nature of the expansion in (9.16), although the estimate for $s=0$ seems to be too large (cf. (9.18)).

9.3. Uniform expansion based on a loop integral. An interesting variant of (9.16) is obtained by using a contour integral for $I_x(p,q)$ in the complex plane, and by applying the method of § 8. Consider the integral

$$I=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}t^{-p}(1-t)^{-q}\frac{dt}{t-x},\qquad 0<c<1,$$

with $p+q>0$ and $0<x<c$. When $t\in(0,1)$ the phases of t, $(1-t)$ and of the multi-valued functions are zero. By deforming the contour around the negative axis, we obtain for $p<1$

$$\begin{aligned}I&=x^{-p}(1-x)^{-q}-\frac{\sin\pi p}{\pi}\int_0^\infty\tau^{-p}(1+\tau)^{-q}\frac{d\tau}{\tau+x}\\&=x^{-p}(1-x)^{-q}I_x(p,q),\end{aligned}$$

where we used formulas 6.1.17, 15.3.1, 15.3.3, 15.3.4, 26.5.23 and 26.5.2 of [1]. It follows that

$$(9.20)\qquad I_x(p,q)=\frac{x^p(1-x)^q}{2\pi i}\int_{c-i\infty}^{c+i\infty}\tau^{-p}(1-\tau)^{-q}\frac{d\tau}{\tau-x},\qquad 0<c<1,$$

with $0<x<c$. The restriction $p<1$, which is needed to evaluate the contour integral for I, can be dropped by using the principle of analytic continuation. We still need the condition $p+q>0$ for convergence at infinity.

Representation (9.20) is the analogue of (9.1). To obtain standard form (8.5), we use the transformation $\tau\to\exp(-\tau)$. This gives

$$(9.21)\qquad I_x(p,q)=\frac{x^p(1-x)^q}{2\pi i}\int_{-\infty}^{(0^+)}e^{pt}(1-e^{-\tau})^q\frac{d\tau}{1-xe^{\tau}},$$

which is the analogue of (9.4). The contour cuts the positive real τ-axis at a point τ_0 such that $\tau_0<-\log x$. A final transformation

$$(9.22)\qquad \tau-\mu\log(1-e^{-\tau})=t-\mu\log t+A(\mu),$$

where $\mu=q/p$ and $A(\mu)$ is given in (9.8), gives

$$(9.23)\qquad I_x(p,q)=x^p(1-x)^q e^{-q}(1+q/p)^{p+q}\frac{1}{2\pi i}\int_{-\infty}^{(0^+)}e^{pt}t^{-q}\frac{g(t)}{t-\alpha}\,dt,$$

where α is defined in (9.9), and

$$(9.24)\qquad g(t)=\frac{t-\alpha}{1-xe^{\tau}}\frac{d\tau}{dt}.$$

A relation for $d\tau/dt$ is given in (9.12). The contour in (9.23) cuts the positive t-axis at a point t_0 satisfying $t_0<\alpha$. Splitting off the pole, we obtain for (9.23)

$$(9.25)\qquad I_x(p,q)=Q(q,\alpha p)+x^p(1-x)^q e^{-q}(1+q/p)^{p+q}G_q(p)/\Gamma(q+1),$$

with $G_q(p)$ as in (8.8). Here we used $g(\alpha)=-1$, which follows from (9.24) by l'Hôpital's rule (observe that $t=\alpha$ corresponds with $\tau=-\log x$ in (9.22)).

Observe that the transformations (9.5) and (9.22) are exactly the same. The real corresponding points in (9.7) determine the mapping in the complex plane. Therefore, no new correspondences have to be defined for (9.22). In fact, the mapping has a very global character. As remarked earlier, we have investigated the mapping (9.5) in [12], with emphasis on what is happening in a neighbourhood of $\mathbb{R}^+$. However, the domain of regularity extends to the full half-plane Re $t\leqq 0$. To understand the mapping in the complex plane, it is instructive to see that the contours of steepest descent in (9.21), (9.23) are mapped onto each other. On these contours the imaginary parts in the left- and right-hand side of (9.22) vanish.

Comparing (9.15) and (9.25), we conclude that the function $B_q(p,\alpha)$ has (in the special case of this section) an asymptotic expansion which corresponds to that of a "complete" integral (8.2). We have

$$(9.26)\qquad B_q(p,\alpha)=\frac{\Gamma(p+1)\,e^{-q}(1+q/p)^{p+q}}{q\Gamma(p+q)}G_q(p),$$

where $G_q(p)$ is given in (8.8), with expansion as in (8.9); g is defined in (9.24).

By using (7.10) and (8.9) we can write (9.26) in the form

$$(9.27)\quad B_q(p,\alpha)=\frac{\sqrt{\mu+1}}{\mu}\frac{\Gamma^*(p)}{\Gamma^*(p+q)}p^{-q}G_q(p),\qquad p^{-q}G_q(p)\sim\sum_{s=0}^{\infty}(-1)^s h_s(\mu)p^{-s}.$$

Therefore, the approach based on a complex contour gives for $I_x(p, q)$ a simpler asymptotic expansion and simpler error bounds than the approach based on, say, (9.11). The computation of the coefficients $h_s(\mu)$ is not a simple problem. Also, the bound for $|h_n(t)|$ in (8.13) has to be computed on a contour in the complex plane. But the form of the error bound (8.14) is much simpler than those obtained in § 7.

To show some of the steps needed to evaluate the coefficients $h_s(\mu)$ in (9.27), we compute $h_0(\mu)$ and its limiting form for $\alpha \to \mu$. We have, using h of (8.8) and g of (9.24),

$$g(\mu)=\frac{\mu-\alpha}{1-(1+\mu)x}\frac{d\tau}{dt}, \tag{9.28}$$

where the derivative is evaluated at $t=\mu$. From (9.12), we have

$$\lim_{t\to\mu}\frac{d\tau}{dt}=\frac{1}{\mu+1}\lim_{t\to\mu}\frac{t-\mu}{1-(1+\mu)\,e^{-\tau}}=\frac{1}{\mu+1}\frac{dt}{d\tau},$$

by using l'Hôpital's rule. Hence

$$\frac{d\tau}{dt}=\frac{1}{\sqrt{\mu+1}}\quad \text{at } t=\mu.$$

The square root has a + sign, since τ is an increasing function of t on $[0,\infty)$. Earlier we computed $g(\alpha)=-1$. So we have

$$h_0(\mu)=\frac{\mu}{\mu-\alpha}\left[\frac{(\mu-\alpha)/\sqrt{\mu+1}}{1-(\mu+1)x}+1\right],\qquad \mu\neq\alpha. \tag{9.29}$$

To evaluate this at $\alpha=\mu$, we have to investigate the relation between x and α in more detail. From (9.9) it follows that

$$\alpha[(\mu+1)x-1]\frac{dx}{d\alpha}=x(1-x)(\alpha-\mu).$$

Substituting the expansion

$$x=\frac{1}{\mu+1}+c_1(\alpha-\mu)+c_2(\alpha-\mu)^2+\cdots,$$

we obtain

$$c_1=-(1+\mu)^{-3/2},\qquad c_1c_2=\frac{1}{3\mu(1+\mu)^3}\left[\frac{1-\mu}{\sqrt{1+\mu}}-1\right].$$

When $\alpha\to\mu$, $h_0(\mu)$ of (9.28) has the expansion

$$h_0(\mu)=-\mu c_2/c_1+\mathcal{O}(\mu-\alpha).$$

Hence

$$\lim_{\alpha\to\mu}h_0(\mu)=\frac{1}{3}\left[1-\frac{1-\mu}{\sqrt{\mu+1}}\right].$$

It follows that $B_0(\alpha)$ of (9.17) and $h_0(\mu)$ of (9.29) are related by

$$B_0(\alpha)=\frac{\sqrt{\mu+1}}{\mu}h_0(\mu).$$

A relation between the higher coefficients is obtained as follows. Comparing (9.16) with (9.27), we arrive at the formal identity

$$\sum_{s=0}^{\infty} B_s(\alpha)p^{-s} = \frac{\sqrt{\mu+1}}{\mu}\frac{\Gamma^*(p)}{\Gamma^*(p+q)}\sum_{s=0}^{\infty}(-1)^s h_s(\mu)p^{-s}.$$

We can expand

$$\frac{\Gamma^*(p)}{\Gamma^*(p+q)} \sim \sum_{s=0}^{\infty} d_s(\mu)p^{-s}, \qquad d_0(\mu)=1,$$

by dividing the two expansions for $\Gamma^*(p)$ and $\Gamma^*(p(1+\mu))$ (see also [12, § 4]). Thus we obtain

$$B_s(\alpha) = \frac{\sqrt{\mu+1}}{\mu}\sum_{r=0}^{s}(-1)^r h_r(\mu)d_{s-r}(\mu),$$

which is (9.30) for $s=0$.

9.4. Some numerical results. We used the method of § 7.2 to compute an upper bound for $B_q(p, \alpha)$. That is, we computed M_0 and τ_0 for (7.16), (7.18), and we evaluated the expression at the right-hand sides of these inequalities. For $z = p = 10\ (10)\ 100$ these expressions were evaluated at the x, μ-grid with $x = 0.05\ (0.05)\ 0.95$, $\mu = 1\ (1)\ 10$. For each z the maximal value occurred at $x = 0.05$, $\mu = 1$. The corresponding α-value for this x, μ-pair is $4.06254 \cdots$. The upper bounds show an interesting feature. For $z = 10$ it is 0.97111 and it steadily increases to 0.97371 for $z = 100$. The ratio of this last upper bound and $B_0(\alpha)$ of (9.17) $(=0.64933\cdots)$ equals $1.500\cdots$, which is the number M_0 that we used in (7.15), (7.16) and (7.18). We observe that the computed upper bounds are slightly less than $M_0B_0(\alpha)$, and tend to this value when z increases. The choice $M_0 = 1.1$ showed the same features: the upper bound of $B_q(p, \alpha)$ is slightly less than $M_0B_0(\alpha)$.

The numerical experiments yield the following conclusion:

$$\sup B_q(p, \alpha) = 1, \quad p \text{ fixed},$$

where the supremum is taken over $x \in [0, 1]$ (or $\alpha \geqq 0$) and $\mu = q/p \in [0, \infty)$. The maximum is assumed at $x = 0$ $(\alpha = \infty)$ and $\mu = 0$. See also (5.11); for $n = 0$ this limit L equals 1. Incidentally, (5.9) gives for $n = 0$

$$B_q(p, \alpha) = p[\log p - \psi(p)] \tag{9.30}$$

at $\alpha = 0$, $\mu = 0$, where $\psi(p)$ is the logarithmic derivative of the gamma function. This follows from well-known representations of this function, and from the fact that f of (9.12) equals $t/(1-\exp(-t))$ at $\mu = 0$. From (9.13) and the asymptotic expansion of the ψ-function, it follows that

$$B_0(p, 0) = \tfrac{1}{2} + \mathcal{O}(p^{-1}) \quad \text{as } p \to \infty.$$

We conjecture that $\frac{1}{2} \leqq B_q(p, \alpha) \leqq 1$ for $\alpha \in [0, \infty)$ (or $x \in [0, 1]$), and $\mu = q/p \in [0, \infty)$, and for all p sufficiently large (say $p \geqq 10$).

REFERENCES

[1] M. A. Abramowitz and I. A. Stegun, *Handbook of mathematical functions*, National Bureau of Standards Applied Mathematics Series 55, Washington D.C., 1964.

[2] N. Bleistein, *Uniform asymptotic expansion of integrals with stationary point near algebraic singularity*, Comm. Pure Appl. Math., 19 (1970), pp. 353-370.

[3] A. Erdélyi and M. Wyman, *The asymptotic evaluation of certain integrals*, Arch. Rational Mech. Anal., 14 (1963), pp. 217-260.

[4] A. Erdélyi, *Asymptotic evaluation of integrals involving a fractional derivative*, this Journal, 14 (1974), pp. 159-171.

[5] W. Gautschi, *A computational procedure for the incomplete gamma functions*, ACM Trans. Math. Software, 5 (1979), pp. 466-481.

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

[7] ———, *Unsolved problems in the asymptotic estimation of special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 99-142.

[8] K. Soni, *A note on uniform asymptotic expansions of incomplete Laplace integrals*, this Journal, 14 (1983), pp. 1015-1018.

[9] N. M. Temme, *Remarks on a paper of A. Erdélyi*, this Journal, 7 (1976), pp. 767-770.

[10] ———, *The uniform asymptotic expansion of a class of integrals related to cumulative distribution functions*, this Journal, 13 (1982), pp. 239-253.

[11] ———, *Uniform asymptotic expansions of Laplace integrals*, Analysis, 3 (1983), pp. 221-249.

[12] ———, *Laplace type integrals: transformation to standard form and uniform asymptotic expansions*, Quart. Appl. Math., 43 (1985), pp. 103-123.

[13] R. Wong, *On uniform asymptotic expansion of definite integrals*, J. Approx. Theory, 7 (1973), pp. 76-86.

[14] A. S. Zil'bergleit, *Uniform asymptotic expansions of some definite integrals*, U.S.S.R. Comput. Math. and Math. Phys., 16 (1977), pp. 36-44.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

012

SOME PROPERTIES OF THE ZEROS OF POLYNOMIAL SOLUTIONS OF STURM-LIOUVILLE EQUATIONS*

R. G. CAMPOS†

Abstract. In this paper certain equalities holding for the zeros of a polynomial satisfying a linear differential equation of the second order are obtained in terms of the coefficients of this equation using an elementary approach. They are applied to the polynomial solutions of a Sturm-Liouville equation. As an example, a new expression for the zeros of the classical polynomials is presented.

Key words. differential equations, polynomials, zeros

AMS(MOS) subject classifications. 65(01)H05, 30(01)A08

1. Introduction. In the past few years, a number of remarkable properties of the zeros of the classical polynomials have been obtained mainly as byproducts of the study of certain integrable many-body problems (Ahmed et al. [2] and Calogero [3]). The purpose of this paper is to present certain results of this kind obtained also as a spin-off, but in this case of the developing of a technique to compute the eigenvalues of the Schrödinger Equation (Campos [5]).

The algebraic relations between the zeros of a polynomial that satisfies an ordinary differential equation (ODE) of the second order and the coefficients of this equation are obtained through an elementary method in § 2. These results are applied to a Sturm-Liouville (SL) equation to illustrate some interesting features of its polynomial solutions. An example is given in § 3 and some final remarks are made in § 4.

2. Properties of the zeros. There are various methods for studying the zeros of polynomials. The method that will be followed in this section is based on an approach due to Laguerre (see Szëgo [9, p. 117]).

We begin by defining

$$\sum_{k=1}^{N}{}' (x_i - x_k)^{-l} \equiv S_N^l(x_i), \qquad l=1,2,\cdots$$

where the prime appended to the sum indicates the omission of the singular term and the points x_k, $k=1,2,\cdots,N$, are the N (complex) zeros of a polynomial $f(x)$ of degree N that satisfies the ODE

$$f''(x)+p(x)f'(x)+q(x)f(x)=0. \tag{1}$$

Then we have the following proposition.

PROPOSITION 1. *If $p(x)$ and $q(x)$ are analytic at x_k, $k=1,2,\cdots,N$, then*

$$S_N^1(x_i) = -\tfrac{1}{2}p(x_i), \tag{2}$$

$$S_N^2(x_i) = \tfrac{1}{3}p'(x_i)+\tfrac{1}{3}q(x_i)-\tfrac{1}{12}[p(x_i)]^2, \tag{3}$$

$$\begin{aligned} S_N^3(x_i) &= \tfrac{1}{8}p(x_i)p'(x_i)-\tfrac{1}{8}p''(x_i)-\tfrac{1}{4}q'(x_i) \\ &= \tfrac{1}{8}\{\tfrac{1}{2}[p(x_i)]^2-p'(x_i)-2q(x_i)\}', \end{aligned} \tag{4}$$

$$\begin{aligned} S_N^4(x_i) = {} & \tfrac{1}{720}[p(x_i)]^4+\tfrac{1}{180}p'(x_i)[p(x_i)]^2-\tfrac{1}{90}q(x_i)[p(x_i)]^2 \\ & -\tfrac{2}{45}[p'(x_i)]^2+\tfrac{1}{45}[q(x_i)]^2-\tfrac{1}{45}p'(x_i)q(x_i) \\ & -\tfrac{1}{20}p''(x_i)p(x_i)+\tfrac{1}{30}p'''(x_i)+\tfrac{1}{10}q''(x_i). \end{aligned} \tag{5}$$

Proof. The idea that we will follow is very simple and it requires only some algebra. A derivative in this proof will be denoted by primes or numbers enclosed by parentheses

* Received by the editors June 18, 1986; accepted for publication (in revised form) December 15, 1986.

† Escuela de Física y Matemáticas, Universidad Michoacana, 58000 Morelia, Michoacán, México.

appended to the function. For the sake of simplicity we will prove only (3). Equations (2), (4) and (5) can be proved along these lines.

Let $f(x)=\prod_{k=1}^N (x-x_k)$ be a solution of (1). Let us define

$$g(x)=(x-x_1)(x-x_2)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_N).$$

Thus,

$$\sum_{\substack{k=1\\k\neq i}}^{N}(x-x_k)^{-2}=\left[\frac{g'(x)}{g(x)}\right]^2-\frac{g''(x)}{g(x)}. \tag{6}$$

Since $p(x)$ and $q(x)$ are analytic at x_i for all x_i, these points are all distinct and $g^{(l)}(x_i)=(l+1)^{-1}f^{(l+1)}(x_i)$ for $l=0,1,\cdots$. Therefore, (6) becomes

$$S_N^2(x_i)=\frac{1}{4}\left[\frac{f''(x_i)}{f'(x_i)}\right]^2-\frac{1}{3}\frac{f'''(x_i)}{f'(x_i)}.$$

Now, using (1), we get (3).

It should be noted that (4) can be interpreted as an extremal property of these zeros: the function of the N continuous variables $z_1, z_2, \cdots, z_N$,

$$H(z_1,z_2,\cdots,z_N)\equiv H(z)=\sum_{\substack{i,j=1\\i\neq j}}^{N}\frac{1}{(z_i-z_j)^2}+\sum_{j=1}^{N}u(z_j)$$

where

$$u(z)=\tfrac{1}{4}[p(z)]^2-\tfrac{1}{2}p'(z)-q(z)+\lambda',$$

and λ' is an arbitrary constant, takes a stationary value at $z_i=x_i$, $i=1,2,\cdots,N$, whenever z_i is in the domain in which the polynomial $f(x)$ satisfies (1). Furthermore, it is not difficult to see that if $u''(x_i)>0$ for all x_i, the function $H(z)$ has an absolute minimum. On the other hand, if $u(x)$ is a symmetric function, the zeros are symmetrically located about the origin.

If an SL equation, like Schrödinger's, is considered, the function $H(x)$ can be interpreted in some cases as the trace of certain finite representation of the Hamiltonian operator with potential $u(x)$ (Campos [5]).

Some interesting features of relations (2)-(5) can be made apparent if we consider zeros of polynomial solutions of an SL equation. The following corollary is an immediate application of Proposition 1.

COROLLARY 1. *Consider the normal* SL *equation*[1]

$$\psi''(x)+[\lambda-V(x)]\psi(x)=0$$

with a solution in the form $\psi(x)=F(x)f(x)$, where $f(x)$ is a polynomial of degree N with zeros x_i, $i=1,2,\cdots,N$, and $F(x_i)\neq 0$ for all x_i. Then, the relation

$$S_N^1(x_i)=-\frac{F'(x_i)}{F(x_i)}$$

holds for all x_i and the eigenvalues are related with the zeros through

$$\lambda=3S_N^2(x_i)-3\frac{F''(x_i)}{F(x_i)}+3\left[\frac{F'(x_i)}{F(x_i)}\right]^2+V(x_i),\qquad i=1,2,\cdots,N. \tag{7}$$

[1] The general SL differential equation can be considered along the same lines.

Moreover, the function

$$(8)\qquad H(z)=\sum_{\substack{i,j\\ i\neq j}}^{N}\frac{1}{(z_i-z_j)^2}+\sum_{i=1}^{N}u(z_i)$$

with

$$u(x)=V(x)-2\frac{F''(x)}{F(x)}+2\left[\frac{F'(x)}{F(x)}\right]^2+(\lambda'-\lambda),$$

where λ' is an arbitrary constant, takes a stationary value in $z_i=x_i$, $i=1,2,\cdots,N$. This value is an absolute minimum if $u''(x_i)>0$ for all x_i. If $u(x)$ is a symmetric function, the zeros are symmetrically located about the origin. Furthermore, they satisfy

$$S_N^4(x_i)=\frac{1}{15}\frac{F^{(IV)}(x_i)}{F(x_i)}-\frac{7}{15}\frac{F'''(x_i)F'(x_i)}{[F(x_i)]^2}+\frac{9}{5}\frac{F''(x_i)[F'(x_i)]^2}{[F(x_i)]^3}$$
$$-\frac{2}{5}\left[\frac{F''(x_i)}{F(x_i)}\right]^2-\frac{7}{9}\left[\frac{F'(x_i)}{F(x_i)}\right]^4-\frac{4}{45}[\lambda-V(x_i)]\left[\frac{F'(x_i)}{F(x_i)}\right]^2+[\lambda-V(x_i)]^2.$$

3. Examples. In this section we consider the zeros of the Hermite, Laguerre and Jacobi polynomials; we use the definitions of these functions given in [7]. The relations obtained through (2)-(4) ((2)-(5)) for Laguerre and Jacobi (Hermite) zeros are already known (see [1] and [2]). Of these equations, only (3) and (4) will be rewritten in this section using the expressions of Corollary 1 for the three cases. The relations corresponding to (5) is given only for Laguerre and Jacobi zeros.

A. Hermite zeros. The differential equation for the Hermite functions $\psi(x)=\exp(-x^2/2)H_N(x)$, is the Schrödinger equation for the harmonic potential $V(x)=(1/2)x^2$:

$$\psi''(x)+(\lambda-x^2)\psi(x)=0,\qquad \lambda=2N+1.$$

If x_k, $k=1,2,\cdots,N$, are the N zeros of $H_N(x)$, then (7) becomes

$$2(N-1)=3S_N^2(x_i)+x_i^2,\qquad i=1,2,\cdots,N,$$

and the function given by (8), with $u(z)=(1/2)z^2$, and $-\infty<z_i<\infty$ for all z_i, has an absolute minimum $H(x)=N(N-1)/2$ at these points (Campos [6]) and, additionally, they are symmetrically located about the origin, as is well known.

B. Laguerre zeros. The differential equation for the Laguerre functions

$$\psi(x)=\exp(-x/2)x^{(\alpha+1)/2}L_N^\alpha(x),$$

where $\alpha>-1$, $x>0$ and $L_N^\alpha(x)$ is the associated Laguerre polynomial, is the Schrödinger-type equation for the potential $V(\alpha,x)=-(2N+\alpha+1)/2x+(\alpha^2-1)/4x^2$:

$$\psi''(x)+\left(\frac{2N+\alpha+1}{2x}+\frac{1-\alpha^2}{4x^2}+\lambda\right)\psi(x)=0,\qquad \lambda=-\frac{1}{4}.$$

Let x_k, $k=1,2,\cdots,N$, be the N zeros of $L_N^\alpha(x)$. Then (7) becomes

$$-\frac{1}{4}=3S_N^2(x_i)+\frac{(\alpha+5)(\alpha+1)}{4x_i^2}-\frac{2N+\alpha+1}{2x_i},\qquad i=1,2,\cdots,N.$$

The function $H(z)$ given by (8) with $z_i > 0$ for all z_i has a stationary value at these points if $u(z) = V(\alpha+2, z) + (1/z) + \lambda'$, where λ' is an arbitrary constant. There also holds

$$S_N^4(x_i) = \left[\frac{\alpha^4 - 110\alpha^2 - 360\alpha - 251}{720}\right]x_i^{-4} - \left[\frac{\alpha^3 + (2N+1)\alpha^2 - 19\alpha - (38N+19)}{180}\right]x_i^{-3} \tag{9}$$

$$+\left[\frac{3\alpha^2 + 4(2N+1)\alpha + (8N^2+8N+1)}{360}\right]x_i^{-2} - \left[\frac{\alpha+2N+1}{180}\right]x_i^{-1} + \frac{1}{720}.$$

C. Jacobi zeros. The SL form of the differential equation for the Jacobi polynomials $f(x) = P_N^{(\alpha,\beta)}(x)$, $\alpha > -1$, $\beta > -1$, is

$$[(1-x)^{\alpha+1}(1+x)^{\beta+1}f'(x)]' + \lambda(1-x)^{\alpha}(1+x)^{\beta}f(x) = 0, \qquad \lambda = N(N+\alpha+\beta+1).$$

Let x_k, $k = 1, 2, \cdots, N$ be the zeros of $P_N^{(\alpha,\beta)}(x)$. The corresponding form of (7) is:

$$N(N+\alpha+\beta+1) = 3(1-x_i^2)S_N^2(x_i) + \frac{(\alpha+5)(\alpha+1)}{4}\frac{(1+x_i)}{(1-x_i)}$$

$$+\frac{(\beta+5)(\beta+1)}{4}\frac{(1-x_i)}{(1+x_i)} - \frac{(\alpha+1)(\beta+1)}{2}, \qquad i = 1, 2, \cdots, N.$$

The function $H(z)$ defined by (8) where $-1 < z_i < 1$ for all z_i has a stationary value at these points if

$$u(z) = \frac{(\alpha+1)(\alpha+3)}{4(1-z)^2} + \frac{(\beta+1)(\beta+3)}{4(1+z)^2} - \frac{2\lambda + (\alpha+1)(\beta+1)}{2(1-z^2)} + \lambda',$$

where λ' is an arbitrary constant. Defining $a = \alpha - \beta$ and $b = \alpha + \beta + 2$, there also holds

$$S_N^4(x_i) = \frac{c_4(x_i)}{(1-x_i^2)^4} + \frac{c_3(x_i)}{(1-x_i^2)^3} + \frac{c_2(x_i)}{(1-x_i^2)^2} + \frac{c_1(x_i)}{(1-x_i^2)} \tag{10}$$

where

$$c_4(x) = \left(\frac{b^4}{720} - \frac{b^3}{90} - \frac{8b^2}{45} - \frac{48b}{30}\right)x^4 - \left(\frac{ab^3}{180} - \frac{ab^2}{30} - \frac{34ab}{45} - \frac{2b^2}{5} - \frac{8a}{5}\right)x^3$$

$$+\left(\frac{a^2b^2}{120} + \frac{7ab}{18} - \frac{a^2b}{45} - \frac{26a^2}{45}\right)x^2 - \frac{a^3(b-2)}{180}x + \frac{a^4}{720},$$

$$c_3(x) = -\left(\frac{b^3}{180} + \frac{8b^2}{45} + \frac{48b}{30} - \frac{2\lambda b}{45} - \frac{4\lambda}{5}\right)x^2$$

$$+\left(\frac{ab^2}{90} + \frac{43ab}{90} - \frac{3b^2}{10} + \frac{12a}{15} - \frac{2\lambda a}{45}\right)x - \left(\frac{a^2b}{180} + \frac{a^2 - ab}{10}\right),$$

$$c_2(x) = -\frac{\lambda}{90}x^2 + \frac{\lambda ab}{45}x + \left[\frac{\lambda^2}{45} - (a^2 + 2b - 18)\frac{\lambda}{90} - \frac{b}{5}\right],$$

$$c_1(x) = -\frac{2b^2}{45}.$$

4. Final remarks. The method used in this paper to obtain the sums $S_N^l(x_i)$ for zeros of polynomials satisfying a second order linear ODE is quite elementary. However, it is applicable to any polynomial (belonging to certain family or not, with complex or real zeros) whenever the coefficients of the ODE are analytic functions at its zeros.

It makes it possible to relate the sums $S_N^l(x_i)$ to the coefficients evaluated at x_i. An interesting feature of these relations is that $S_N^3(x_i)$ turns out to be the value of the derivative of a certain function evaluated at x_i. This can be interpreted as an extremal property of the zeros.

In the case of a polynomial solution of an SL equation, these expressions give a direct relation between the zeros and the eigenvalue of the equation and it is interesting to note that in the Hermite and Laguerre cases, the function whose derivative is related to $S_N^3(x_i)$ is the "potential function" appearing when the differential equation is written in its normal form.

The sums $S_N^l(x_i)$ can be related to the coefficients of the ODE just through (6), the expression that follows for $g^{(l)}(x_i)$ in terms of $f^{(l+1)}(x_i)$, and the differential equation for $f(x)$. An alternative technique to obtain these relationships has been used by S. Ahmed (see [1]); however, some interesting aspects of these relations, like the minimal property of the zeros, are not shown in that work.

The extremal property of the zeros of the classical polynomials can also be interpreted as the equilibrium condition for certain classical one-dimensional many-body problems (Calogero [4]) and, closely related to this, it is found their electrostatic interpretation (Szëgo [9, § 6.7]; Forrester and Rogers [8]).

All of the equations reported in this paper in § 3 concerning the sums $S_N^l(x_i)$ are already known (Ahmed et al. [2]), with the exception of (9) and (10), which are presumably new.

REFERENCES

[1] S. Ahmed, *A general technique to obtain nonlinear equations for the zeros of classical orthogonal polynomials*, Lett. Nuovo Cimento, 26 (1979), pp. 285-288.

[2] S. Ahmed, M. Bruschi, F. Calogero, M. A. Olshanetsky and A. M. Perelomov, *Properties of the zeros of the classical polynomials and of the Bessel functions*, Nuovo Cimento B, 49 (1979), pp. 173-199.

[3] F. Calogero, *Matrices, differential operators, and polynomials*, J. Math. Phys., 22 (1981), pp. 919-934.

[4] ———, *Integrable dynamical systems and related mathematical results*, in Lecture Notes in Physics, 189, Springer-Verlag, Berlin, 1983, pp. 88-96.

[5] R. G. Campos, *A non-perturbative method for the $\kappa x^2+\beta x^4$ interaction*, Rev. Mexicana Fis., 32 (1986), pp. 379-400.

[6] ———, *A nonlocal equation for the wave function of the harmonic oscillator when the position spectrum is to be made discrete*, Rev. Mexicana Fis., 29 (1983), pp. 217-236.

[7] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[8] P. J. Forrester and J. B. Rogers, *Electrostatics and the zeros of the classical polynomials*, this Journal, 17 (1986), pp. 461-468.

[9] G. Szëgo, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1976.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

013

ON AN EXPANSION THEOREM OF F. V. ATKINSON AND P. BINDING*

HANS VOLKMER†

Abstract. We study an expansion theorem of F. V. Atkinson and P. Binding for the right definite multiparameter eigenvalue problem

$$x_m = T_m x_m + \sum_{n=1}^{k} \lambda_n V_{mn} x_m, \qquad m = 1, \cdots, k$$

where $\lambda_1, \cdots, \lambda_k$ are real, x_m is a nonzero element of some separable Hilbert space and T_m, V_{mn} are compact symmetric. We present a new version of the theorem which is easier to use in order to prove completeness of eigenvectors under some additional assumptions. In particular, we prove completeness of eigenvectors (i) under Minkowski's definiteness condition and (ii) for arbitrary two parameter problems.

Key words. multiparameter problem, completeness of eigenfunctions, Minkowski matrices, integral equations

AMOS(MOS) subject classifications. 47A70, 45C05

1. Introduction. This paper is devoted to an open question in the theory of right definite multiparameter eigenvalue problems that is the fundamental question of the completeness of eigenvectors. We know the expansion theorem of Atkinson [1, Thms. 11.8.1, 11.10.1] which has been generalized and improved by several authors. Binding, Källström and Sleeman [5, Thm. 4.2] derived an expansion theorem under a definiteness condition weaker than that used by Atkinson. Binding [3, Thm. 4.1] gave another version of the expansion theorem for right definite problems on the basis of his oscillation theorem. However, it turns out that it is often difficult to decide on the completeness of eigenvectors by means of these expansion theorems (see [14]).

In this paper we modify known proofs of the theorem in order to obtain simpler sufficient conditions for the completeness of eigenvectors. There are two main results. In § 2 we prove the completeness of eigenvectors under Minkowski's definiteness condition. In § 3 we demonstrate the completeness of eigenvectors for the general two parameter right definite problem. Thus some progress is achieved in the question of completeness but, for instance, the question remains open for the general three parameter problem.

We now formulate the multiparameter eigenvalue problem under consideration and give a short survey on its theory. The notation and the results are mainly taken from Binding [3].

Let T_m, V_{mn} be compact symmetric operators on separable Hilbert spaces H_m for $m, n = 1, \cdots, k$. Then define the compact symmetric operators

$$W_m(\lambda) = T_m + \sum_{n=1}^{k} \lambda_n V_{mn}$$

for $m = 1, \cdots, k$, $\lambda = (\lambda_1, \cdots, \lambda_k) \in \mathbb{R}^k$ and consider the k-parameter eigenvalue problem

$$x_m = W_m(\lambda) x_m, \qquad 0 \neq x_m \in H_m, \quad m = 1, \cdots, k. \tag{1.1}$$

λ is called an eigenvalue of (1.1) if there exist nonzero vectors x_m in H_m such that $x_m = W_m(\lambda) x_m$ for all $m = 1, \cdots, k$. The tensor $x_1 \otimes \cdots \otimes x_k$ in the Hilbert space tensor

* Received by the editors July 15, 1985; accepted for publication (in revised form) December 1, 1986.

† Fachbereich 6, Mathematik, Universität, GHS Essen, Universitätsstr. 3, D-4300 Essen 1, West Germany.

product $H=H_1\otimes\cdots\otimes H_k$ is called an eigenvector of (1.1) corresponding to the eigenvalue λ. The tensor product setting is necessary in order to give any sense to the question of the completeness of the eigenvectors. The eigenvalue problem (1.1) has been studied mostly under the "right definiteness" condition

$$\text{(1.2)}\qquad \det_{1\leqq m,n\leqq k}(V_{mn}x_m, x_m)>0 \quad \text{for all } 0\neq x_m\in H_m.$$

We shall assume (1.2) throughout this paper.

In § 4, the theory of (1.1) will be applied to eigenvalue problems for integral operators

$$\text{(1.3)}\quad x_m(\xi_m)=\int_{a_m}^{b_m} t_m(\xi_m,\eta_m)x_m(\eta_m)\,d\eta_m+\sum_{n=1}^{k}\lambda_n\int_{a_m}^{b_m} v_{mn}(\xi_m,\eta_m)x_m(\eta_m)\,d\eta_m$$

where t_m and v_{mn} are real symmetric functions continuous on $[a_m, b_m]\times[a_m, b_m]$. The corresponding Hilbert spaces are

$$H_m=L^2(a_m,b_m)\quad\text{and}\quad H=H_1\otimes\cdots\otimes H_k=L^2\left(\prod_{m=1}^{k}(a_m,b_m)\right).$$

The eigenvalue problems (1.1) and (1.3) have also been discussed by Anna Pell [11] in the special case of two parameters. Pell assumes that $H_1=H_2=l^2$ and $T_1=T_2=0$. The operators V_{mn} are defined by real symmetric Hilbert-Schmidt matrices. The eigenvalue problem is assumed to be left-definite, i.e., V_{12} and $-V_{22}$ are positive definite. I intend to generalize the results of Pell in the left-definite case to more than two parameters in another paper.

The theory of (1.1) can also be applied to multiparameter problems involving unbounded operators. Let T'_m and V'_{mn} be selfadjoint operators on H_m, with V'_{mn} bounded and T'_m bounded above with compact resolvent.

Then the multiparameter eigenvalue problem

$$\text{(1.4)}\qquad T'_m x_m+\sum_{n=1}^{k}\lambda_n V'_{mn}x_m=0,\quad 0\neq x_m\in D(T'_m),\quad m=1,\cdots,k,$$

can be transformed into a problem of the form (1.1) (see [3, § 1]).

A special case of (1.4) is the k-parameter Sturm-Liouville eigenvalue problem

$$\text{(1.5)}\qquad \frac{d^2x_m}{d\xi_m^2}+q_m x_m+\sum_{n=1}^{k}\lambda_n v_{mn}x_m=0,\qquad m=1,\cdots,k,$$

with the boundary conditions

$$\alpha_{m1}x_m(a_m)+\alpha_{m2}\frac{dx_m}{d\xi_m}(a_m)=0,\qquad (0,0)\neq(\alpha_{m1},\alpha_{m2})\in\mathbb{R}^2,$$

$$\beta_{m1}x_m(b_m)+\beta_{m2}\frac{dx_m}{d\xi_m}(b_m)=0,\qquad (0,0)\neq(\beta_{m1},\beta_{m2})\in\mathbb{R}^2.$$

This problem appears when we solve boundary value problems for partial differential equations by the method of separation of variables. For example, if we consider the vibration problem of an elliptic membrane with clamped boundary then we obtain a two-parameter eigenvalue problem of the form (1.5) where the ordinary differential equations are Mathieu equations (see [10, § 4.3]).

The eigenvalue problems (1.4) and (1.5) have been investigated by many authors. Cordes [7] proved the completeness of eigenfunctions in the two-parameter case if V'_{11}, V'_{12}, $-V'_{21}$, V'_{22} are positive definite and $V'_{11}+V'_{12}=Id_1$, $-V'_{21}+V'_{22}=Id_2$ where Id_m denotes the identity operator on H_m. This result has been generalized by Volkmer [14] to more than two parameters.

Faierman [8] proved the completeness of eigenfunctions of problem (1.5) under the assumption of continuous v_{mn} satisfying

$$\det_{1\leqq m,n\leqq k} v_{mn}(\xi_m)>0 \quad \text{for all } \xi_m\in[a_m, b_m].$$

More generally, the combined results of Browne [6], Källström and Sleeman [9] and Volkmer [15] establish the completeness of eigenvectors if the eigenvalue problem (1.4) is strongly definite. By a result of Binding [2] this means that

$$(1.6) \qquad \det_{1\leqq m,n\leqq k} (V'_{mn}x_m, x_m)\geqq \varepsilon \prod_{m=1}^{k} (x_m, x_m) \quad \text{for all } x_m\in H_m.$$

It should be noted that the eigenvalue problem (1.1) cannot be strongly definite (i.e., (1.6) holds with V_{mn} in place of V'_{mn}) unless H has finite dimensions because the operators V_{mn} are compact.

The following oscillation theorem of Binding [3, Cor. 3.3] proves the existence of eigenvalues of (1.1).

THEOREM 1.1. *Let $i=(i_1,\cdots,i_k)$ be a multi-index where $1\leqq i_m\leqq \dim H_m$ for each m. Then there exists a unique eigenvalue of* (1.1), *denoted by λ^i, such that the i_mth greatest eigenvalue of $W_m(\lambda)$ is equal to* 1 *for each m. Of course, the eigenvalues of $W_m(\lambda)$ are counted according to multiplicity.*

In order to formulate the expansion theorem we need some further notation. The operator determinant Δ_0 is defined by

$$(1.7) \qquad \Delta_0=\otimes\begin{vmatrix} V_{11}\cdots V_{1k}\\ \vdots \qquad \vdots\\ V_{k1}\cdots V_{kk}\end{vmatrix}$$

where $\otimes$ indicates that the tensor product of operators is used in the expansion of the determinant. For $n=1,\cdots,k$, the operator determinant Δ_n is defined as in (1.7) with the nth column of the determinant replaced by Id_m-T_m, $m=1,\cdots,k$. Then $\Delta_0,\cdots,\Delta_k$ are bounded symmetric operators on the tensor product H. We write

$$(1.8) \qquad [z_1, z_2]=(\Delta_0 z_1, z_2)$$

for $z_1, z_2\in H$. It follows from (1.2) that the form [,] is positive semi-definite (see [1, p. 194]).

We can choose eigenvectors e^i of (1.1) corresponding to the eigenvalues λ^i such that the e^i form an orthonormal system in $(H, [\ ,\ ])$. If all the spaces H_m have finite dimensions then the completeness of the e^i follows easily from the oscillation theorem by counting the number of the indices i. This completeness result was shown earlier by Atkinson [1, Thm. 7.9.1]. If at least one of the spaces H_m has infinite dimensions then the e^i form an infinite orthonormal system in $(H, [\ ,\ ])$. The expansion theorem of Binding [3, Thm. 4.1] which is closely related to a theorem of Atkinson [1, Thm. 1.10.1] now reads as follows.

THEOREM 1.2. *Let $y=y_1\otimes\cdots\otimes y_k$ be a decomposable tensor in H and assume that*

$$(1.9) \qquad \Delta_n y\in\Delta_0(H)$$

for all $n=1,\cdots,k$. Then there holds the Parseval's equality

$$[y,y]=\sum_i |[y,e^i]|^2.$$

In general, it is difficult to check condition (1.9) which guarantees that y can be expanded in a series of eigenvectors. In the next section we shall reprove Theorem 1.2 in order to point out the modifications we need to obtain conditions which are simpler to check than (1.9).

2. Expansion theorems for k-parameter eigenvalue problems. Consider the eigenvalue problem (1.1) under the definiteness condition (1.2) and let $y=y_1\otimes\cdots\otimes y_k$ be a decomposable tensor in H. We shall discuss the question of whether y can be expanded in a series of eigenvectors of (1.1). As for the expansion theorems of Atkinson and Binding our results will be based on sequences of orthoprojectors P_m^j on H_m strongly convergent to Id_m as $j\to\infty$ for all $m=1,\cdots,k$. We choose these projectors so that the range spaces

$$H_m^j:=P_m^j(H_m)$$

have finite dimensions, contain y_m and are montone increasing

$$y_m\in H_m^1\subseteq H_m^2\subseteq H_m^3\subseteq\cdots.$$

Then we consider the approximating eigenvalue problems

$$x_m=P_m^jW_m(\lambda)x_m,\quad 0\neq x_m\in H_m^j,\quad m=1,\cdots,k. \tag{2.1}$$

We denote the eigenvalues of (2.1) by λ^{ij} according to Theorem 1.1 where

$$i\in I_j:=\{(i_1,\cdots,i_k)\colon 1\leqq i_m\leqq \dim H_m^j\}.$$

Choose corresponding unit vectors u_m^{ij} so that

$$u_m^{ij}=P_m^jW_m(\lambda^{ij})u_m^{ij}$$

and the tensors $u^{ij}=u_1^{ij}\otimes\cdots\otimes u_k^{ij}$, $i\in I_j$, form an [,]-orthogonal system of eigenvectors of (2.1).

Binding [3, Cor. 3.3] has shown that the sequences λ^{ij} converge to the unique eigenvalue λ^i of Theorem 1.1 as $j\to\infty$ and that the weak limit points of the sequences u_m^{ij} are also strong limit points and furnish corresponding eigenvectors. Therefore, by taking subsequences, we may assume that

$$\lambda^{ij}\to\lambda^i,\qquad u_m^{ij}\to u_m^i\quad\text{as } j\to\infty\quad\text{for } i\in I:=\bigcup_j I_j.$$

Condition (1.2) allows us to renormalize the eigenvectors

$$e^{ij}=u^{ij}/[u^{ij},u^{ij}]^{1/2},$$
$$e^i=u^i/[u^i,u^i]^{1/2}\quad\text{where } u^i=u_1^i\otimes\cdots\otimes u_k^i$$

so that the e^{ij} form an orthonormal basis in the finite dimensional space $(H_1^j\otimes\cdots\otimes H_k^j$, [,]) and the e^i form a possibly incomplete orthonormal system in (H, [,]). Hence there holds the expansion

$$y=\sum_{i\in I_j}\alpha^{ij}e^{ij},\qquad \alpha^{ij}=[y,e^{ij}] \tag{2.2}$$

and Parseval's equality

$$[y,y]=\sum_{i\in I_j}|\alpha^{ij}|^2 \tag{2.3}$$

for every j. We see that $\alpha^{ij} \to [y, e^i]$ as $j \to \infty$. Hence (2.3) will give the desired Parseval's equality

$$[y, y] = \sum_{i \in I} |[y, e^i]|^2 \tag{2.4}$$

provided that (empty sum $= 0$)

$$\sum_{i \in I_j \setminus I_{j_0}} |\alpha^{ij}|^2 \to 0 \quad \text{as } j_0 \to \infty \text{ uniformly in } j. \tag{2.5}$$

The next lemma yields a convenient method in order to verify this condition. We denote by Q the set of all $\lambda \in \mathbb{R}^k$ such that

$$1 = (W_m(\lambda)u_m, u_m), \qquad m = 1, \cdots, k \tag{2.6}$$

for some unit vectors u_m in H_m. Obviously, all eigenvalues λ^{ij} and λ^i belong to Q.

LEMMA 2.1. *Assume that there exists a functional $f: Q \to \mathbb{R}$ bounded from below so that*

(i) *the set*

$$\{\lambda \in Q \,|\, f(\lambda) \leqq c\}$$

is bounded for all real c and

(ii) *the sequence*

$$\sum_{i \in I_j} |\alpha^{ij}|^2 f(\lambda^{ij}), \qquad j = 1, 2, \cdots$$

is bounded.

Then (2.5) *is satisfied and hence Parseval's equality* (2.4) *holds.*

Proof. Conditions (i) and (ii) remain valid if we replace $f(\lambda)$ by $f(\lambda) + c_0$ where c_0 is a constant. Hence we may assume that $f(\lambda) \geqq 0$ on Q. It follows from (ii) that there is a real number c_1 independent of j and j_0 so that

$$0 \leqq \sum_{i \in I_j \setminus I_{j_0}} |\alpha^{ij}|^2 f(\lambda^{ij}) \leqq \sum_{i \in I_j} |\alpha^{ij}|^2 f(\lambda^{ij}) \leqq c_1.$$

We know from [3, Thm. 2.5] that the sequence $\lambda^{i(j)j}$, $j = 1, 2, 3, \cdots$, is unbounded whenever the sequence of indices $i(j)$ is unbounded. Hence (i) implies that, for any given $c > 0$, there is a j_0 such that

$$f(\lambda^{ij}) \geqq c \quad \text{if } j \geqq j_0, \qquad i \in I_j \setminus I_{j_0}.$$

It follows that

$$0 \leqq c \cdot \sum_{i \in I_j \setminus I_{j_0}} |\alpha^{ij}|^2 \leqq \sum_{i \in I_j \setminus I_{j_0}} |\alpha^{ij}|^2 f(\lambda^{ij}) \leqq c_1$$

which proves (2.5). □

If we choose

$$f(\lambda) = \lambda_1^2 + \cdots + \lambda_k^2 \tag{2.7}$$

then we obtain the following expansion theorem which slightly improves Theorem 1.2.

THEOREM 2.2. (i) *If*

$$\Delta_n y \in \Delta_0^{1/2}(H) \tag{2.8}$$

for some n, then the sequence

$$\sum_{i \in I_j} |\alpha^{ij}|^2 |\lambda_n^{ij}|^2, \qquad j = 1, 2, 3, \cdots$$

is bounded.

(ii) *Assume that* (2.8) *holds for all $1 \leqq n \leqq k$. Then Parseval's equality* (2.4) *is satisfied hence y can be expanded in a series of eigenvectors of* (1.1).

Proof. Condition (ii) is a consequence of (i) and Lemma 2.1 applied to the functional (2.7). Hence it suffices to prove (i).

Since the e^{ij}'s are eigenvectors of (2.1) we know from [1, Thm. 6.8.1] that

$$(P_1^j \otimes \cdots \otimes P_k^j)\Delta_n e^{ij} = \lambda_n^{ij}(P_1^j \otimes \cdots \otimes P_k^j)\Delta_0 e^{ij}. \tag{2.9}$$

We multiply (2.9) from the left by y and obtain

$$(\Delta_n y, e^{ij}) = \lambda_n^{ij}\alpha^{ij}$$

from (2.2) and $(P_1^j \otimes \cdots \otimes P_k^j)y = y$. Hence

$$\sum_{i \in I_j} |\alpha^{ij}|^2 |\lambda_n^{ij}|^2 = \sum_{i \in I_j} |(\Delta_n y, e^{ij})|^2. \tag{2.10}$$

By assumption, we can write $\Delta_n y = \Delta_0^{1/2} x$.

Without loss of generality, we may assume that x is orthogonal to the kernel of $\Delta_0^{1/2}$. Then there exists a sequence x^q in H such that $\Delta_0^{1/2} x^q \to x$ as $q \to \infty$. Consequently, $\Delta_0 x^q \to \Delta_0^{1/2} x = \Delta_n y$. It follows from Bessel's inequality that

$$\sum_{i \in I_j} |[x^q, e^{ij}]|^2 \leqq [x^q, x^q].$$

This inequality gives

$$\sum_{i \in I_j} |(\Delta_n y, e^{ij})|^2 \leqq (x, x) \tag{2.11}$$

as $q \to \infty$. Now (2.10) and (2.11) imply (i). □

The condition (2.8) of Theorem 2.2 is weaker than the corresponding condition (1.9) of Theorem 1.2 because the range of Δ_0 is contained in the range of $\Delta_0^{1/2}$. This weakening will be important in the next section in connection with Lemma 3.4.

However, the main idea of this paper is to replace the functional (2.7) by some other functionals.

THEOREM 2.3. *Let f be a linear functional on $\mathbb{R}^k$.*

(i) *Then condition* (ii) *of Lemma* 2.1 *is automatically satisfied.*

(ii) *If f satisfies condition* (i) *of that lemma then there exists an orthonormal basis of eigenvectors of* (1.1) *in* $(H, [\ ,\ ])$.

Proof. (i) Multiply (2.9) from the right by $e^{i_0 j}$. It follows that

$$(\Delta_n e^{ij}, e^{i_0 j}) = \lambda_n^{ij}(\Delta_0 e^{ij}, e^{i_0 j}) = \begin{cases} 0 & \text{if } i \neq i_0, \\ \lambda_n^{ij} & \text{if } i = i_0. \end{cases}$$

Now (2.2) gives

$$(\Delta_n y, y) = \sum_{i \in I_j} |\alpha^{ij}|^2 \lambda_n^{ij}.$$

Therefore, the numbers $\sum_{i \in I_j} |\alpha^{ij}|^2 f(\lambda^{ij})$ are independent of j and hence bounded.

(ii) Since f is linear and satisfies (i) of Lemma 2.1, it follows that f is bounded from below on Q. Therefore, we can use Lemma 2.1 to prove Parseval's equality (2.4) for the given decomposable tensor y. Now the decomposable tensors form a total subset of H, i.e., their linear hull is dense in H. Since y can be chosen arbitrarily in this total subset of H we see that (2.4) holds for all y in H and every choice of the $[\ ,\ ]$-orthonormal system e^i $(i \in I)$ of eigenvectors of (1.1). This proves the theorem. □

Let us apply this theorem to an eigenvalue problem (1.1) which satisfies Minkowski's definiteness condition, i.e.,

(i) V_{mn} is negative definite for all $m \neq n$, and

(ii) $\sum_{n=1}^{k} V_{mn}$ is positive definite for all m.

This condition has been introduced in multiparameter theory by Binding and Browne [4]. They showed that, under Minkowski's condition, the operator determinant Δ_0 and all the cofactors Δ_{omn} of V_{mn} in the expansion of Δ_0 are positive definite. These results and Theorem 2.3 lead to the following theorem.

THEOREM 2.4. *If the eigenvalue problem* (1.1) *satisfies Minkowski's definiteness condition then there exists an orthonormal basis of eigenvectors of* (1.1) *in* $(H, [\ ,\ ])$.

Proof. Let $\lambda \in Q$. Then Cramer's rule applied to the linear system (2.6) yields

$$\lambda_n = \frac{(\Delta_n u, u)}{(\Delta_0 u, u)}, \qquad n = 1, \cdots, k \tag{2.12}$$

where $u = u_1 \otimes \cdots \otimes u_k$. In addition to the hypothesis of the theorem let us assume that $Id_m - T_m$ is positive semi-definite for all m. Then

$$\Delta_n = \sum_{m=1}^{k} (Id_m - T_m) \otimes \Delta_{omn}$$

shows that $(\Delta_n u, u)$ is nonnegative for all n. Hence, by (2.12), $Q \subseteq [0, \infty[^k$ and, therefore, condition (i) of Lemma 2.1 is satisfied if we choose $f(\lambda) = \sum_{n=1}^{k} \lambda_n$. By Theorem 2.3(ii), this completes the proof under our additional assumption. If the operator T_m has eigenvalues greater than 1 for some m then we apply the following simple lemma to transform the eigenvalue problem in such a way that our additional assumption is satisfied. □

LEMMA 2.5. *There exists a translation* $\lambda = \hat{\lambda} + \mu$ *in the parameter space* $\mathbb{R}^k$ *which transforms* (1.1) *into*

$$x_m = \hat{T}_m x_m + \sum_{n=1}^{k} \hat{\lambda}_n V_{mn} x_m, \qquad m = 1, \cdots, k$$

such that all eigenvalues of $\hat{T}_m$ *are smaller than* 1.

Proof. Let μ be the eigenvalue of the problem

$$x_m = 2 W_m(\lambda) x_m, \quad 0 \neq x_m \in H_m, \quad m = 1, \cdots, k$$

corresponding to the multiindex $i = (1, \cdots, 1)$. Then the greatest eigenvalue of $2 W_m(\mu)$ is equal to one. Hence the greatest eigenvalue of $W_m(\mu) = \hat{T}_m$ is equal to $\frac{1}{2}$. □

3. Two-parameter eigenvalue problems. In this section we want to prove the completeness of eigenvectors for the eigenvalue problem (1.1) in the special case of two parameters

$$\begin{aligned} x_1 &= T_1 x_1 + \lambda_1 V_{11} x_1 + \lambda_2 V_{12} x_1, \qquad 0 \neq x_1 \in H_1, \\ x_2 &= T_2 x_2 + \lambda_1 V_{21} x_1 + \lambda_2 V_{22} x_2, \qquad 0 \neq x_2 \in H_2, \end{aligned} \tag{3.1}$$

under the definiteness condition

$$\begin{vmatrix} (V_{11} x_1, x_1) & (V_{12} x_1, x_1) \\ (V_{21} x_2, x_2) & (V_{22} x_2, x_2) \end{vmatrix} > 0 \quad \text{for all } 0 \neq x_1 \in H_1, 0 \neq x_2 \in H_2. \tag{3.2}$$

As a first step we reduce (3.1) to an appropriate canonical form by an affine transformation of the eigenvalues

$$\lambda_1 = \gamma_{10} + \gamma_{11} \hat{\lambda}_1 + \gamma_{12} \hat{\lambda}_2, \qquad \lambda_2 = \gamma_{20} + \gamma_{21} \hat{\lambda}_1 + \gamma_{22} \hat{\lambda}_2 \tag{3.3}$$

where

$$\gamma_{11}\gamma_{22}-\gamma_{12}\gamma_{21}=1. \tag{3.4}$$

Such a transformation takes (3.1) into

$$x_1=\hat{T}_1x_1+\hat{\lambda}\hat{V}_{11}x_1+\hat{\lambda}_2\hat{V}_{12}x_1, \qquad x_2=\hat{T}_2x_2+\hat{\lambda}_1\hat{V}_{21}x_2+\hat{\lambda}_2\hat{V}_{22}x_2, \tag{3.5}$$

where

$$\hat{T}_m:=T_m+\gamma_{10}V_{m1}+\gamma_{20}V_{m2} \qquad \hat{V}_{mn}:=\gamma_{1n}V_{m1}+\gamma_{2n}V_{m2}.$$

It follows from (3.4) that the operator determinant $\hat{\Delta}_0$ for (3.5) is equal to Δ_0:

$$\begin{aligned}\hat{\Delta}_0&=\hat{V}_{11}\otimes\hat{V}_{22}-\hat{V}_{12}\otimes\hat{V}_{21}\\&=(\gamma_{11}V_{11}+\gamma_{21}V_{12})\otimes(\gamma_{12}V_{21}+\gamma_{22}V_{22})\\&\qquad-(\gamma_{12}V_{11}+\gamma_{22}V_{12})\otimes(\gamma_{11}V_{21}+\gamma_{21}V_{22})\\&=(\gamma_{11}\gamma_{22}-\gamma_{12}\gamma_{21})(V_{11}\otimes V_{22}-V_{12}\otimes V_{21})\\&=\Delta_0.\end{aligned}$$

Moreover, (3.1) and (3.5) have the same eigenvectors. Hence it will be sufficient to prove the completeness of eigenvectors for (3.5). The reduction to canonical form will be achieved by the next lemma.

LEMMA 3.1. *There exists an affine transformation* (3.3), (3.4) *such that the transformed eigenvalue problem* (3.5) *satisfies* (i) *and* (ii).

(i) *There holds one of the following six sets of sign conditions*:

(1) $\hat{V}_{11}>0,\quad \hat{V}_{21}=0,\quad \hat{V}_{22}>0$;

(2) $\hat{V}_{11}=0,\quad \hat{V}_{21}<0,\quad \hat{V}_{12}>0$;

(3) $\hat{V}_{11}>0,\quad \hat{V}_{21}\leqq 0,\quad \hat{V}_{12}\geqq 0,\quad \hat{V}_{22}>0$;

(4) $\hat{V}_{11}\geqq 0,\quad \hat{V}_{21}<0,\quad \hat{V}_{12}>0,\quad \hat{V}_{22}\geqq 0$;

(5) $\hat{V}_{11}>0,\quad \hat{V}_{21}\leqq 0,\quad \hat{V}_{12}>0,\quad \hat{V}_{22}\geqq 0,\quad -\hat{V}_{21}+\hat{V}_{22}>0$;

(6) $\hat{V}_{11}\geqq 0,\quad \hat{V}_{21}<0,\quad \hat{V}_{12}\geqq 0,\quad \hat{V}_{22}>0,\quad \hat{V}_{11}+\hat{V}_{12}>0$.

(ii) $\hat{\Delta}_2=\hat{V}_{11}\otimes(Id_2-\hat{T}_2)-(Id_1-\hat{T}_1)\otimes\hat{V}_{21}$ *is positive definite.*

Proof. First we construct $\gamma_{11},\gamma_{21},\gamma_{12},\gamma_{22}$ such that (i) holds. By [1, Thm. 9.2.2] there exist nonzero $\alpha=(\alpha_1,\alpha_2)$, $\beta=(\beta_1,\beta_2)\in\mathbb{R}^2$ such that

$$\begin{aligned}\alpha_1V_{11}+\alpha_2V_{12}&\geqq 0, \qquad &\beta_1V_{11}+\beta_2V_{12}&\geqq 0,\\ \alpha_1V_{21}+\alpha_2V_{22}&\leqq 0, \qquad &\beta_1V_{21}+\beta_2V_{22}&\geqq 0.\end{aligned}$$

Suppose that $\alpha=\mu\beta$ for some real μ. Then

$$\begin{aligned}\alpha_1V_{11}+\alpha_2V_{12}&=\beta_1V_{11}+\beta_2V_{12}=0\quad\text{if } \mu<0,\\ \alpha_1V_{21}+\alpha_2V_{22}&=\beta_1V_{21}+\beta_2V_{22}=0\quad\text{if } \mu>0.\end{aligned}$$

Set $\gamma_{11}=\alpha_1$, $\gamma_{21}=\alpha_2$ and choose γ_{21}, γ_{22} such that (3.4) holds. If $\mu>0$ then it follows that $\hat{V}_{11}\geqq 0$ and $\hat{V}_{21}=0$. Hence $\Delta_0=\hat{\Delta}_0=\hat{V}_{11}\otimes\hat{V}_{22}$ and (3.2) yield $\hat{V}_{11}>0$ and $\hat{V}_{22}>0$. Therefore, (i) (1) is satisfied. Similarly, if $\mu<0$ then (i) (2) holds.

Now suppose that α,β are linearly independent. Then we set $\gamma_{11}=\alpha_1$, $\gamma_{21}=\alpha_2$, $\gamma_{12}=\beta_1$, $\gamma_{22}=\beta_2$ where we may assume (3.4). We obtain

$$\hat{V}_{11}\geqq 0,\quad \hat{V}_{21}\leqq 0,\quad \hat{V}_{12}\geqq 0,\quad \hat{V}_{22}\geqq 0.$$

Since $(\hat{V}_{11}x_1, x_1)(\hat{V}_{22}x_2, x_2) - (\hat{V}_{12}x_1, x_1)(\hat{V}_{21}x_2, x_2) > 0$ for every $0 \neq x_1 \in H_1$, $0 \neq x_2 \in H_2$, we conclude that

$$\hat{V}_{11} > 0 \quad \text{or} \quad \hat{V}_{21} < 0$$

and

$$\hat{V}_{12} > 0 \quad \text{or} \quad \hat{V}_{22} > 0.$$

Further, we see that

$$\hat{V}_{11} + \hat{V}_{12} > 0 \quad \text{and} \quad -\hat{V}_{21} + \hat{V}_{22} > 0.$$

Hence, there holds one of the last four sets of sign conditions.

Now we construct γ_{10}, γ_{20} such that (ii) holds. By Lemma 2.5, we can choose γ_{10} and γ_{20} such that $Id_1 - \hat{T}_1 > 0$ and $Id_2 - \hat{T}_2 > 0$. By (i), we know that $\hat{V}_{11} \geqq 0$ and $\hat{V}_{21} \leqq 0$ where at least one of these two operators is definite.

Now we use the easily established facts that

$$A_1 \geqq 0, \quad A_2 \geqq 0 \quad \text{imply } A_1 \otimes A_2 \geqq 0$$

and

$$A_1 > 0, \quad A_2 > 0 \quad \text{imply } A_1 \otimes A_2 > 0$$

whenever A_m are bounded symmetric operators on H_m. Hence $\hat{V}_{11} \otimes (Id_2 - \hat{T}_2) \geqq 0$, $-(Id_1 - \hat{T}_1) \otimes \hat{V}_{21} \geqq 0$ and one of these operators is positive definite. Therefore, (ii) is true. □

The following definiteness result shows that, for $k = 2$, the form (1.5) is not only positive semi-definite but positive definite on H. Hence $(H, [\ ,\ ])$ is a pre-Hilbert space under the assumptions of this section. A similar result for general k greater than two is not known.

THEOREM 3.2. *Condition* (3.2) *implies that* Δ_0 *is positive definite on* H.

Proof. We may assume that (3.1) is already reduced to one of the six canonical forms of Lemma 3.1. In case (1) we have $\Delta_0 = V_{11} \otimes V_{22}$ and V_{11} and V_{22} are positive definite. Hence Δ_0 is positive definite. Case (2) is similar. If (3) holds then $\Delta_0 \geqq V_{11} \otimes V_{22} > 0$. Case (4) is similar. Suppose that (5) holds. Since $\Delta_0 \geqq 0$ it suffices to show that $(\Delta_0 z, z) = 0$ implies $z = 0$. Hence assume that $(\Delta_0 z, z) = 0$ for some $z \in H$. Then

$$(V_{11} \otimes V_{22} z, z) = (V_{12} \otimes V_{21} z, z) = 0,$$

consequently $V_{11} \otimes V_{22} z = V_{12} \otimes V_{21} z = 0$.

Since $(V_{11} \otimes Id_2)(Id_1 \otimes V_{22}) z = V_{11} \otimes V_{22} z$ and $V_{11} > 0$ it follows that $Id_1 \otimes V_{22} z = 0$. Similarly, $Id_1 \otimes V_{21} z = 0$. By assumption, $-V_{21} + V_{22} > 0$, hence $Id_1 \otimes (-V_{21} + V_{22}) z = 0$ implies $z = 0$. Finally, case (6) is similar to case (5). □

We need some further lemmas.

LEMMA 3.3. *Let T and V be compact symmetric operators on a separable Hilbert space H. Let V be positive definite. Then there is a sequence of eigenpairs (λ^p, x^p) of the eigenvalue problem*

$$x = Tx + \lambda Vx, \qquad 0 \neq x \in H \tag{3.6}$$

such that x^p is total in H.

Proof. By Lemma 2.5, we may assume, without loss of generality, that $Id - T$ is boundedly invertible. By Theorem 2.4 with $k = 1$, there exists a sequence of eigenpairs (λ^p, x^p) of (3.6) such that x^p forms an orthonormal basis in the pre-Hilbert space $(H, [\ ,\ ])$ where $[x, y] = (Vx, y)$ and $(\ ,\ )$ denotes the original inner product in H. The sequence x^p is total in $(H, [\ ,\ ])$ and it remains to show that x^p is total in H, also. Since x^p is total in $(H, [\ ,\ ])$ and $V^{1/2}$ is an isometry from $(H, [\ ,\ ])$ onto $(V^{1/2}(H), (\ ,\ ))$ the sequence $V^{1/2}x^p$ is total in $(V^{1/2}(H), (\ ,\ ))$. Now, $V^{1/2}$ is continuous on H and hence $Vx^p = V^{1/2}(V^{1/2}x^p)$ is total in $(V(H), (\ ,\ ))$. It follows that Vx^p is total in H because the range $V(H)$ of the positive definite operator V is dense in H. By assumption, the operator $Id - T$ is boundedly invertible, in particular, the eigenvalues λ^p are different from zero. It follows that $x^p = \lambda^p(Id - T)^{-1}Vx^p$ is total in H. □

LEMMA 3.4. *Let A and B be bounded symmetric operators on a Hilbert space H such that $0 \leqq A \leqq B$. Then the range of A is contained in the range of the square root $B^{1/2}$ of B.*

For the proof see [12, Hilfssatz 8].

LEMMA 3.5. *Assume that the eigenvalue problem* (3.1) *has the canonical form described in Lemma* 3.1.

Then there is a total subset of decomposable tensors $y_1 \otimes y_2$ in H which satisfy

$$\Delta_1(y_1 \otimes y_2) \in \Delta_0^{1/2}(H). \tag{3.7}$$

Proof. We give the proof in the cases (1), (3) and (5). In the other three cases the proof is similar and will be omitted. Under our assumptions, the operator $-V_{21} + V_{22}$ is positive definite. Hence, by Lemma 3.3, there are sequences λ_2^p and x_2^p such that

$$x_2^p = T_2x_2^p + \lambda_2^p(-V_{21} + V_{22})x_2^p \tag{3.8}$$

and the x_2^p are total in H_2. Now consider the one-parameter eigenvalue problem

$$x_1 = (T_1 + \lambda_2^p V_{12})x_1 + \lambda_1 V_{11}x_1$$

where p is fixed. Since V_{11} is positive definite we can apply Lemma 3.3 to this problem. We obtain sequences λ_1^{pq} and x_1^{pq} such that

$$x_1^{pq} = (T_1 + \lambda_2^p V_{12})x_1^{pq} + \lambda_1^{pq} V_{11}x_1^{pq} \tag{3.9}$$

and the vectors x_1^{pq}, $q = 1, 2, \cdots$, are total in H_1 for every p. Now (3.8), (3.9) and

$$\Delta_1 = (Id_1 - T_1) \otimes V_{22} - V_{12} \otimes (Id_2 - T_2)$$

give

$$\begin{aligned}\Delta_1(x_1^{pq} \otimes x_2^p) &= \lambda_1^{pq} V_{11}x_1^{pq} \otimes V_{22}x_2^p + \lambda_2^p V_{12}x_1^{pq} \otimes V_{21}x_2^p \\ &\in (V_{11} \otimes V_{22})(H) + (V_{12} \otimes V_{21})(H).\end{aligned}$$

Since $0 \leqq V_{11} \otimes V_{22} \leqq \Delta_0$ and $0 \leqq -V_{12} \otimes V_{21} \leqq \Delta_0$ it follows from Lemma 3.4 that

$$(V_{11} \otimes V_{22})(H) \subseteq \Delta_0^{1/2}(H), \qquad (V_{12} \otimes V_{21})(H) \subseteq \Delta_0^{1/2}(H).$$

This proves that $\Delta_1(x_1^{pq} \otimes x_2^p)$ is in the range of $\Delta_0^{1/2}$ for all p and q. Since x_2^p, $p = 1, 2, \cdots$, is total in H_2 and x_1^{pq}, $q = 1, 2, \cdots$, is total in H_1 for every p the set of tensors $x_1^{pq} \otimes x_2^p$, $p, q = 1, 2, \cdots$, is total in $H = H_1 \otimes H_2$. The proof of the last statement is easy and similar to the proof of [13, Thm. 2.3]. This completes the proof. □

Now we are in a position to prove the main result of this section.

THEOREM 3.6. *There exists an orthonormal basis in the pre-Hilbert-space $(H_1 \otimes H_2, [\ ,\ ])$ consisting of eigenvectors of the two-parameter problem* (3.1) *provided the definiteness condition* (3.2) *holds.*

Proof. Without loss of generality, we shall assume that (3.1) is reduced to one of the six canonical forms of Lemma 3.1. Let $y_1 \otimes y_2$ be a tensor in H which satisfies

(3.7). Now we choose sequences of orthoprojectors P_m^j as at the beginning of the second section. Then we apply Lemma 2.1 to the functional

$$f(\lambda)=\lambda_1^2+\lambda_2. \tag{3.10}$$

Since $\Delta_2>0$ it follows from (2.12) that $\lambda_2>0$ whenever $(\lambda_1,\lambda_2)\in Q$. Hence the functional (3.10) satisfies (i) of Lemma 2.1. Now Theorem 2.3 (i) shows that the sequence

$$\sum_{i\in I_j}|\alpha^{ij}|^2\lambda_2^{ij},\qquad j=1,2,3,\cdots$$

is bounded. Theorem 2.2 (i) and (3.7) show that

$$\sum_{i\in I_j}|\alpha^{ij}|^2|\lambda_1^{ij}|^2,\qquad j=1,2,3,\cdots$$

is bounded. Hence the hypothesis of Lemma 2.1 is satisfied and, therefore, Parseval's equality (2.4) holds for the given $y_1\otimes y_2$. This completes the proof because the set of these $y_1\otimes y_2$ is total in H by Lemma 3.5. □

4. An expansion theorem for k-parameter integral equations. In this section we consider the eigenvalue problem (1.3) under the assumption of right definiteness (1.2). The operator determinant Δ_0 is now an integral operator

$$(\Delta_0x)(\xi)=\int_\pi d_0(\xi,\eta)x(\eta)\,d\eta \tag{4.1}$$

where $\xi=(\xi_1,\cdots,\xi_k)$, $\eta=(\eta_1,\cdots,\eta_k)\in\pi=\prod_{m=1}^k[a_m,b_m]$ and

$$d_0(\xi,\eta)=\det_{1\leqq m,n\leqq k}v_{mn}(\xi_m,\eta_m).$$

The sesquilinear form

$$[x,y]=\int_\pi\int_\pi d_0(\xi,\eta)x(\eta)\overline{y(\xi)}\,d\xi\,d\eta \tag{4.2}$$

is positive semidefinite on $L^2(\pi)$. Therefore, there holds Bessel's inequality

$$\sum_p|[x^p,y]|^2\leqq[y,y] \tag{4.3}$$

for every $y\in L^2(\pi)$ and every finite system x^p which is orthonormal with respect to [,].

Now we fix a point $\tilde{\xi}$ in π and choose a sequence $y^1,y^2,\cdots$ of continuous nonnegative functions on π such that

$$\int_\pi y^q(\eta)\,d\eta=1\quad\text{and}\quad y^q(\xi)=0\quad\text{if }\|\xi-\tilde{\xi}\|\geqq\frac{1}{q}$$

where $\|\ \|$ denotes Euclidean norm. Then the representations (4.1) and (4.2) and some elementary analysis show that

$$[x,y^q]\to(\Delta_0x)(\tilde{\xi})\quad\text{for }x\in L^2(\pi),$$
$$[y^q,y^q]\to d_0(\tilde{\xi},\tilde{\xi}),$$

as $q\to\infty$. We substitute y^q for y in (4.3) and let $q\to\infty$. Then we obtain

$$\sum_p|(\Delta_0x^p)(\tilde{\xi})|^2\leqq d_0(\tilde{\xi},\tilde{\xi}). \tag{4.4}$$

This inequality can be used to prove the following expansion theorem.

THEOREM 4.1. *Let e^i, $i \in I$, be a system of eigenfunctions of* (1.3) *corresponding to the eigenvalues λ^i such that e^i is orthonormal with respect to $[\ ,\ ]$. Assume that the function y in $L^2(\pi)$ can be expanded in a Fourier series*

$$y = \sum_{i \in I} [y, e^i] e^i. \tag{4.5}$$

Then the expansion

$$\Delta_0 y = \sum_{i \in I} [y, e^i] \Delta_0 e^i \tag{4.6}$$

holds where the series converges absolutely and uniformly on π.

Proof. The Cauchy-Schwarz inequality and (4.4) show that

$$\left(\sum_{i \in J} |[y, e^i](\Delta_0 e^i)(\xi)| \right)^2 \leqq d_0(\xi, \xi) \sum_{i \in J} |[y, e^i]|^2 \tag{4.7}$$

for every finite subset J of I and every ξ in π. Now d_0 is bounded on π and $\sum_{i \in I} |[y, e^i]|^2 < \infty$ by Bessel's inequality. Hence it follows from (4.7) that, for every positive ε, there is a finite subset K of I such that

$$\sum_{i \in J} |[y, e^i](\Delta_0 e^i)(\xi)| \leqq \varepsilon$$

for every ξ in π and every finite subset J of I disjoint from K. This means that the series in (4.6) is absolutely and uniformly convergent. It is clear from (4.5) that the uniform limit of this series is $\Delta_0 y$. □

By Theorem 2.4 and Theorem 3.6 the condition (4.5) is satisfied for all y in $L^2(\pi)$ provided that Minkowski's definiteness condition holds or that $k = 2$.

We remark that Pell [11, Thm. 4] proved a theorem similar to Theorem 4.1 for a left-definite two-parameter integral equation.

REFERENCES

[1] F. V. ATKINSON, *Multiparameter Eigenvalue Problems*, Vol. 1, Academic Press, New York, 1972.

[2] P. BINDING, *Another positivity result for determinantal operators*, Proc. Roy. Soc. Edinburgh Sect. A, 86 (1980), pp. 333-337.

[3] ———, *Nonuniform right definiteness*, J. Math. Anal. Appl., 102 (1984), pp. 233-243.

[4] P. BINDING AND P. J. BROWNE, *A definiteness result for determinantal operators*, in Ordinary Differential Equations and Operators, Lecture Notes in Mathematics 1032, 1983, pp. 17-30.

[5] P. BINDING, A. KÄLLSTRÖM AND B. D. SLEEMAN, *An abstract multiparameter spectral theory*, Proc. Roy. Soc. Edinburgh Sect. A, 92 (1982), pp. 193-204.

[6] P. J. BROWNE, *Abstract multiparameter theory* I, J. Math. Anal. Appl., 60 (1977), pp. 259-273.

[7] H. O. CORDES, *Der Entwicklungssatz nach Produkten bei singulären Eigenwertproblemen partieller Differentialgleichungen, die durch Separation zerfallen*, Nachr. Akad. Wiss. Göttingen, (1954), pp. 51-69.

[8] M. FAIERMAN, *The completeness and expansion theorem associated with the multiparameter eigenvalue problem in ordinary differential equations*, J. Differential Equations, 5 (1969), pp. 197-213.

[9] A. KÄLLSTRÖM AND B. D. SLEEMAN, *Solvability of a linear operator system*, J. Math. Anal. Appl., 55 (1976), pp. 785-793.

[10] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Sphäroid-funktionen*, Springer, Berlin, 1954.

[11] A. PELL, *Linear equations with two parameters*, Trans. Amer. Math. Soc., 23 (1922), pp. 198-211.

[12] F. RELLICH, *Störungstheorie der Spektralzerlegung* V, Math. Ann., 118 (1941/43), pp. 462-484.

[13] B. D. SLEEMAN, *Multiparameter Theory in Hilbert Space*, Pitman Press, London, 1978.

[14] H. VOLKMER, *On the completeness of eigenvectors of right definite multiparameter problems*, Proc. Roy. Soc. Edinburgh Sect. A, 96 (1984), pp. 69-78.

[15] ———, *On multiparameter theory*, J. Math. Anal. Appl., 86 (1982), pp. 44-53.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

014

A PERIODIC WAVE AND ITS STABILITY TO A CIRCULAR CHAIN OF WEAKLY COUPLED OSCILLATORS*

YOSHIHISA MORITA†

Abstract. A circular chain of weakly coupled oscillators with nearest neighbor and isotropic coupling is considered. The identical oscillator is represented by an ordinary differential equation which has a stable limit cycle bifurcating from a steady state. This equation of coupled oscillators has two parameter and a high degenerate singularity for a parameter value. Studying a bifurcation problem around the singularity shows that periodic traveling wave solutions, together with a homogeneous periodic one, bifurcate from the steady state for suitable parameter values. Furthermore, in a certain parameter region, a classification of stability for those solutions is presented. These results can also be extended to a class of retarded functional differential equations.

Key words. periodic wave, weakly coupled oscillators, bifurcation, center manifold, linearized stability

AMS(MOS) subject classifications. 34C15, 34C25, 34D05, 34K15

1. Introduction. We are able to observe a variety of oscillatory phenomena in models of coupled oscillators arising in the fields of biology, biochemical, electric circuits, etc. For instance, synchronization (or entrainment), phase-locking, periodic wave, chaos and so forth are found in many works including [3], [10], [11], [15]-[19].

In this paper, as a model of coupled oscillators, we shall study a circular chain of n weakly coupled equation which is represented by

$$(1.1)\qquad \begin{aligned}&\frac{d}{dt}u_k(t)=F(\mu,u_k)+\nu\{N(u_{k-1}-u_k)+N(u_{k+1}-u_k)\},\qquad k=0,1,\cdots,n-1,\\ &u_{-1}=u_{n-1},\qquad u_n=u_0\end{aligned}$$

where $u_k\in\mathbf{R}^m(m\geqq 2)$, ν is nonnegative parameter. We assume that $F:I_0\times\mathbf{R}^m\to\mathbf{R}^m$($I_0$: an interval containing the origin) and $N:\mathbf{R}^m\to\mathbf{R}^m$ are sufficiently smooth mappings satisfying $F(\mu,0)=0$ and $N(0)=0$, respectively. The coupling of (1.1) is nearest neighbor and isotropic. Moreover it is assumed that in the absence of the coupling each component of (1.1), that is,

$$(1.2)\qquad \frac{d}{dt}u(t)=F(\mu,u(t)),$$

represents an oscillator; more precisely we assume that there exists a family of periodic solutions for small $\mu>0$, such that they bifurcate from the steady state, $u=0$, at $\mu=0$. This bifurcation (called Hopf bifurcation) occurs under the condition that the linearized equation of (1.2) around $u=0$ has a pair of simple conjugate eigenvalues, $\lambda_0(\mu)$ and $\overline{\lambda_0(\mu)}$, satisfying $\lambda_0(0)=i\omega_0$ and $\operatorname{Re}(d\lambda_0/d\mu)(0)>0$.

Here we shall approach (1.1) along the lines of studying a bifurcation problem associated with (1.1) under the above conditions. We first see that at $(\mu,\nu)=(0,0)$ the equation (1.1) has high degenerate singularity, that is, the linearized equation around $u_k=0$, $k=0,\cdots,n-1$, has a pair of conjugate pure imaginary eigenvalues, $\pm i\omega_0$, with multiplicity n. A structure of bifurcation around such a singularity may be complicated and it is not easy to make clear the complete bifurcation picture. Therefore, in this

* Received by the editors February 18, 1986; accepted for publication (in revised form) December 1, 1986.

† Research Institute for Mathematical Sciences, Kyoto University, Kitashirakawa, Sakyo-Ku, Kyoto 606, Japan.

article, as the first step of the investigation, we shall study the primary bifurcation from the steady state and the stability of the bifurcating solution in a neighborhood of $(\mu, \nu) = (0, 0)$.

It is clear that for any ν and small $\mu > 0$ (1.1) has a homogeneous (or synchronized) periodic solution whose components are given by the periodic solution of (1.2). Moreover we will show the existence of open $n-1$ periodic solutions bifurcating from the steady state. Those $n-1$ periodic solutions, $\mathbf{u}^l = (u_0^l, \cdots, u_{n-1}^l)$, $l = 1, \cdots, n-1$, can be expressed as follows:

$$(1.3) \qquad u_k^l(t) = \phi^l\left(t + lT^l \cdot \frac{k}{n}\right), \qquad k = 0, \cdots, n-1$$

where $\phi^l(t)$ is a periodic function with period T^l, and satisfies $\phi^{n-l}(t) = \phi^l(t)$. This expression is due to symmetry associated with (1.1), that is, covariance of the vector field of (1.1) with respect to the transformations $u_k \to u_{k+1} (k = 0, \cdots, n-1)$ and $u_k \to u_{n-k} (k = 0, \cdots, n-1)$; for the details see §§ 3 and 4. By (1.3) we have

$$(1.4) \qquad u_{k+1}^l(t) = u_k^l\left(t + \frac{lT^l}{n}\right), \quad u_{k+1}^{n-l}(t) = u_k^{n-l}\left(t - \frac{lT^l}{n}\right), \quad k = 0, \cdots, n-1,$$

so these solutions are called periodic (traveling) wave solutions to (1.1). Note that the homogeneous periodic solution satisfies (1.3) (or (1.4)) for $l = 0$; we may call it a solution of (1.3) for $l = 0$.

Next we shall state the stability of the above solutions. Let us consider the situation that all the periodic wave solutions are unstable sufficiently near the corresponding bifurcation points except for the homogeneous one. This implies that for fixed $\nu > 0$, in μ increasing, the first bifurcation from the steady state occurs for the homogeneous solution (at $\mu = 0$). Then there are two typical cases classifying the stability of the solution in a domain of (μ, ν) satisfying $\nu/\mu \ll 1$. One case is that the solutions of (1.3) for l and $n-l, 0 \leqq l \leqq n/4$ (resp. $n/4 < l \leqq n/2$) are unstable (resp. stable) in the domain. The other is that those for l and $n-l, 0 \leqq l < n/4$ (resp. $n/4 \leqq l \leqq n/2$) are stable (resp. unstable) (see Theorem 5.2 in § 5). Hence we can see the stability change of periodic waves. In fact, as we vary μ or ν, we might observe several (or a few) times of secondary bifurcation along each branch of the periodic solution. In spite of the complex feature of bifurcation, as mentioned above, we can classify the stability of the primary bifurcating periodic solutions in the domain satisfying $\nu/\mu \ll 1$.

Our program to obtain these results is along the following lines. In § 2 we reduce the equation (1.1) on a center manifold constructed around the singularity of $(\mu, \nu, \mathbf{u}) = (0, 0, \mathbf{0})$, and then we transform this equation into a simpler one (called the reduced equation) by a nonlinear transformation. Considering the symmetry of this equation, we can obtain the periodic solutions satisfying (1.3) by the standard Lyapunov-Schmidt method (see § 4). In § 5 we shall discuss the linearized stability of (1.3). By virtue of the reduced equation, it is possible to do this. The reader will see an application to a specific equation in § 6. As a further application, we can extend our results to some class of retarded functional differential equations, which is described in § 7.

We remark that several kinds of coupled oscillators involve abundant phenomena as first described, while there are many unsolved mathematical problems. For the equation (1.1) a further step in its study is expected in the future.

We also find in [3], [15], [17], [18] that another approach to a class of the equation (1.1) is possible by investigating an equation on an invariant n-dimensional torus associated with the original one. This method has an advantage in that we can apply a large amplitude periodic solution of (1.2) to (1.1), while difficulty for an analytic expression of the periodic solution restricts its application to specific examples. Moreover it is difficult to apply it to the retarded functional differential equation as in § 7. (We also refer to [4] as a related work.)

This study was motivated by the book [11]. The reduced equation (4.19) in § 4 is proposed as a discrete space version of the Ginzburg–Landau equation which is obtained by a formal perturbation method in the book.

2. The reduced equation on a center manifold. Let $A(\mu)\equiv(\partial F/\partial u)(\mu,0)$ and $D\equiv(dN/du)(0)$. As described in § 1, we assume that $A(\mu)$ has a pair of simple eigenvalues, $\lambda_0(\mu)$ and $\overline{\lambda_0(\mu)}$, satisfying $\lambda_0(0)=i\omega_0$ and that all the remaining eigenvalues have negative real parts for $\mu=0$. We denote an eigenvector of $A(\mu)$ (resp. ${}^tA(\mu)$) corresponding to $\lambda_0(\mu)$ (resp. $\overline{\lambda_0(\mu)}$) by $\zeta_0(\mu)$ (resp. $\zeta_0^*(\mu)$), where ${}^tA(\mu)$ is the transpose of $A(\mu)$. Let $(\cdot,\cdot)$ be the hermite product in $\mathbf{C}^m$. We simply write $\zeta_0\equiv\zeta_0(0)$ and $\zeta_0^*\equiv\zeta_0^*(0)$; moreover we may assume that $(\zeta_0,\zeta_0^*)=1$ by normalizing it.

The equation (1.1) can be written as follows:

$$\begin{aligned}&\dot u_k=A(0)u_k+G(\mu,u_k)+\nu\tilde N_k(\mathbf{u}),\qquad k=0,1,\cdots,n-1,\\ &\dot\mu=0,\\ &\dot\nu=0\end{aligned}\tag{2.1}$$

where $\cdot$ denotes d/dt, and

$$\begin{aligned}&G(\mu,u)\equiv F(\mu,u)-A(0)u,\\ &\tilde N_k(\mathbf{u})\equiv N(u_{k-1}-u_k)+N(u_{k+1}-u_k),\qquad k=0,\cdots,n-1.\end{aligned}\tag{2.2}$$

Note that $(\partial/\partial u)G(0,0)=0$ and $(\partial^j/\partial\mu^j)G(\mu,0)=0, j=1,2,\cdots$. It is always understood that $u_{-1}=u_{n-1}$ and $u_n=u_0$, even if they are not written explicitly.

Using the projections, P and Q, defined by

$$u_k^P=Pu_k\equiv(u_k,\zeta_0^*)\zeta_0+(u_k,\bar\zeta_0^*)\bar\zeta_0,\qquad u_k^Q=Qu_k\equiv u-u_k^P,$$

we have the decomposition of (2.1) as follows:

$$\begin{aligned}&\dot u_k^P=A(0)u_k^P+PG(\mu,u_k^P+u_k^Q)+\nu P\tilde N_k(\mathbf{u}^P+\mathbf{u}^Q),\\ &\dot\mu=0,\\ &\dot\nu=0,\end{aligned}\tag{2.3a}$$

$$\dot u_k^Q=A(0)u_k^Q+QG(\mu,u_k^P+u_k^Q)+\nu Q\tilde N_k(\mathbf{u}^P+\mathbf{u}^Q)\tag{2.3b}$$

where

$$\mathbf{u}^P\equiv\begin{pmatrix}u_0^P\\ \vdots\\ u_{n-1}^P\end{pmatrix},\qquad \mathbf{u}^Q\equiv\mathbf{u}-\mathbf{u}^P.$$

All the eigenvalues of the linear part of (2.3a) (resp. (2.3b)) have zero real parts (resp. negative real parts) by the above assumption. Therefore we easily see that the center manifold theorem can apply to (2.3) (refer to [1]). This theorem implies that there exists a local invariant manifold in a neighborhood of $(\mu,\nu,\mathbf{u})=(0,0,\mathbf{0})$, and it is tangent to the space of $(\mu,\nu,\mathbf{u}^P)=(0,0,\mathbf{0})$. Moreover we see from the attractivity of the center manifold that asymptotic behaviors of all solutions of (2.3) with small amplitudes are determined by an equation which is reduced on this manifold.

Now we adopt complex variables, $z_k=(u_k, \zeta_0^*)$, $k=0, 1, \cdots, n-1$ (or $\mathbf{z}={}^t(z_0, \cdots, z_{n-1})$). Then the center manifold is represented by a sufficiently smooth mapping, $u_k^Q=h_k(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})$, $k=0, \cdots, n-1$, defined in a neighborhood of $(\mu, \nu, \mathbf{z})=(0, 0, \mathbf{0})$, and the equation on the manifold is given by

$$\begin{aligned}\dot z_k &= i\omega_0 z_k+(G(\mu, z_k\zeta_0+\overline{z_k\zeta_0}+h_k(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})), \zeta_0^*)\\ &\quad+\nu(\tilde N_k(\mathbf{z}\zeta_0+\overline{\mathbf{z}\zeta_0}+\mathbf{h}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})), \zeta_0^*),\\ z_{-1}&=z_{n-1}, \qquad z_n=z_0\end{aligned} \tag{2.4}$$

where $\mathbf{h}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})={}^t(h_0(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}), \cdots, h_{n-1}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}))$ and $\mathbf{z}\zeta_0$ denotes ${}^t(z_0\zeta_0, \cdots, z_{n-1}\zeta_0)$. We note that $\overline{h_k(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})}=h_k(\mu, \nu, \bar{\mathbf{z}}, \mathbf{z})$, $h_{-1}(\cdot)=h_{n-1}(\cdot)$, $h_n(\cdot)=h_0(\cdot)$ and $h_k(0, 0, \mathbf{0}, \mathbf{0})=0$; moreover the first derivatives, together with $(\partial^j/\partial^{j_1}\mu\partial^{j_2}\nu)h_k, j=j_1+j_2\geqq 1$, vanish for $(\mu, \nu, \mathbf{z})=(0, 0, \mathbf{0})$.

We let

$$C_1\equiv(D\zeta_0, \zeta_0^*), \qquad C_2\equiv(D\bar\zeta_0, \zeta_0^*), \tag{2.5}$$

and we can define

$$h(\mu, z_k, \bar z_k)\equiv h_k(\mu, 0, \mathbf{z}, \bar{\mathbf{z}}) \tag{2.6}$$

for any $k, 0\leqq k\leqq n-1$; virtually, $h(\mu, \cdot, \cdot)$ represents a center manifold of (1.2) constructed around $(\mu, u)=(0, 0)$. Then (2.4) is written as

$$\begin{aligned}\dot z_k &= i\omega_0 z_k+(G(\mu, z_k\zeta_0+\overline{z_k\zeta_0}+h(\mu, z_k, \bar z_k)), \zeta_0^*)\\ &\quad+\nu C_1(z_{k-1}-2z_k+z_{k+1})+\nu C_2(\bar z_{k-1}-2\bar z_k+\bar z_{k+1})+R_k^1(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}),\\ R_k^1&=O(\nu\{|(\mathbf{z}, \bar{\mathbf{z}})|^2+|\mu(\mathbf{z}, \bar{\mathbf{z}})|\}).\end{aligned} \tag{2.7}$$

The Taylor expansion of (2.7) is as follows:

$$\begin{aligned}\dot z_k &= i\omega_0 z_k+g(\mu, z_k, \bar z_k)+\nu C_1(z_{k-1}-2z_k+z_{k+1})\\ &\quad+\nu C_2(\bar z_{k-1}-2\bar z_k+\bar z_{k+1})+R_k^2(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}),\end{aligned} \tag{2.8}$$

$$\begin{aligned}g(\mu, z_k, \bar z_k) &= \mu(a_{10}z_k+a_{01}\bar z_k)+a_{20}z_k^2+a_{11}|z_k|^2+a_{02}\bar z_k^2\\ &\quad+a_{30}z_k^3+a_{21}|z_k|^2z_k+a_{12}|z_k|^2\bar z_k+a_{03}\bar z_k^3,\end{aligned} \tag{2.9}$$

$$R_k^2(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})=R_k^1+O(|(z_k, \bar z_k)|\{|(z_k, \bar z_k)|^3+|\mu(\mu, z_k, \bar z_k)|\}) \tag{2.10}$$

where the coefficients, $a_{ij}(i, j=0, 1, 2, 3)$ are given by $(\alpha 4)$ in Appendix A.

Now we shall transform equation (2.8) into a simpler one. Since the coefficients of g depend only on $F(\cdot, \cdot)$ (independent of ν and $N(\cdot)$), we can use the nonlinear transformation found in [8], that is,

$$z_k=w_k+\mu b_{01}\bar w_k+b_{20}w_k^2+b_{11}|w_k|^2+b_{02}\bar w_k^2+b_{03}w_k^3+b_{12}|w_k|^2\bar w_k+b_{30}\bar w_k^3, \tag{2.11}$$

whose coefficients are determined by a_{ij}, $i, j=0, 1, 2, 3$, in an appropriate manner. By (2.11), the equation (2.8) is transformed into

$$\begin{aligned}\dot w_k &= i\omega_0 w_k+\mu A_1 w_k+B_1|w_k|^2w_k+\nu C_1(w_{k-1}-2w_k+w_{k+1})\\ &\quad+\nu C_2(\bar w_{k-1}-2\bar w_k+\bar w_{k+1})+R_k^3(\mu, \nu, \mathbf{w}, \bar{\mathbf{w}})\end{aligned} \tag{2.12}$$

where

$$A_1\equiv a_{10}=\frac{d\lambda_0}{d\mu}(0) \qquad \text{(by } (\alpha 1) \text{ and } (\alpha 4) \text{ in Appendix A),} \tag{2.13}$$

$$B_1 \equiv a_{21} + \frac{i}{\omega_0} a_{20} a_{11} - \frac{i}{\omega_0} |a_{11}|^2 - \frac{2i}{3\omega_0} |a_{02}|^2, \tag{2.14}$$

and R_k^3 is of the same order as R_k^2 in (2.10).

Further calculation in Appendix A shows a formula about the coefficient, B_1, i.e.,

$$B_1 = (F_{uu}(0)(\zeta_0, \hat{\zeta}_2), \zeta_0^*) + (F_{uu}(0)(\bar{\zeta}_0, \zeta_2), \zeta_0^*) + \tfrac{1}{2}(F_{uuu}(0)(\zeta_0, \zeta_0, \bar{\zeta}_0), \zeta_0^*) \tag{2.15}$$

where ζ_2 and $\hat{\zeta}_2$ are defined by

$$\zeta_2 = (2i\omega_0 - A(0))^{-1} \tfrac{1}{2} F_{uu}(0)(\zeta_0, \zeta_0), \qquad \hat{\zeta}_2 = -A(0)^{-1} F_{uu}(0)(\zeta_0, \bar{\zeta}_0) \tag{2.16}$$

(note that $F_{uu}(0) \equiv \partial^2/\partial u^2 F(0,0)$ and $F_{uuu}(0) \equiv \partial^3/\partial u^3 F(0,0)$).

Applying the following transformation to (2.12):

$$w_k = z_k - \frac{\nu C_2}{2i\omega_0} (\bar{z}_{k-1} - 2\bar{z}_k + \bar{z}_{k+1}), \tag{2.17}$$

we obtain the equation with a simpler form:

$$\begin{aligned} \dot{z}_k &= i\omega_0 z_k + \mu A_1 z_k + B_1 |z_k|^2 z_k + \nu C_1 (z_{k-1} - 2z_k + z_{k+1}) + R_k^4(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}), \\ z_{-1} &= z_{n-1}, \qquad z_n = z_0 \end{aligned} \tag{2.18}$$

where R_k^4 also has the same order as R_k^2. (Note that $\overline{R_k^4(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})} = R_k^4(\mu, \nu, \bar{\mathbf{z}}, \mathbf{z})$.) The equation (2.18) is expressed in vector form as

$$\begin{aligned} \dot{\mathbf{z}} &= i\omega_0 \mathbf{z} + \mu A_1 \mathbf{z} + B_1 \mathbf{V}(\mathbf{z}, \bar{\mathbf{z}}) \mathbf{z} + \nu C_1 \Delta_d \mathbf{z} + \mathbf{R}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}) \\ &\equiv \mathbf{f}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}), \end{aligned} \tag{2.19}$$

$$\mathbf{V}(\mathbf{z}, \bar{\mathbf{z}}) \equiv \begin{pmatrix} |z_0|^2 & & \\ & \ddots & \\ & & |z_{n-1}|^2 \end{pmatrix},$$

$$\Delta_d \equiv \begin{pmatrix} -2 & 1 & 0 & \cdot & 0 & 1 \\ 1 & -2 & 1 & 0 & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & 0 & \cdot & 0 & 1 & -2 \end{pmatrix}, \qquad \mathbf{R}(\cdot) \equiv \begin{pmatrix} R_0^4(\cdot) \\ \vdots \\ R_{n-1}^4(\cdot) \end{pmatrix}. \tag{2.20}$$

Hence we obtain the next lemma.

Lemma 2.1. *Consider the equation* (1.1). *The asymptotic behavior of any solution to* (1.1) *having small amplitude in a neighborhood of* $(\mu, \nu, \mathbf{u}) = (0, 0, \mathbf{0})$ *is determined by the* $2n$*-dimensional system of* (2.18) (*or* (2.19)). *This reduced equation has the symmetry such that*

$$S_j \mathbf{f}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}) = \mathbf{f}(\mu, \nu, S_j \mathbf{z}, \overline{S_j \mathbf{z}}), \qquad j = 1, 2 \tag{2.21}$$

where

$$S_1 \equiv \begin{pmatrix} 0 & 1 & 0 & \cdot & 0 \\ \cdot & 0 & 1 & 0 & \cdot \\ \cdot & & \cdot & \cdot & 0 \\ 0 & 0 & & \cdot & 1 \\ 1 & 0 & \cdot & \cdot & 0 \end{pmatrix}, \qquad S_2 \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & 1 \\ \vdots & & \iddots & \\ & & & \\ 0 & 1 & & 0 \end{pmatrix}. \tag{2.22}$$

The latter part of the lemma about the symmetry will be shown in the next section.

Remark 2.1. In the case of $m=2$, it is clear that the center manifold is not necessary to get the above reduced equation (2.19). For the case of $m \geqq 3$, we also see from (2.13) and (2.15) that the leading terms of (2.19) can be obtained by no explicit computation about the center manifold.

Remark 2.2. In § 7 the above argument will be extended to a case of coupled oscillators represented by a functional differential equation. Then an equation quite similar to (2.18) will be obtained together with a formula for its coefficients corresponding to A_1, B_1 and C_1 in (2.18) (see (7.10)).

3. Symmetry of the reduced equation. We show (2.21) in Lemma 2.1. Let

$$\mathbf{F}(\mu, \nu, \mathbf{u}) \equiv \begin{pmatrix} F(\mu, u_0) + \nu \tilde{N}_0(\mathbf{u}) \\ \vdots \\ F(\mu, u_{n-1}) + \nu \tilde{N}_{n-1}(\mathbf{u}) \end{pmatrix}$$

where $\tilde{N}_k$ is defined in (2.2). By this definition of $\mathbf{F}$,

$$\Sigma_j \mathbf{F}(\mu, \nu, \mathbf{u}) = \mathbf{F}(\mu, \nu, \Sigma_j \mathbf{u}), \qquad j=1,2 \tag{3.1}$$

where

$$\Sigma_1 \mathbf{u} = \Sigma_1 \begin{pmatrix} u_0 \\ \vdots \\ u_{n-1} \end{pmatrix} \equiv \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_0 \end{pmatrix}, \qquad \Sigma_2 \mathbf{u} = \Sigma_2 \begin{pmatrix} u_0 \\ \vdots \\ u_{n-1} \end{pmatrix} \equiv \begin{pmatrix} u_0 \\ u_{n-1} \\ u_{n-2} \\ \vdots \\ u_2 \\ u_1 \end{pmatrix}.$$

Uniqueness of the solution to (1.1) and (3.1) imply that the center manifold constructed in § 2 satisfies

$$\Sigma_j \mathbf{h}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}) = \mathbf{h}(\mu, \nu, S_j \mathbf{z}, \overline{S_j \mathbf{z}}), \qquad j=1,2.$$

Hence it is easily seen that the right-hand side of (2.8), say $\tilde{\mathbf{g}}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}})$, is covariant for $S_j, j=1,2$, that is,

$$S_j \tilde{\mathbf{g}}(\mu, \nu, \mathbf{z}, \bar{\mathbf{z}}) = \tilde{\mathbf{g}}(\mu, \nu, S_j \mathbf{z}, \overline{S_j \mathbf{z}}), \qquad j=1,2.$$

This property is invariant under the transformation of (2.11). Moreover since S_j commutes with Δ_d defined in (2.20), it is also invariant under the transformation of (2.17) which is expressed in vector form as $\mathbf{w} = \mathbf{z} + \alpha \Delta_d \bar{\mathbf{z}}$, $\alpha = -\nu C_2 / 2i\omega_0$. This implies the completion of the proof of Lemma 2.1.

Next we shall define some matrices and observe how they change by a certain transformation. This observation will be useful for the later argument in § 5.

Let

$$\mathbf{P} \equiv (p_{k,m})_{k,m=0,\cdots,n-1}, \qquad p_{k,m} \equiv e^{ik \cdot 2m\pi/n} / \sqrt{n} \tag{3.2}$$

be a matrix whose kth row and mth column is $p_{k,m}$, and let

$$\mathbf{P}_l \equiv \begin{pmatrix} p_{0,l} & & 0 \\ & \ddots & \\ 0 & & p_{n-1,l} \end{pmatrix}. \tag{3.3}$$

It is easily seen that

$$\mathbf{P}^{-1}\Delta_d\mathbf{P}=-\Lambda\equiv-\begin{pmatrix}\sigma_0 & & & 0\\ & \sigma_1 & & \\ & & \ddots & \\ 0 & & & \sigma_{n-1}\end{pmatrix} \tag{3.4}$$

where

$$\sigma_k\equiv 4\sin^2\frac{k\pi}{n},\qquad k=0,\cdots,n-1. \tag{3.5}$$

Moreover since $\mathbf{P}^{-1}=\mathbf{P}^*\equiv{}^t\mathbf{P}$ and $\mathbf{P}_l\mathbf{P}=(1/\sqrt{n})(p_{k,l+m})_{k,m=0,\cdots,n-1}$, we have

$$(\mathbf{P}_l\mathbf{P})^{-1}\Delta_d(\mathbf{P}_l\mathbf{P})=-\Lambda_l\equiv-\begin{pmatrix}\sigma_l & & & 0\\ & \sigma_{l+1} & & \\ & & \ddots & \\ 0 & & & \sigma_{l+n-1}\end{pmatrix}, \tag{3.6}$$

$$(\mathbf{P}_l\mathbf{P})^{-1}\mathbf{P}_l^2\overline{\mathbf{P}_l\mathbf{P}}=\frac{1}{n}\mathbf{P}^{-1}\bar{\mathbf{P}}=\frac{1}{n}S_2 \tag{3.7}$$

(S_2 is as in (2.22)). Note that $p_{k+n,m}=p_{k,n+m}=p_{k,m}$ and $\sigma_k=\sigma_{n+k}$, so we understand that $p_{k,m}=p_{k-n,m}$, $\sigma_{k-n}=\sigma_k$ (resp. $p_{k,m}=p_{k,m-n}$) for $k\geqq n$ (resp. $m\geqq n$).

4. Existence of the periodic wave solutions. We shall consider (2.19). If we truncate R_k^4, and if Re $A_1\neq 0$, Re $B_1\neq 0$, then there exists a solution for $\mu=\mu_2^l$ such as

$$z_k^l=\varepsilon e^{i\omega_2^l t}e^{ik\cdot 2l\pi/n}\qquad (k=0,\cdots,n-1) \tag{4.1}$$

where μ_2^l and ω_2^l are defined as

$$\begin{aligned}\mu_2^l&\equiv-\frac{b_1}{a_1}\varepsilon^2+\frac{D_1\sigma_l}{a_1}\nu,\\ \omega_2^l&\equiv\omega_0+\mu_2^l a_2+b_2\varepsilon^2-\nu\sigma_l D_2,\end{aligned} \tag{4.2}$$

and

$$\begin{aligned}&a_1\equiv\operatorname{Re}A_1,\quad a_2\equiv\operatorname{Im}A_2,\quad b_1\equiv\operatorname{Re}B_1,\quad b_2\equiv\operatorname{Im}B_2,\\ &D_1\equiv\operatorname{Re}C_1,\qquad D_2\equiv\operatorname{Im}C_2\end{aligned} \tag{4.3}$$

(σ_l is as in (3.5)). We, however, have to discuss more precisely the existence of periodic solutions whose leading terms are given by (4.1), because we may never neglect the terms included in R_k^4.

To seek a periodic solution of (2.19), we shall consider

$$\omega\frac{d}{ds}\mathbf{y}(s)=\mathbf{f}(\mu,\nu,\mathbf{y}(s),\overline{\mathbf{y}(s)}),\qquad \mathbf{y}(s)\equiv\mathbf{z}\left(\frac{s}{\omega}\right), \tag{4.4}$$

on the space,

$$\mathbf{P}_{2\pi,l}\equiv\{\mathbf{y}={}^t(y_0,\cdots,y_{n-1});\ y_0(s)\text{ is a }2\pi\text{-periodic continuous function with values in }\mathbf{C}^m\text{ and }y_k(s)=y_0(s+k\cdot 2l\pi/n)\}.$$

A solution of (4.4) in $\mathbf{P}_{2\pi,l}(l\geqq 1)$ gives a traveling periodic wave solution to (1.1) mentioned in the Introduction. By virtue of the symmetry of $\mathbf{f}$ in (2.21), the equation (4.4) is well defined on the space, $\mathbf{P}_{2\pi,l}$; this yields that (4.4) can be reduced to the following 2-dimensional equation with respect to $y_0(s)$ on a space of 2π-periodic functions:

$$\omega \frac{d}{ds} y_0(s) = f_0\left(\mu, \nu, y_0(s), \cdots, y_0\left(s+(n-1)\cdot \frac{2l\pi}{n}\right), \overline{y_0(s)}, \cdots\right). \tag{4.5}$$

Applying the standard Lyapunov-Schmidt method to (4.5), we obtain the following theorem.

THEOREM 4.1. *Consider* (2.19) *under the assumption,*

$$\text{Re}\, \frac{d\lambda_0}{d\mu}(0) = \text{Re}\, A_1 > 0. \tag{A1}$$

Then for sufficiently small positive numbers, $\bar{\varepsilon}$ and $\bar{\nu}$, there exist smooth functions, $\mu = \mu^l(\varepsilon, \nu)$ and $\omega = \omega^l(\varepsilon, \nu)$ $(l = 0, \cdots, n-1)$, defined in $I_{\varepsilon,\nu} \equiv (0, \bar{\varepsilon}) \times (-\bar{\nu}, \bar{\nu})$, such that

$$\mu^l(\varepsilon, \nu) = \mu_2^l(\varepsilon, \nu) + \hat{\mu}^l(\varepsilon, \nu), \qquad \omega^l(\varepsilon, \nu) = \omega_0 + \omega_2^l(\varepsilon, \nu) + \hat{\omega}^l(\varepsilon, \nu), \tag{4.6}$$

where $\mu_2^l(\varepsilon, \nu)$ and $\omega_2^l(\varepsilon, \nu)$ satisfy (4.2), *and $\hat{\mu}^l(\varepsilon, \nu) = O(|(\varepsilon, \nu)||(\varepsilon^2, \nu)|) = \hat{\omega}^l(\varepsilon, \nu)$. For each $\mu = \mu^l(\varepsilon, \nu)$ and $(\varepsilon, \nu) \in I_{\varepsilon,\nu}$, there exists a periodic solution of* (2.19) *bifurcating from the origin with period, $T^l(\varepsilon, \nu) \equiv 2\pi/\omega^l(\varepsilon, \nu)$.*

This bifurcating periodic solution is expressed as follows:

$$\mathbf{z}^l(t) = \begin{pmatrix} z_0^l(t) \\ \vdots \\ z_{n-1}^l(t) \end{pmatrix} = \varepsilon\, e^{i\omega^l(\varepsilon,\nu)t} \begin{pmatrix} 1 \\ e^{i(2l\pi/n)} \\ \vdots \\ e^{i(n-1)\cdot 2l\pi/n} \end{pmatrix} + \mathbf{w}^l(t), \tag{4.7}$$

$$\mathbf{w}^l(\cdot) = O(|(\varepsilon, \nu)|^2),$$

and satisfies, $z_k^l(t) = z_0^l(\omega^l(\varepsilon, \nu)t + k\cdot 2l\pi/n)$, $k = 1, \cdots, n-1$. Furthermore $\mu^{n-l}(\varepsilon, \nu) = \mu^l(\varepsilon, \nu)$, $\omega^{n-l}(\varepsilon, \nu) = \omega^l(\varepsilon, \nu)$ and $z_k^{n-l}(t) = z_{n-k}^l(t)$, $k = 0, \cdots, n-1$ (by a suitable phase shift) hold for l, $1 \leqq l \leqq n/2$.

We shall omit the proof of the theorem, since it will be easily shown by the standard Hopf bifurcation theory. For example, we refer to the works [5], [8], [9]. We only note that $z_k^{n-l}(t) = z_{n-k}^l(t)$, $k = 0, \cdots, n-1$, is shown by the symmetry of $\mathbf{f}$ with respect to S_2.

The next corollary follows immediately from the above theorem.

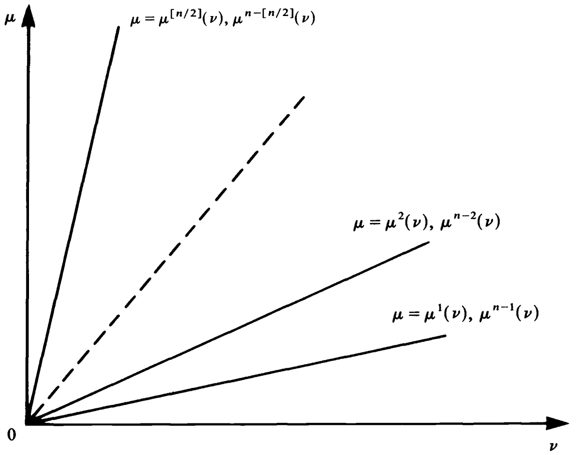
COROLLARY 4.2. *Consider* (1.1) *under the same condition of Theorem* 4.1. *Then, for $\mu = \mu^l(\varepsilon, \nu)$, $(\varepsilon, \nu) \in I_{\varepsilon,\nu}$ and $1 \leqq l \leqq n/2$, there exist traveling periodic wave solutions, $\mathbf{u}^l = (u_0^l, \cdots, u_{n-1}^l)$ and $\mathbf{u}^{n-l} = (u_0^{n-l}, \cdots, u_{n-1}^{n-l})$ satisfying* (1.4), *where $\mu^l(\varepsilon, \nu)$ and $T^l(\varepsilon, \nu)$ are as in Theorem* 4.1. *Moreover for each $l = 1, \cdots, n/2$, $u_0^l(t) = u_0^{n-l}(t + \rho)$, where ρ is a suitable shift in phase.*

Hereafter we will assume that

$$\text{Re}\, B_1 = b_1 < 0, \tag{A2}$$

in addition to (A1). The assumption (A2) implies that bifurcation occurs supercritically with respect to μ, that is, each periodic solution exists for $\mu > \mu^l(0, \nu)$. In Figs. 1-3, we find bifurcation diagrams of this case with an additional condition $\text{Re}\, C_1 = D_1 > 0$.

5. Stability of the periodic wave solutions. We shall discuss the stability of the periodic solutions to (1.1) obtained in the previous section. By virtue of the attractivity of the center manifold, it is enough to investigate the solution of (4.7) to (2.19). The reduced form of (2.19) will play a central role in the argument below.



μ
$\mu=\mu^{[n/2]}(\nu),\ \mu^{n-[n/2]}(\nu)$
$\mu=\mu^{2}(\nu),\ \mu^{n-2}(\nu)$
$\mu=\mu^{1}(\nu),\ \mu^{n-1}(\nu)$
0
ν

FIG. 1. *Bifurcation lines,* $\mu^l(\nu)=\mu^l(0,\nu)$, *are figured for the case,* $D_1=\text{Re }C_1>0$. *The periodic wave solutions,* $\mathbf{u}^l$ and $\mathbf{u}^{n-l}$, *in Corollary* 4.2 *bifurcate at* $\mu=\mu^l(\nu)$.

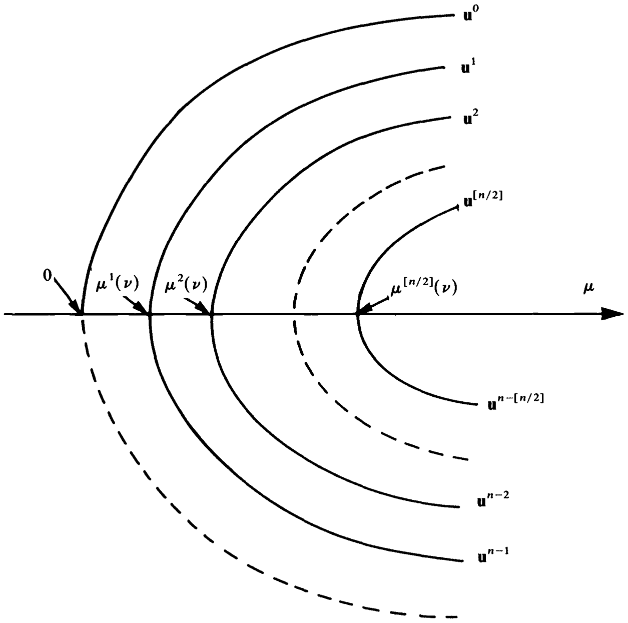


FIG. 2. *A bifurcation diagram for fixed* $\nu>0$, *under the same condition as in Fig.* 1. *Vertical direction indicates a function space schematically. The branches of* $\mathbf{u}^l$ *and* $\mathbf{u}^{n-l}$ *are drawn symmetric with respect to the horizontal axis, which symbolizes that* $\mathbf{u}^l$ *and* $\mathbf{u}^{n-l}$ *are symmetric each other in the sense of Corollary* 4.2.

The linearized equation around the solution (4.7) is given by:

$$\begin{aligned}\dot{y}_k&=(i\omega_0+\mu^lA_1+2B_1|z_k^l(t)|^2)y_k+B_1\{z_k^l(t)\}^2\bar{y}_k\\ &\quad+\nu C_1(y_{k-1}-2y_k+y_{k+1})+\frac{\partial}{\partial\mathbf{z}}R_k^4(\mu^l,\nu,\mathbf{z}^l(t),\bar{\mathbf{z}}^l(t))\mathbf{y}\qquad(5.1)\\ &\quad+\frac{\partial}{\partial\bar{\mathbf{z}}}R_k^4(\mu^l,\nu,\mathbf{z}^l(t),\bar{\mathbf{z}}^l(t))\bar{\mathbf{y}}\end{aligned}$$

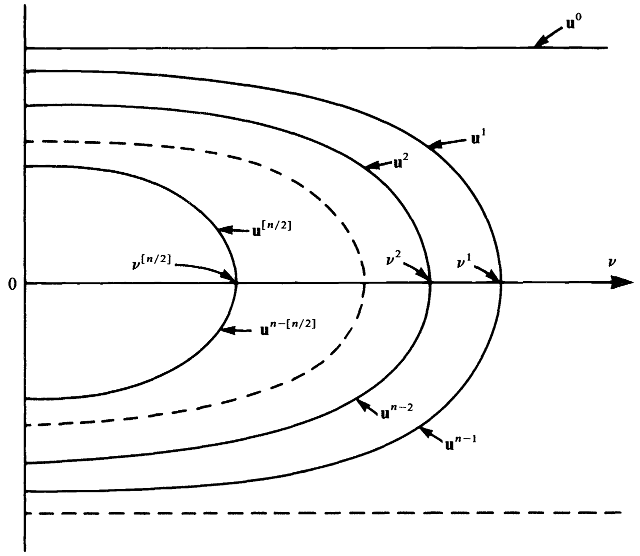


FIG. 3. *A bifurcation diagram for fixed $\mu > 0$, say $\tilde{\mu}$. The value, ν^l, is determined by $\mu^l(\nu^l) = \tilde{\mu}$. This picture is illustrated in the similar manner as in Fig. 2.*

where, $y_{-1} = y_{n-1}$, $y_n = y_0$ and $\mu^l = \mu^l(\varepsilon, \nu)$. Let

$$\alpha_1 \equiv i\omega^l(\varepsilon, \nu) + \varepsilon^2 B_1 + \nu\sigma^l C_1, \qquad \alpha_2(t) \equiv \varepsilon^2 B_1\, e^{2i\omega^l(\varepsilon,\,\nu)t}. \tag{5.2}$$

Then, considering the expression of (4.7), we can rewrite (5.1) in the following vector form:

$$\dot{\mathbf{y}} = \alpha_1\mathbf{y} + n\alpha_2(t)(\mathbf{P}_l)^2\bar{\mathbf{y}} + \nu C_1\Delta_d\mathbf{y} + \mathbf{w}^1(\varepsilon, \nu, t)\mathbf{y} + \mathbf{W}^2(\varepsilon, \nu, t)\bar{\mathbf{y}} \tag{5.3}$$

where $\mathbf{P}_l$ is defined by (3.3) and $\mathbf{W}^j, j = 1, 2$, are matrices whose elements are periodic with period, $T^l(\varepsilon, \nu) = 2\pi/\omega^l(\varepsilon, \nu)$, and are $O(\varepsilon|(\varepsilon, \nu)|^2 + |\nu(\varepsilon, \nu)|)$. Using (3.6), (3.7) and

$$\mathbf{y} = \mathbf{P}_l\mathbf{P}\mathbf{v}, \tag{5.4}$$

we transform (5.3) into

$$\begin{aligned} &\dot{\mathbf{v}} = \alpha_1\mathbf{v} + \alpha_2(t)S_2\bar{\mathbf{v}} - \nu C_1\mathbf{\Lambda}_l\mathbf{v} + \tilde{\mathbf{W}}^1(\varepsilon, \nu, t)\mathbf{v} + \tilde{\mathbf{W}}^2(\varepsilon, \nu, t)\bar{\mathbf{v}}, \\ &\tilde{\mathbf{W}}^j \equiv (\mathbf{P}_l\mathbf{P})^{-1}\mathbf{W}^j(\mathbf{P}_l\mathbf{P}), \qquad j = 1, 2 \end{aligned} \tag{5.5}$$

where S_2 and Λ_l are as in (2.22) and (3.6), respectively.

In general, let $\Phi(t)$ be a fundamental matrix solution (with $\Phi(0) = I$, the identity matrix) of a linear periodic system such as (5.5). Then Floquet theory (for example, see [6]) implies that $\Phi(t)$ can be expressed as $\Phi(t) = X(t)\, e^{\mathbf{M}t}$, where $X(t)$ is periodic with the same period as the original system, and $\mathbf{M}$ is a certain matrix. By virtue of this expression, the real part of an eigenvalue of $\mathbf{M}$ gives a criteria for stability of the zero solution. The eigenvalue of $\mathbf{M}$ is called a Floquet exponent.

In the case of (5.5), after the change of variables, $v_k = w_k\, e^{i\omega^l(\varepsilon,\,\nu)t}$, $k = 0, \cdots, n-1$, (5.5) is written in the form,

$$\begin{pmatrix}\mathbf{w}\\ \dot{\bar{\mathbf{w}}}\end{pmatrix} = (\mathbf{M}_0 + \mathbf{M}_1(\varepsilon, \nu, t))\begin{pmatrix}\mathbf{w}\\ \bar{\mathbf{w}}\end{pmatrix} \tag{5.6}$$

where $\mathbf{M}_0$ is a constant matrix with $O(\varepsilon^2 + \nu)$ and $\mathbf{M}_1 = O(\nu|(\varepsilon, \nu)| + \varepsilon|(\varepsilon, \nu)|^2)$. This fundamental matrix solution, $\Phi(t)$, satisfies

$$\Phi(t)=e^{\mathbf{M}_0 t}\left(I+\int_0^t e^{-\mathbf{M}_0 s}\mathbf{M}_1(\varepsilon,\nu,s)\Phi(s)\,ds\right),$$

from which we see that

$$e^{\mathbf{M}\cdot T^l(\varepsilon,\nu)}=e^{\mathbf{M}_0\cdot T^l(\varepsilon,\nu)}\{I+O(\nu|(\varepsilon,\nu)|+\varepsilon|(\varepsilon,\nu)|^2)\}.$$

This implies that an eigenvalue of $\mathbf{M}_0$ gives the principal part of the one corresponding to $\mathbf{M}$, if ε and ν are sufficiently small. Since in the absence of $\mathbf{M}_1$ the components of the equation (5.6) are given by:

$$\begin{aligned}\dot{w}_k&=\{\varepsilon^2B_1+\nu C_1(\sigma_l-\sigma_{l+k})\}w_k+\varepsilon^2B_1\bar{w}_{n-k},\\ \dot{\bar{w}}_{n-k}&=\{\varepsilon^2\bar{B}_1+\nu\bar{C}_1(\sigma_l-\sigma_{l-k})\}\bar{w}_{n-k}+\varepsilon^2\bar{B}_1w_k,\end{aligned}\qquad k=0,\cdots,n-1 \tag{5.7}$$

(note $\sigma_{l+n-k}=\sigma_{l-k}$), the next lemma immediately follows.

LEMMA 5.1. *The linearized equation of* (2.19) *around the periodic solution,* $\mathbf{z}^l(t)$, *given by* (4.7) *has the Floquet exponent such that*

$$\gamma=\gamma_2(\varepsilon,\nu)+\hat{\gamma}(\varepsilon,\nu)$$

where $\hat{\gamma}(\varepsilon,\nu)=O(\varepsilon|(\varepsilon,\nu)|^2+\nu|(\varepsilon,\nu)|)$ *and* $\gamma_2(\varepsilon,\nu)$ *is a root of the equation,*

$$\gamma_2^2-\xi_k^l\gamma_2+\eta_k^l=0,\qquad k=0,\cdots,n-1, \tag{5.8}$$

whose coefficients are defined by

$$\begin{aligned}\xi_k^l&\equiv 2\varepsilon^2b_1-\nu D_1(\sigma_{l-k}-2\sigma_l+\sigma_{l+k})+i\nu D_2(\sigma_{l-k}-\sigma_{l+k}),\\ \eta_k^l&\equiv\nu^2(D_1^2+D_2^2)(\sigma_l-\sigma_{l+k})(\sigma_l-\sigma_{l-k})\\ &\qquad-\nu\varepsilon^2(D_1b_1+D_2b_2)(\sigma_{l-k}-2\sigma_l+\sigma_{l+k})\\ &\qquad-i\nu\varepsilon^2(D_1b_2-D_2b_1)(\sigma_{l-k}-\sigma_{l+k})\end{aligned} \tag{5.9}$$

($b_j, D_j, j=1,2$ *are defined in* (4.3)). *Furthermore,*

$$\xi_{n-k}^l=\bar{\xi}_k^l,\quad \eta_{n-k}^l=\bar{\eta}_k^l,\quad \xi_k^{n-l}=\bar{\xi}_k^l\quad\text{and}\quad \eta_k^{n-l}=\bar{\eta}_k^l, \tag{5.10}$$

hold.

Now consider the equation (5.8). For $k=0$, (5.8) is $\gamma_2^2-2\varepsilon^2b_1\gamma_2=0$, which has roots $0, 2\varepsilon^2b_1(<0$ by (A2)). A zero Floquet exponent always exists for the linearized equation around a periodic solution. The above zero corresponds to this.

Next consider the case $k\geqq 1$. In the specific cases $l=0$ and $n/2$ (for n: even), the coefficients of (5.7) are real valued such as:

$$\begin{aligned}\xi_k^0&=2(\varepsilon^2b_1-\nu D_1\sigma_k),\\ \eta_k^0&=\nu\sigma_k(D_1^2+D_2^2)\left\{\nu\sigma_k-\frac{2\varepsilon^2E_1}{D_1^2+D_2^2}\right\},\\ \xi_k^{n/2}&=2\{\varepsilon^2b_1+\nu D_1(\sigma_{n/2}-\sigma_{n/2-k})\},\\ \eta_k^{n/2}&=\nu(\sigma_{n/2}-\sigma_{n/2-k})(D_1^2+D_2^2)\left\{\nu(\sigma_{n/2}-\sigma_{n/2-k})+\frac{2\varepsilon^2E_1}{D_1^2+D_2^2}\right\},\end{aligned}$$

where, $E_1=D_1b_1+D_2b_2$. Note that $\sigma_{n/2}-\sigma_{n/2-k}=\sigma_{n/2}-\sigma_{n/2+k}>0$, $k\geqq 1$. We let $D_1>0$, $E_1>0$. Then both roots for $l=0$ exist in the left half of the complex plane for ε near the zero, and one of them crosses the origin from left to right as ε increases, while for $l=n/2$, the pair of roots in the right for small ε cross the imaginary axis from right to left as ε increases. If $D_1>0$ and $E_1<0$, then for $l=0$, both roots remain in the left half plane, while for $l=n/2$, there are two cases. For $D_1E_1+D_2E_2<0$ ($E_2=D_1b_2-D_2b_1$), one of them always has positive real part. In the case of $D_1E_1+D_2E_2>0$, the roots in the right half for ε small cross the imaginary axis from right to left, moreover thereafter one of them crosses the origin from left to right.

We can also check the change of roots in a similar manner for the case $D_1<0$. For general l, the movement of the root of (5.8) in the complex plane is more complicated when ε increases (or ν decreases). In a specific region of the parameter space (μ, ν), however, we can see from Appendix B the stability of all the periodic solutions in (4.7) as follows.

THEOREM 5.2. *Consider the periodic solutions of* (1.1) *which bifurcate from the origin, under the assumptions,* (A1), (A2) *and*

$$\text{(A3)} \qquad D_1=\operatorname{Re}(D\zeta_0, \zeta_0^*)>0.$$

If $\operatorname{Re}((D\zeta_0, \zeta_0^*)\bar{B}_1)=D_1b_1+D_2b_2$ *is positive* (*i.e.,* $D_2b_2>-D_1b_1>0$) *and fixed, there exists a sufficiently small* $\theta_1>0$ *such that for* $(\mu, \nu), 0<\nu<\theta_1\mu$, *the periodic solution in* (1.3) *with indices, l and $n-l$, $0\leqq l\leqq n/4$* (*resp. $n/4<l\leqq n/2$*) *are unstable* (*resp. stable*); *if $D_1b_1+D_2b_2<0$, then in a region satisfying, $0<\nu<\theta_2\mu$ and $0<\theta_2\ll 1$, those solutions with l and $n-l$, $0\leqq l<n/4$* (*resp. $n/4\leqq l\leqq n/2$*) *are stable* (*resp. unstable*). *The numbers, $\theta_j, j=1, 2$, depend on l, $0\leqq l\leqq n/2$ and the values of $D_1b_1+D_2b_2$.*

Remark 5.1. The assumptions (A1) and (A3) imply that the steady state is stable (resp. unstable) for $\mu<0$ (resp. $\mu>0$). The above discussion for the Floquet exponents of (5.1) (or (5.5)) with $l=0$ yields that the homogeneous periodic solution is always stable for $E_1<0$ and that for $E_1>0$ it becomes unstable in a region, if (A1)-(A3) hold. On the other hand, under (A1)-(A3) the bifurcation of all the periodic wave solutions occurs in the unstable region of the steady state (see Fig. 1), which implies that these solutions are unstable fairly near the bifurcation points. Theorem 5.2 says that recovery of stability occurs for several of the periodic solutions and that they have the stable region as described in the statement.

Remark 5.2. For the case $D_1=\operatorname{Re}(D\zeta_0, \zeta_0^*)<0$, we can also discuss the stability. To avoid repeating a similar statement, we omit the case in the above theorem. The condition $D_1<0$ is supposed to be a more specific case than $D_1>0$ (see the examples in §§ 6 and 7), which is another reason to omit it here. We also note that $D_1>0$ implies $\operatorname{Re}(\partial\lambda_l/\partial\nu)(0, 0)<0, l=1, \cdots, n-1$, by ($\alpha$3) in Appendix A.

Remark 5.3. When the eigenvalue of (5.1) crosses the imaginary axis, a secondary bifurcation occurs and an invariant torus or a new periodic solution might appear. For the specific case $n=2$ (i.e., two coupled oscillators) the structure of a secondary bifurcation of a periodic solution to a class of equation (1.1) has already been studied in [14].

To recover the stability, some kinds of secondary bifurcation occur along the branch of the periodic solution. Therefore the complicated structure of (1.1) for a large system (having many oscillators) is imagined. In spite of this, Theorem 5.2 states that the stability of the primary bifurcating solution is clearly classified in a parameter region.

6. An example. The following equation which is called the Brusselator is known as a simple biochemical reaction model:

$$(6.1) \qquad \dot{x}(t)=A-(B+1)x+x^2y, \qquad \dot{y}(t)=Bx-x^2y$$

where A and B are positive parameters. The equation (6.1) has a steady state, $(x, y)=(A, B/A)$, and at $B=A^2+1$ a Hopf bifurcation occurs; a family of stable limit cycles bifurcate from the steady state (see [19]). Hence $F(\mu, \cdot)$ in (1.1) (or (1.2)) is given by

$$F(\mu, u, v)=\begin{pmatrix} A^2+\mu & A^2 \\ -(A^2+1+\mu) & -A^2 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}+\left(\frac{A^2+1+\mu}{A}u^2+2Auv+u^2v\right)\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

where $B=A^2+1+\mu$, $x=A+u$ and $y=B/A+v$.

After a calculation (referring to [11, Appendix]), we have

$$\zeta_0=\begin{pmatrix} 1 \\ \dfrac{i-A}{A} \end{pmatrix}, \qquad \zeta_0^*=\frac{1+Ai}{2}\begin{pmatrix} 1 \\ \dfrac{A}{A-i} \end{pmatrix},$$

$$B_1=-\frac{1}{2}\left(\frac{A^2+2}{A^2}+i\frac{4A^4-7A^2+4}{3A^2}\right),$$

and if

$$D=\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}, \quad d_1, d_2 \geqq 0, \quad d_1+d_2>0,$$

then

$$(D\zeta_0, \zeta_0^*)=\frac{1}{2}\{(d_1+d_2)+iA(d_2-d_1)\},$$

$$\operatorname{Re}(\bar{B}_1(D\zeta_0, \zeta_0^*))=-\frac{1}{12A^2}\{3(d_1+d_2)(A^2+2)+(d_2-d_1)(4A^4-7A^2+4)\}.$$

Since $\operatorname{Re}(D\zeta_0, \zeta_0^*)>0$, we can apply Theorem 5.2 to this case. According to the choice of d_1, d_2 and A, $\operatorname{Re}(\bar{B}_1(D\zeta_0, \zeta_0^*))$ takes negative or positive value.

7. Application to a retarded functional differential equation. One may encounter a model of an oscillator represented by a differential equation with time delay. As a specific example, we shall take the following equation:

$$\dot{u}(t)=-\left(\frac{\pi}{2}+\mu\right)(1+u(t))u(t-1). \tag{7.1}$$

This equation has a periodic solution which bifurcates from the origin at $\mu=0$ (for example, see [8]). Using this equation, we get to a chain of weakly coupled equations such as

$$\begin{gathered} \dot{u}_k(t)=-\left(\frac{\pi}{2}+\mu\right)(1+u_k(t))u_k(t-1)+\nu(u_{k-1}(t)-2u_k(t)+u_{k+1}(t)), \\ k=0, \cdots, n-1, \\ u_{-1}=u_{n-1}, \qquad u_n=u_0. \end{gathered} \tag{7.2}$$

The equations (7.1) and (7.2) are included in a class of retarded functional differential equations. In this section, first we shall state the formulation for a functional differential equation of general form (by following [6]), and then we shall show that the results obtained in the preceding sections can be extended to this case. Finally we shall give the values corresponding to A_1, B_1 and C_1 in (2.19) for the above equation (7.2). (We refer to [8] about the formulation below.)

Let $C=C([-r, 0]; \mathbf{R}^m)$ be a set of all continuous functions defined on $[-r, 0]$ with the values in $\mathbf{R}^m$; C is a Banach space equipped with supremum norm. We define $u_t \in C$, $t \in [a, b]$ by $u_t(\theta)=u(t+\theta)$, $-r \leqq \theta \leqq 0$, for a continuous function $u(t) \in$

$C([a-r, b]; \mathbf{R}^m)$. Let $F(\cdot, \cdot): I_0 \times C \to \mathbf{R}^m$ be a sufficiently smooth operator satisfying $F(\mu, 0) = 0$, $\mu \in I_0$. Using this operator, we consider the equation,

$$\dot{u}(t) = F(\mu, u_t) = L(\mu)u_t + M(\mu, u_t) \tag{7.3}$$

where $L(\mu)$ and $M(\mu, \cdot)$ are the linear and nonlinear parts of $F(\mu, \cdot)$, respectively.

Riesz's representation theorem implies that there is an $m \times m$ matrix function $\eta(\theta; \mu)$ whose elements have bounded variation in θ such that

$$L(\mu)\phi = \int_{-r}^{0} [d\eta(\theta; \mu)]\phi(\theta), \qquad \phi \in C.$$

The generator $A(\mu)$ of a semigroup $\{T(t)\}_{t \geqq 0}$, associated with the equation

$$\dot{u}(t) = L(\mu)u_t$$

is defined by

$$A(\mu)\phi = \begin{cases} \dfrac{d\phi}{d\theta}, & -r \leqq \theta < 0 \\ L(\mu)\phi, & \theta = 0 \end{cases} \quad \text{for } \phi \in \mathscr{D}(A(\mu)). \tag{7.4}$$

Then the eigenvalue is given by a root of the characteristic equation

$$\det\left(\lambda I - \int_{-r}^{0} e^{\lambda\theta}[d\eta(\theta; \mu)]\right) = 0. \tag{7.5}$$

We assume that $A(\mu)$ has a pair of simple eigenvalues, $\lambda_0(\mu)$ and $\overline{\lambda_0(\mu)}$, satisfying $\lambda_0(0) = i\omega_0$, $\operatorname{Re}(d\lambda_0/d\mu)(0) > 0$. The function, $\zeta_1(\theta)$, $-r \leqq \theta \leqq 0$, denotes an eigenfunction corresponding to $i\omega_0$. Define $\tilde{G}(\mu, \cdot)$ by:

$$\tilde{G}(\mu, \phi)(\theta) \equiv \begin{cases} 0, & -r \leqq \theta < 0, \\ G(\mu, \phi) \equiv F(\mu, \phi) - L(0)\phi, & \theta = 0. \end{cases}$$

By this and (7.4) the equation (7.3) can be written as

$$\begin{aligned} \frac{d}{dt} u_t &= A(0)u_t + \tilde{G}(\mu, u_t) \\ &= \begin{cases} \dfrac{du(t+\theta)}{d\theta}, & -r \leqq \theta < 0, \\ L(0)u_t + G(\mu, u_t), & \theta = 0. \end{cases} \end{aligned} \tag{7.6}$$

We shall introduce a bilinear form:

$$\begin{aligned} \langle \phi, \psi \rangle = (\phi(0), \psi(0)) - \int_{-r}^{0} \int_{0}^{\theta} {}^t\overline{\psi(\xi - \theta)}[d\eta(\theta; 0)]\phi(\xi)\, d\xi \\ \text{for } \phi \in C([-r, 0]; \mathbf{C}^m) \text{ and } \psi \in C^* = C([0, r]; \mathbf{C}^m). \end{aligned} \tag{7.7}$$

The adjoint operator A^* of $A(0)$ is defined by

$$[A^*\psi](s) \equiv \begin{cases} -\dfrac{d\psi}{ds}(s), & 0 < s \leqq r, \\ \displaystyle\int_{-r}^{0} {}^t[d\eta(\theta; 0)]\psi(-\theta), & s = 0, \end{cases} \qquad \psi \in C^*,$$

that is, A^* satisfies $\langle A(0)\phi, \psi \rangle = \langle \phi, A^*\psi \rangle$. We let $\zeta_1^*(s)$, $0 \leqq s \leqq r$, be an eigenfunction of A^* corresponding to $-i\omega_0$, and let $\langle \zeta_1, \zeta_1^* \rangle = 1$ (note $\langle \bar{\zeta}_1, \zeta_1^* \rangle = 0$). We can decompose (7.6) as

$$
\begin{aligned}
\dot z &= i\omega_0 z + \langle \tilde G(\mu, z\zeta_1 + \overline{z\zeta_1} + u_t^Q), \zeta_1^* \rangle \\
&= i\omega_0 z + (G(\mu, z\zeta_1 + \overline{z\zeta_1} + u_t^Q), \zeta_0^*) \qquad (\zeta_0^* = \zeta_1^*(0)), \\
\frac{d}{dt} u_t^Q &= A(0)u_t^Q + Q\tilde G(\mu, z\zeta_1 + \overline{z\zeta_1} + u_t^Q)
\end{aligned} \tag{7.8}
$$

where $z(t) \equiv \langle u_t, \zeta_1^* \rangle$, $u^Q \equiv u - u^P = u - (z\zeta_1 + \overline{z\zeta_1})$. The existence of a center manifold of (7.8) was shown in [2], by which we are able to reduce (7.8) to an equation on the manifold. This calculation is similar to the case of an ordinary differential equation (see [8]).

We shall consider the following coupled equations of (7.3):

$$
\begin{aligned}
&\dot u^k = F(\mu, u_t^k) + \nu\{N(u_t^{k-1} - u_t^k) + N(u_t^{k+1} - u_t^k)\}, \\
&u^{-1} = u^{n-1}, \qquad u^n = u^0
\end{aligned} \tag{7.9}
$$

where $N(\cdot)$ is a smooth operator of C into $\mathbf{R}^m$ with $N(0) = 0$ and $D = (dN/du)(0)$. By (7.8) the equation (7.9) can be decomposed as seen in (2.3), and the center manifold theorem yields an equation on the manifold. After computations quite similar to those described in § 2 and Appendix A, we have a slightly modified formula of the reduced equation; that is, in this case, A_1, B_1 and C_1 in (2.19) are given by:

$$
\begin{aligned}
A_1 &= \left(\frac{d}{d\mu} L(0)\zeta_1, \zeta_0^*\right) = \frac{d\lambda_0}{d\mu}(0), \\
B_1 &= (F_{uu}(0)(\zeta_1, \hat\zeta_2), \zeta_0^*) + (F_{uu}(0)(\bar\zeta_1, \zeta_2), \zeta_0^*) + \tfrac{1}{2}(F_{uuu}(0)(\zeta_1, \zeta_1, \overline{\zeta_1}), \zeta_0^*), \\
C_1 &= (D\zeta_1, \zeta_0^*)
\end{aligned} \tag{7.10}
$$

where

$$
\begin{aligned}
&\zeta_2 \equiv (2i\omega_0 - A(0))^{-1}\tfrac{1}{2}\tilde F_{uu}(0)(\zeta_1, \zeta_1), \\
&\hat\zeta_2 \equiv -A(0)^{-1}\tilde F_{uu}(0)(\zeta_1, \bar\zeta_1), \\
&[\tilde F_{uu}(0)(\cdot,\cdot)](\theta) \equiv \begin{cases} 0, & -r \leqq \theta < 0, \\ F_{uu}(0)(\cdot,\cdot), & \theta = 0. \end{cases}
\end{aligned}
$$

Now consider (7.2). Then we have

$$
F(\mu, \phi) = -\left(\frac{\pi}{2} + \mu\right)(1 + \phi(-1))\phi(0), \quad D\phi = \phi(0), \quad \phi \in C[-1, 0].
$$

In this case, A_1, B_1 and C_1 in (7.10) are easily calculated as follows:

$$
A_1 = \left(\frac{\pi}{2} + i\right) \Big/ \left(1 + \frac{\pi^2}{4}\right), \qquad B_1 = \frac{\pi}{10}(1 - 3i)\left(1 - \frac{\pi}{2}i\right) \Big/ \left(1 + \frac{\pi^2}{4}\right),
$$

$$
C_1 = \left(1 - \frac{\pi}{2}i\right) \Big/ \left(1 + \frac{\pi^2}{4}\right)
$$

(refer to [12]). Hence we obtain $\operatorname{Re}(\bar B_1 C_1) = \pi/10/(1 + \pi^2/4) > 0$.

Appendix A. Calculation of coefficients for equations (2.8)–(2.10). First we note that

$$
\frac{d\lambda_0}{d\mu}(0) = (A(0)\zeta_0, \zeta_0^*), \tag{α1}
$$

which is derived from $A(\mu)\zeta_0(\mu) = \lambda_0(\mu)\zeta_0(\mu)$. The linearized equation of (1.1) around

$u_k = 0$, $k = 0, \cdots, n-1$, is given by

$$(\alpha 2) \qquad \begin{aligned} &\dot{u}_k = A(\mu)u_k + \nu D(u_{k-1} - 2u_k + u_{k+1}), \qquad k = 0, \cdots, n-1,\\ &u_{-1} = u_{n-1}, \qquad u_n = u_0. \end{aligned}$$

It is easily checked that (α2) has the eigenvalues $\lambda_l(\mu, \nu)$, $l = 1, \cdots, n-1$ satisfying $\lambda_l(\mu, 0) = \lambda_0(\mu)$ in a neighborhood of $(\mu, \nu) = (0, 0)$. Furthermore a corresponding eigenvector to $\lambda_l(\mu, \nu)$ is given by

$$\xi_k = e^{ik \cdot 2l\pi/n} \zeta_l(\mu, \nu), \qquad k = 0, \cdots, n-1$$

where $\zeta_l(\mu, \nu)$ satisfies

$$(A(\mu) - \nu\sigma_l D)\zeta_l(\mu, \nu) = \lambda_l(\mu, \nu)\zeta_l(\mu, \nu)$$

(σ_l is as in (3.5)). Hence, in addition to (α1), we obtain the relation

$$(\alpha 3) \qquad \frac{\partial \lambda_l}{\partial \nu}(0, 0) = -\sigma_l(D\zeta_0, \zeta_0^*),$$

by differentiating it with respect to ν.

The coefficients of (2.9) which are necessary for the later calculation are as follows:

$$(\alpha 4) \qquad \begin{aligned} &a_{10} = (F_{\mu u}(0)\zeta_0, \zeta_0^*), \qquad a_{20} = \tfrac{1}{2}(F_{uu}(0)(\zeta_0, \zeta_0), \zeta_0^*),\\ &a_{11} = (F_{uu}(0)(\zeta_0, \bar{\zeta}_0), \zeta_0^*), \qquad a_{02} = (F_{uu}(0)(\bar{\zeta}_0, \bar{\zeta}_0), \zeta_0^*),\\ &a_{21} = (F_{uu}(0)(\zeta_0, h_{z\bar{z}}(0)), \zeta_0^*) + \tfrac{1}{2}(F_{uu}(0)(\bar{\zeta}_0, h_{zz}(0)), \zeta_0^*)\\ &\qquad + \tfrac{1}{2}(F_{uuu}(0)(\zeta_0, \zeta_0, \bar{\zeta}_0), \zeta_0^*) \end{aligned}$$

where $F_{\mu u}(0) \equiv (\partial^2/\partial\mu\, \partial u)F(0, 0)$, $F_{uu}(0) \equiv (\partial^2/\partial u^2)F(0, 0)$, $h_{zz}(0) \equiv (\partial^2/\partial z^2)h(0, 0, 0)$ and so forth. We can also check that

$$(\alpha 5) \qquad \begin{aligned} &h_{zz}(0) = (2i\omega_0 - A(0))^{-1}QF_{uu}(0)(\zeta_0, \zeta_0),\\ &h_{z\bar{z}}(0) = -A(0)^{-1}QF_{uu}(0)(\zeta_0, \bar{\zeta}_0),\\ &h_{\bar{z}\bar{z}}(0) = \overline{h_{zz}(0)}. \end{aligned}$$

We shall consider (2.14). By the relations (α4),

$$(\alpha 6) \qquad \begin{aligned} B_1 &= a_{21} + \frac{i}{\omega_0}(2a_{20}a_{11} - a_{20}a_{11}) - \frac{i}{\omega_0}\overline{a_{11}}a_{11} - \frac{2i}{3\omega_0}\overline{a_{02}}a_{02}\\ &= a_{21} + \frac{i}{\omega_0}a_{11}(F_{uu}(0)(\zeta_0, \zeta_0), \zeta_0^*) - \frac{i}{\omega_0}\overline{a_{11}}(F_{uu}(0)(\zeta_0, \bar{\zeta}_0), \zeta_0^*)\\ &\quad - \frac{i}{\omega_0}a_{20}(F_{uu}(0)(\zeta_0, \overline{\zeta_0}), \zeta_0^*) - \frac{i}{3\omega_0}\overline{a_{02}}(F_{uu}(0)(\overline{\zeta_0}, \overline{\zeta_0}), \zeta_0^*)\\ &= \left(F_{uu}(0)\left(\zeta_0, h_{z\bar{z}}(0) + \frac{i}{\omega_0}a_{11}\zeta_0 - \frac{i}{\omega_0}\overline{a_{11}\zeta_0}\right), \zeta_0^*\right)\\ &\quad + \left(F_{uu}(0)\left(\bar{\zeta}_0, h_{zz}(0) - \frac{i}{\omega_0}a_{20}\zeta_0 - \frac{i}{3\omega_0}\overline{a_{02}\zeta_0}\right), \zeta_0^*\right)\\ &\quad + \tfrac{1}{2}(F_{uuu}(0)(\zeta_0, \zeta_0, \overline{\zeta_0}), \zeta_0^*). \end{aligned}$$

On the other hand, by (α5), ζ_2 defined in (2.16) is written as:

$$(\alpha 7) \qquad \begin{aligned} \zeta_2 &= \tfrac{1}{2}h_{zz}(0) + \tfrac{1}{2}(2i\omega_0 - A(0))^{-1}PF_{uu}(0)(\zeta_0, \zeta_0)\\ &= \tfrac{1}{2}h_{zz}(0) + (2i\omega_0 - A(0))^{-1}(a_{20}\zeta_0 + \overline{a_{02}\zeta_0}). \end{aligned}$$

Similarly,

$$(\alpha 8)\qquad \begin{aligned}\hat{\zeta}_2 &= h_{z\bar{z}}(0) - A(0)^{-1}PF_{uu}(0)(\zeta_0, \overline{\zeta_0})\\ &= h_{z\bar{z}}(0) - A(0)^{-1}(a_{11}\zeta_0 + \overline{a_{11}\zeta_0}).\end{aligned}$$

We easily check that

$$(\alpha 9)\qquad \begin{aligned}&(2i\omega_0 - A(0))^{-1}\zeta_0 = -\frac{i}{\omega_0}\zeta_0, \qquad (2i\omega_0 - A(0))^{-1}\overline{\zeta_0} = -\frac{i}{3\omega_0}\overline{\zeta_0},\\ &-A(0)^{-1}\zeta_0 = \frac{i}{\omega_0}\zeta_0, \qquad -A(0)^{-1}\overline{\zeta_0} = -\frac{i}{\omega_0}\overline{\zeta_0}.\end{aligned}$$

Combining (α6), (α7), (α8) and (α9), we obtain (2.15).

Appendix B. Proof of Theorem 5.2: Stability of periodic solutions. Let $\gamma_\alpha = \gamma_1 + i\gamma_2$ and $\gamma_\beta = \gamma_3 + i\gamma_4$ be the roots of (5.8). Then,

$$(\beta 1)\qquad \begin{aligned}&\gamma_1 + \gamma_3 = 2\varepsilon^2 b_1 - \nu D_1 \Delta_k^l \sigma,\\ &\gamma_2 + \gamma_4 = \nu D_2(\sigma_{l-k} - \sigma_{l+k}),\\ &\gamma_1\gamma_3 - \gamma_2\gamma_4 = \nu^2(D_1^2 + D_2^2) S_k^l\sigma - \nu\varepsilon^2 E_1 \Delta_k^l\sigma,\\ &\gamma_1\gamma_4 + \gamma_2\gamma_3 = -\nu\varepsilon^2 E_2(\sigma_{l-k} - \sigma_{l+k})\end{aligned}$$

where $E_1 \equiv D_1 b_1 + D_2 b_2$, $E_2 \equiv D_1 b_2 - D_2 b_1$ and

$$\Delta_k^l\sigma \equiv \sigma_{l-k} - 2\sigma_l + \sigma_{l+k} = 8\sin^2\frac{k\pi}{n}\cos\frac{2l\pi}{n},$$

$$S_k^l\sigma \equiv (\sigma_l - \sigma_{l+k})(\sigma_l - \sigma_{l-k}) = 16\sin^2\frac{k\pi}{n}\sin\frac{k+2l}{n}\pi\sin\frac{k-2l}{n}\pi.$$

We are only interested in the signs of γ_1 and γ_3, so by (5.10) it is enough to consider the cases $1 \leqq l < n/2$, $1 \leqq k \leqq n/2$.

First we note that

$$(\beta 2)\qquad \Delta_k^l\sigma = \begin{cases} >0, & 1 \leqq l < \frac{n}{4},\\ =0, & l = \frac{n}{4},\\ <0, & \frac{n}{4} < l < \frac{n}{2},\end{cases}$$

$$(\beta 3)\qquad S_k^l\sigma = \begin{cases} <0, & 1 \leqq k < 2l,\\ =0, & k = 2l, \qquad \text{for } l,\ 1 \leqq l < \frac{n}{4},\\ >0, & 2l < k \leqq \frac{n}{2},\end{cases}$$

$$S_k^l\sigma = \begin{cases} <0, & 1 \leqq k < n-2l,\\ =0, & k = n-2l, \qquad \text{for } l,\ \frac{n}{4} < l < \frac{n}{2},\\ >0, & n-2l < k \leqq \frac{n}{2},\end{cases}$$

and

$$(\beta 4)\qquad S_k^{n/4}\sigma = -(\sigma_{n/4}-\sigma_{n/4\pm k})^2 = \begin{cases} =0, & k=\dfrac{n}{2}, \\ <0, & k\neq\dfrac{n}{2}. \end{cases}$$

(1) Assume that $\Delta_k^l\sigma \neq 0$: Let ε vanish in $(\beta 1)$. A little consideration reveals that $\gamma_1\gamma_3>0$ (resp. <0) holds for $\gamma_1\gamma_3-\gamma_2\gamma_4>0$ (resp. <0), and that $\gamma_1\gamma_3=0$ for $\gamma_1\gamma_3-\gamma_2\gamma_4=0$ (i.e., $S_k^l\sigma=0$). Next let $\nu/\varepsilon^2 \ll 1$ in $(\beta 1)$. Putting $\nu=0$ in $(\beta 1)$, we have $\gamma_2=\gamma_4=0$. Therefore, $\gamma_2=O(\nu)=\gamma_4$, which implies that

$$(\beta 5)\qquad \begin{aligned} &\operatorname{sign}(\gamma_1+\gamma_3)=\operatorname{sign}(2b_1\varepsilon^2), \\ &\operatorname{sign}(\gamma_1\gamma_3-\gamma_2\gamma_4)=\operatorname{sign}\{-\nu\varepsilon^2E_1\Delta_k^l\sigma\}=\operatorname{sign}(\gamma_1\gamma_3). \end{aligned}$$

Hence $(\beta 2)$ and $(\beta 5)$ show the stability result in Theorem 5.2 for $l\neq n/4$.

(2) Assume that $\Delta_k^l\sigma=0$ (i.e., $l=n/4$): We see from $(\beta 4)$ that $\gamma_2=\gamma_4=0$, $\gamma_1\gamma_3=0$ and $\gamma_1+\gamma_3<0$ for $k=n/2$. Next let $k\neq n/2$. If $\gamma_1=0$, then

$$\begin{aligned} &\gamma_2+\gamma_4=\nu D_2(\sigma_{l-k}-\sigma_{l+k})=2\nu D_2(\sigma_{l-k}-\sigma_l), \\ &\gamma_2\gamma_4=-\nu^2(D_1^2+D_2^2)S_k^l\sigma, \end{aligned}$$

which yields a contradiction: $(\gamma_2-\gamma_4)^2=(\gamma_2+\gamma_4)^2-4\gamma_2\gamma_4<0$. Thus the roots γ_α and γ_β never cross the imaginary axis in this case. This concludes the proof of Theorem 5.2.

Acknowledgment. The author expresses thanks to Professor Masaya Yamaguti for his suggestion which proved to be fruitful.

REFERENCES

[1] J. Carr, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[2] N. Chafee, *A bifurcation problem for a functional differential equation of finitely retarded type*, J. Math. Anal. Appl., 35 (1971), pp. 312-348.

[3] G. B. Ermentrout and N. Kopell, *Frequency plateaus in a chain of weakly coupled oscillators* I, this Journal, 15 (1984), pp. 215-237.

[4] S. A. Gils and T. Valkering, *Hopf bifurcation and symmetry: standing and travelling waves in a circular chain*, Japan J. Appl. Math., to appear.

[5] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[6] J. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[7] ———, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[8] B. P. Hassard, N. D. Kazarinoff and Y. H. Wan, *Theory and Applications of Hopf Bifurcation*, London Math. Soc. Monographs, Lecture Note 41, 1981.

[9] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, 1980.

[10] M. Kawato and R. Suzuki, *Two coupled newral oscillators as a model of the circadian pacemaker*, J. Theoret. Biol., 86 (1980), pp. 547-575.

[11] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York, 1984.

[12] Y. Morita, *Destabilization of periodic solutions arising in delay-diffusion systems in several space dimensions*, Japan J. Appl. Math., 1 (1984), pp. 39-65.

[13] ———, *Stability changes of periodic solutions to a coupled nonlinear equation with time delay*, Publ. RIMS, Kyoto Univ., 21 (1985), pp. 47-74.

[14] ———, *A secondary bifurcation problem of weakly coupled oscillators with time delay*, Japan J. Appl. Math., to appear.

[15] J. Neu, *Coupled chemical oscillators*, SIAM J. Appl. Math., 37 (1979), pp. 307-315.

[16] ———, *Chemical waves and diffusive coupling of limit cycle oscillators*, SIAM J. Appl. Math., 36 (1979), pp. 509-515.

[17] ———, *Large populations of coupled chemical oscillators*, SIAM J. Appl. Math., 38 (1980), pp. 305-316.

[18] R. H. Rand and P. H. Holmes, *Bifurcation of periodic motions in two weakly coupled Van der Pol oscillators*, Internat. J. Non-Linear Mech., 15 (1980), pp. 387-399.

[19] J. J. Tyson, *Some further studies of nonlinear oscillations in chemical systems*, J. Chem. Phys., 58 (1973), pp. 3919-3930.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

015

MELNIKOV'S METHOD AT A SADDLE-NODE AND THE DYNAMICS OF THE FORCED JOSEPHSON JUNCTION*

STEPHEN SCHECTER†

Abstract. A version of Melnikov's method is developed for time-periodic perturbations of a planar vector field having a separatrix loop at a saddle-node. The method is applied to the forced pendulum, or Josephson junction, equation $\beta\ddot{\phi}+\dot{\phi}+\sin\phi=\rho+\varepsilon\sin\omega t$.

Key words. Melnikov's method, saddle-node separatrix-loop bifurcation, pendulum, Josephson junction

AMS(MOS) subject classification. 58F14

1. Introduction. Melnikov's method [6], [3] is an analytic technique for showing the existence of transverse homoclinic orbits, which imply the complicated dynamics associated with horseshoes. The method applies to time-periodic perturbations of an autonomous planar vector field having a separatrix loop at a saddle point. In this work we shall extend Melnikov's method to the case in which the unperturbed vector field has a separatrix loop at a saddle-node.

Planar vector fields having a saddle-node separatrix loop occur generically in two-parameter families. It therefore seems natural to study three-parameter problems:

$$\begin{aligned}&\dot{x}=\tilde{f}(x,\nu_1,\nu_2)+\tilde{g}(x,\nu_1,\nu_2,\nu_3,t),\\&x\in\mathbb{R}^2,\qquad \nu_1,\nu_2,\nu_3\in\mathbb{R},\end{aligned}\tag{1}$$

with $\tilde{g}(x,\nu_1,\nu_2,0,t)\equiv 0$ and $\tilde{g}$ T-periodic in t. Here $\dot{x}=\tilde{f}(x,0,0)$ has a saddle-node separatrix loop, and $\dot{x}=\tilde{f}(x,\nu_1,\nu_2)$ is a generic unfolding. The analytic condition for such a generic unfolding was derived in [9]. There we studied the analogue of Melnikov's method for autonomous perturbations of a saddle-node separatrix loop.

(Similarly, saddle separatrix loops occur generically in one-parameter families of autonomous planar vector fields. Therefore, in the study of the usual Melnikov method it is natural to consider two-parameter problems in which the second parameter corresponds to a small time-periodic perturbation. This point of view is taken, for example, in [2] and [7].)

The motivating problem for this work is the pendulum equation with linear damping, a constant applied torque, and a small sinusoidal applied torque:

$$\beta\ddot{\phi}+\dot{\phi}+\sin\phi=\rho+\varepsilon\sin\omega t.$$

The same differential equation describes the AC–DC current-driven point Josephson junction. Setting $y=\dot{\phi}$, we have the system

$$\begin{aligned}&\dot{\phi}=y,\\&\dot{y}=\frac{1}{\beta}(-y-\sin\phi+\rho+\varepsilon\sin\omega t).\end{aligned}\tag{2}$$

There is a number $\beta_0>0$ such that the system (2) with $\rho=1$, $\beta=\beta_0$, $\varepsilon=0$ has a saddle-node separatrix loop. It is shown in [9] that the two-parameter autonomous unfolding obtained from (2) by setting $\varepsilon=0$ is generic. Thus (2) has the form (1) with $\nu_1=\rho-1$, $\nu_2=\beta-\beta_0$, $\nu_3=\varepsilon$. It follows from [5] and the calculations of [9] that there is a curve $\beta=\beta(\rho)$, $0<\rho\leqq 1$, having a quadratic tangency with the line $\rho=1$ at $(1,\beta_0)$,

* Received by the editors September 23, 1985; accepted for publication (in revised form) December 10, 1986.

† Mathematics Department, North Carolina State University, Raleigh, North Carolina 27695-8205.

such that (2) with $\beta=\beta(\rho)$, $0<\rho<1$, $\varepsilon=0$, has a saddle separatrix loop. In [8] Salam and Sastry applied Melnikov's method to (2) at these saddle separatrix loops. For fixed ρ, $0<\rho<1$, they found that for most ω (all but a discrete set), there is an $\mathcal{O}(\varepsilon)$-sized β-interval containing $\beta(\rho)$ such that for β in this interval, (2) has a transverse homoclinic orbit. A consequence of the present work is that, roughly speaking, the size of this interval remains uniformly $\mathcal{O}(\varepsilon)$ as $\rho\to 1$. This conclusion is consistent with Fig. 12 of [8]. Salam and Sastry speculate [8, p. 794] that for fixed ε, the size of the interval decreases as $\rho\to 1$. Whether or not this is so can presumably be decided by computing a second-order Melnikov function, but we do not pursue this point in the present paper.

The usual Melnikov method requires the evaluation of an integral around the separatrix loop; the integral involves the first-order approximation of the time-periodic perturbation. The present case is basically the same, except that one perturbation parameter must be rescaled as in [9], and certain boundary terms that go to zero in the saddle case must be retained.

In most applications of Melnikov's method the separatrix loop is known explicitly, and the required integral is explicitly calculated. The saddle separatrix loops of (2), however, are not known explicitly. Nevertheless Salam and Sastry succeed in [8] in determining enough about the integral to derive their results. Here also the saddle-node separatrix loop of (2) with $\rho=1$, $\beta=\beta_0$, $\varepsilon=0$ is not known explicitly, but enough can be calculated to derive the results.

This paper is organized as follows. In § 2 the results about equation (1) are described. The functions in terms of which the results are stated are specified more precisely in § 3. Proofs are given in § 4 through § 7. In § 8 the results are applied to (2).

2. Melnikov's method at a saddle-node. We shall consider the three-parameter problem (1) with $\tilde{g}(x,\nu_1,\nu_2,0,t)\equiv 0$, $\tilde{g}$ T-periodic in t. We assume:

(i) $\tilde{f}$ and $\tilde{g}$ are C^{k+1}, $k\geqq 5$,

(ii) $\tilde{f}(p,0,0)=0$,

(iii) $D_x\tilde{f}(p,0,0)$ has eigenvalues 0 and $-\lambda$, where $\lambda>0$.

Let u be a right eigenvector and w a left eigenvector of the eigenvalue 0, with w chosen so that $wu>0$.

(iv) $wD_x^2\tilde{f}(p,0,0)(u,u)>0$,

(v) $\dot{x}=\tilde{f}(x,0,0)$ has a separatrix loop Γ at p,

(vi) $wD_{\nu_1}\tilde{f}(p,0,0)>0$.

As in [9], these assumptions imply that $\dot{x}=\tilde{f}(x,0,0)$ has a saddle-node at p with one negative eigenvalue. The vector u is one tangent vector to Γ at p. As in [9] we let v denote a right eigenvector of $D_x\tilde{f}(p,0,0)$ for the eigenvalue $-\lambda$, chosen so that v is also tangent to Γ at p. See [9, Fig. 1]. Perturbation in the positive ν_1-direction eliminates the equilibrium p, while perturbation in the negative ν_1-direction splits the equilibrium in two.

For fixed $\nu=(\nu_1,\nu_2,\nu_3)$, let $\tilde{P}_\nu:\mathbb{R}^2\to\mathbb{R}^2$ be the time $(0,T)$ advance map of (2). Then $\tilde{P}_0$ has a fixed point of saddle-node type at p. In fact, assumptions (i) to (iv) and (vi) imply that there is a C^k function $\alpha(\nu_2,\nu_3)$, with $\alpha(0,0)=0$, such that $\tilde{P}_\nu$ has a fixed point of saddle-node type near p if and only if $\nu_1=\alpha(\nu_2,\nu_3)$. A fixed point of $\tilde{P}_\nu$ corresponds to a T-periodic solution of (1).

Let $f(x,\mu_1,\mu_2)=\tilde{f}(x,\nu_1,\nu_2)$, where $\mu_1=\nu_1-\alpha(\nu_2,0)$ and $\mu_2=\nu_2$. Let $g(x,\mu_1,\mu_2,\mu_3,t)=\tilde{g}(x,\nu_1,\nu_2,\nu_3,t)$, where $\mu_1=\nu_1-\alpha(\nu_2,\nu_3)$, $\mu_2=\nu_2$, $\mu_3=\nu_3$. We consider

$$\dot{x}=f(x,\mu_1,\mu_2)+g(x,\mu_1,\mu_2,\mu_3,t). \tag{3}$$

Notice f and g are C^k. For fixed $\mu=(\mu_1, \mu_2, \mu_3)$, let $P_\mu:\mathbb{R}^2\to\mathbb{R}^2$ be the time $(0, T)$ advance map of (3). Then P_μ has a fixed point of saddle-node type near p if and only if $\mu_1=0$. If $\mu_1<0$, then P_μ has two fixed points, one a saddle and one a sink, near p. If $\mu_1>0$ there are no fixed points of P_μ near p.

We shall show in § 3 that the fixed points of the maps P_μ can be parameterized as follows. There is a C^{k-2} mapping $p(\delta, \mu_2, \mu_3)$, from a neighborhood of $(0, 0, 0)$ in $\mathbb{R}^3$ to $\mathbb{R}^2$, such that $p(0, 0, 0)=p$; $p(0, \mu_2, \mu_3)$ is the fixed point of saddle-node type of $P_{(0,\mu_2,\mu_3)}$ near p; $p(\delta, \mu_2, \mu_3)$ is the fixed point of saddle-type of $P_{(-\delta^2,\mu_2,\mu_3)}$ near p if $\delta>0$, and is the sink of $P_{(-\delta^2,\mu_2,\mu_3)}$ near p if $\delta<0$.

We define $p(\delta, \mu_2, \mu_3, t)$ to be the T-periodic solution of

$$\dot{x}=f(x, -\delta^2, \mu_2)+g(x, -\delta^2, \mu_2, \mu_3, t) \tag{4}$$

that passes through $p(\delta, \mu_2, \mu_3)$ at $t=0$. Thus for $\delta\geqq 0$ there might exist solutions of (4) homoclinic to $p(\delta, \mu_2, \mu_3, t)$. For $\delta>0$, a homoclinic orbit is of course an orbit in the intersection of the stable and unstable manifolds of $p(\delta, \mu_2, \mu_3, t)$; for $\delta=0$, we consider only homoclinic orbits in the intersection of the stable and center manifolds of $p(0, \mu_2, \mu_3, t)$. We shall define in § 3 a C^{k-2} function $d(\delta, \mu_2, \mu_3, t_0)$, T-periodic in t_0, such that (4) has such a homoclinic orbit if and only if for $\delta\geqq 0$ and for some t_0, $d(\delta, \mu_2, \mu_3, t_0)=0$. Moreover, $d(\delta, \mu_2, 0, t_0)$ is independent of t_0, and $d(0, 0, 0, t_0)\equiv 0$. The homoclinic orbit is transverse if, in addition, $(\partial d/\partial t_0)(\delta, \mu_2, \mu_3, t_0)\neq 0$.

If $w=(w_1, w_2)$ and $z=(z_1, z_2)$, we define $w\wedge z=w_1z_2-w_2z_1$. The following theorem gives formulas for the partial derivatives of d at $(0, 0, 0, t_0)$.

THEOREM 1.

$$\frac{\partial d}{\partial\delta}(0, 0, 0, t_0)=\frac{\partial p}{\partial\delta}(0, 0, 0)\wedge\lim_{t_1\to\infty} f(q(t_1), 0, 0)\exp\left[-\int_0^{t_1}\operatorname{div} f(q(s), 0, 0)\,ds\right],$$

a negative multiple of $u\wedge v$ *that is independent of* t_0.

$$\begin{aligned}\frac{\partial d}{\partial\mu_2}(0, 0, 0, t_0)=&\frac{\partial p}{\partial\mu_2}(0, 0, 0)\wedge\lim_{t_1\to\infty} f(q(t_1), 0, 0)\exp\left[-\int_0^{t_1}\operatorname{div} f(q(s), 0, 0)\,ds\right]\\&+\int_{-\infty}^{\infty}\exp\left[-\int_0^{t}\operatorname{div} f(q(s), 0, 0)\,ds\right]f(q(t), 0, 0)\\&\wedge\frac{\partial f}{\partial\mu_2}(q(t), 0, 0)\,dt\end{aligned}$$

independent of t_0, *where the limit is a negative multiple of* v *and the integral converges.*

$$\begin{aligned}\frac{\partial d}{\partial\mu_3}(0, 0, 0, t_0)=\lim_{t_1\to\infty}\Bigg\{&\frac{\partial p}{\partial\mu_3}(0, 0, 0, t_1)\wedge f(q(t_1-t_0), 0, 0)\\&\cdot\exp\left[-\int_{t_0}^{t_1}\operatorname{div} f(q(s-t_0), 0, 0)\,ds\right]\\&+\int_{t_0}^{t_1}\exp\left[-\int_{t_0}^{t}\operatorname{div} f(q(s-t_0), 0, 0)\,ds\right]f(q(t-t_0), 0, 0)\\&\wedge\frac{\partial g}{\partial\mu_3}(q(t-t_0), 0, 0, t)\,dt\Bigg\}.\end{aligned}$$

The limit is finite. Each summand is asymptotically periodic of period T in t_1 as $t_1 \to \infty$. In particular, $(\partial p/\partial \mu_3)$ $(0, 0, 0, t_1)$ *has period T in t_1, and*

$$f(q(t_1 - t_0), 0, 0) \exp\left[-\int_{t_0}^{t_1} \operatorname{div} f(q(s - t_0), 0, 0)\, ds \right]$$

approaches a finite negative multiple of v as $t_1 \to \infty$, independent of t_0.

If the constant $(\partial d/\partial \mu_2)$ $(0, 0, 0, t_0) \neq 0$, then there is a C^{k-2} function $\mu_2(\delta)$ such that for δ and μ_2 small, $d(\delta, \mu_2, 0, t_0) = 0$ if and only if $\mu_2 = \mu_2(\delta)$. If $(\partial d/\partial \mu_2)(0, 0, 0, t_0) \neq 0$, we define the Melnikov function

$$M(t_0) = -\frac{\partial d}{\partial \mu_3}(0, 0, 0, t_0) / \frac{\partial d}{\partial \mu_2}(0, 0, 0, t_0).$$

THEOREM 2. *Assume* (1) $(\partial d/\partial \mu_2)(0, 0, 0, t_0)$, *a constant independent of t_0 by Theorem* 1, *is not zero*; (2) $M(t_0)$ *attains its maximum* (*resp. minimum*) *value on* $0 \leqq t_0 < T$ *at a unique* $t_0^{\max}$ (*resp.* $t_0^{\min}$), *and* $M(t_0)$ *has a nondegenerate extremum at* $t_0^{\max}$ (*resp.* $t_0^{\min}$); (3) *if* $0 \leqq t_0 < T$, $t_0 \neq t_0^{\min}$, $t_0^{\max}$, *then* $M'(t_0) \neq 0$. *Then there are functions* $\gamma_*(\delta, \mu_3)$ *and* $\gamma^*(\delta, \mu_3)$, *with* $\gamma_*(\delta, 0) \equiv \gamma^*(\delta, 0) \equiv 0$ *and* $(\partial \gamma_*/\partial \mu_3)(0, 0) = M(t_0^{\min})$, $(\partial \gamma^*/\partial \mu_3)(0, 0) = M(t_0^{\max})$, *such that, for* (δ, μ_2, μ_3) *small, the equation* $d(\delta, \mu_2, \mu_3, t_0) = 0$ *has a solution if and only if either*

(a) $\mu_3 \geqq 0$ *and* $\gamma_*(\delta, \mu_3) \leqq \mu_2 - \mu_2(\delta) \leqq \gamma^*(\delta, \mu_3)$, *or*

(b) $\mu_3 \leqq 0$ *and* $\gamma^*(\delta, \mu_3) \leqq \mu_2 - \mu_2(\delta) \leqq \gamma_*(\delta, \mu_3)$.

Therefore, there is a homoclinic orbit of (3) *asymptotic to the periodic solution* $p(\sqrt{-\mu_1}, \mu_2, \mu_3, t)$ *if and only if either*

(a) $\mu_3 \geqq 0$ *and* $\gamma_*(\sqrt{-\mu_1}, \mu_3) \leqq \mu_2 - \mu_2(\sqrt{-\mu_1}) \leqq \gamma^*(\sqrt{-\mu_1}, \mu_3)$, *or*

(b) $\mu_3 \leqq 0$ *and* $\gamma^*(\sqrt{-\mu_1}, \mu_3) \leqq \mu_2 - \mu_2(\sqrt{-\mu_1}) \leqq \gamma_*(\sqrt{-\mu_1}, \mu_3)$.

There is a transverse homoclinic orbit of (3) *asymptotic to this periodic solution if and only if* (a) *or* (b) *holds with all inequalities replaced by strict inequalities.*

See Fig. 1. The upper and lower curves in Fig. 1 meet the μ_2-axis at $\mu_2 = \gamma^*(0, \mu_3) \approx \mu_3 M(t_0^{\max})$ and $\mu_2 = \gamma_*(0, \mu_3) \approx \mu_3 M(t_0^{\min})$. In drawing Fig. 1 we have assumed $\mu_3 > 0$ and $M(t_0^{\min}) < 0 < M(t_0^{\max})$.

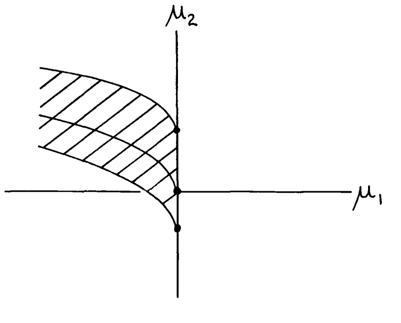


FIG. 1

Theorem 2 remains true if (2) and (3) are replaced by the assumption that $M(t_0)$ is a Morse function (all critical points nondegenerate, no two have the same value of M). Slightly weaker assumptions also suffice. A similar approach to the usual saddle case is in [2] and [7].

3. Functions used in the statements of Theorems 1 and 2. In this section we describe more precisely the functions $p(\delta, \mu_2, \mu_3)$, $p(\delta, \mu_2, \mu_3, t)$, and $d(\delta, \mu_2, \mu_3, t_0)$ used in the statements of Theorems 1 and 2.

We first note that the system

$$\begin{aligned} &\dot{x} = f(x, \mu_1, \mu_2) + g(x, \mu_1, \mu_2, \mu_3, t), \\ &\dot{\mu}_i = 0 \qquad (i = 1, 2, 3) \end{aligned} \tag{5}$$

has the constant solution

$$(x(t), \mu_1(t), \mu_2(t), \mu_3(t)) \equiv (p, 0, 0, 0). \tag{6}$$

This T-periodic solution of (5) has a five-dimensional C^k local center manifold $N_{\text{loc}} \subset \mathbb{R}^2 \times \mathbb{R}^3 \times \mathbb{R}$; the last variable is time t. (Note that a solution of (5) is a curve in $x\mu$-space, but an invariant manifold of (5) is a submanifold of $x\mu t$-space.) N_{loc} meets each space $\mathbb{R}^2 \times \{(\mu_1, \mu_2, \mu_3)\} \times \mathbb{R}$, (μ_1, μ_2, μ_3) small, in a two-dimensional surface. For $(\mu_1, \mu_2, \mu_3) = (0, 0, 0)$, this surface contains

$$(\text{a portion of } \Gamma \text{ tangent at } p \text{ to } u \times \{(0, 0, 0)\}) \times \mathbb{R}.$$

N_{loc} is T-periodic in t [1, § 9.4].

We shall now give a useful parameterization of the center manifold. Let $q(t)$ be a solution of $\dot{x} = f(x, 0, 0)$ with $q(0) \in \Gamma$. Let N denote the "global" center manifold that contains N_{loc}, i.e., the invariant manifold of (5) obtained from N_{loc} by completing solution curves. N meets each space $\mathbb{R}^2 \times \{(\mu_1, \mu_2, \mu_3)\} \times \mathbb{R}$, (μ_1, μ_2, μ_3) small, in a two-dimensional surface, which we identify with a surface $N(\mu_1, \mu_2, \mu_3)$ in $\mathbb{R}^2 \times \mathbb{R}$ (xt-space). Let L be a line segment in $\mathbb{R}^2$ perpendicular to Γ at $q(0)$. Then the surface $N(0, 0, 0)$ meets $L \times \mathbb{R}$ transversally in $\mathbb{R}^2 \times \mathbb{R}$ along the line $\{q(0)\} \times \mathbb{R}$. Therefore for (μ_1, μ_2, μ_3) small there is a C^k function $x(\mu_1, \mu_2, \mu_3, t_0)$, T-periodic in t_0, such that $x(0, 0, 0, t_0) \equiv q(0)$, and $x(\mu_1, \mu_2, \mu_3, t_0) \in N(\mu_1, \mu_2, \mu_3) \cap L \times \{t_0\}$. Since a C^k vector field has a C^k flow, there is a C^k family of solutions of (3):

$$q^c(\mu_1, \mu_2, \mu_3, t_0, t), \qquad (\mu_1, \mu_2, \mu_3) \text{ small},$$

such that $q^c(\mu_1, \mu_2, \mu_3, t_0, t_0) = x(\mu_1, \mu_2, \mu_3, t_0)$. Thus

$$q^c(\mu_1, \mu_2, \mu_3, t_0 + T, t + T) = q^c(\mu_1, \mu_2, \mu_3, t_0, t).$$

Notice $q^c(0, 0, 0, t_0, t) = q(t - t_0)$, and the curve

$$\{(q^c(\mu_1, \mu_2, \mu_3, t_0, t), t) \colon t \in \mathbb{R}\} \tag{7}$$

lies in $N(\mu_1, \mu_2, \mu_3)$. For $\mu_1 < 0$, the curve (7) lies in the unstable manifold of a T-periodic solution of saddle-type of (3); for $\mu_1 = 0$, it lies in the center manifold of a T-periodic solution of saddle-node type of (3). (Similar terminology is used in [9], except that the problem studied there is autonomous. See Fig. 3 of [9].)

We next describe coordinates that simplify the differential equation. According to [1, § 9.4] there is a C^k change of coordinates

$$y(x, \mu_1, \mu_2, \mu_3, t) = (y_1(x, \mu_1, \mu_2, \mu_3, t), y_2(x, \mu_1, \mu_2, \mu_3, t)), \tag{8}$$

T-periodic in t, defined for $\|x - p\|$, μ_1, μ_2, μ_3 small, such that (1) $y(p, 0, 0, 0, t) \equiv 0$; (2) $N_{\text{loc}} \cap \mathbb{R}^2 \times \{(\mu_1, \mu_2, \mu_3, t)\}$ is transformed into the line $y_2 = 0$; (3) the lines $y_1 =$ constant in $\mathbb{R}^2 \times \{(\mu_1, \mu_2, \mu_3)\}$ are mapped into each other by the (t_1, t_2)-advance maps. In other words, in the new coordinates (3) becomes the C^k differential equation

$$\begin{aligned} \dot{y}_1 &= a(y_1, \mu_1, \mu_2) + \mu_3 b(y_1, \mu_1, \mu_2, \mu_3, t), \\ \dot{y}_2 &= y_2[c(y_1, y_2, \mu_1, \mu_2) + \mu_3\, e(y_1, y_2, \mu_1, \mu_2, \mu_3, t)]. \end{aligned}$$

Since the stable manifold of the constant solution (6) necessarily is transformed into the plane $y_1=0$ in $\mathbb{R}^2\times\{(0,0,0)\}\times\mathbb{R}$, it is easy to arrange that

$$D_x y(p,0,0,0,t)u\equiv(1,0),\qquad D_x y(p,0,0,0,t)v\equiv(0,1).$$

We may also assume the coordinates are chosen so that for each (μ_2,μ_3), the fixed point of $P_{(0,\mu_2,\mu_3)}$ near p corresponds to $y_1=y_2=0$.

Consider the system

$$\begin{aligned}\dot y_1&=a(y_1,\mu_1,\mu_2)+\mu_3 b(y_1,\mu_1,\mu_2,\mu_3,t),\\ \dot y_2&=y_2[c(y_1,y_2,\mu_1,\mu_2)+\mu_3\, e(y_1,y_2,\mu_1,\mu_2,\mu_3,t)],\\ \dot\mu_i&=0,\qquad i=1,2,3.\end{aligned}\tag{9}$$

As in [9] it can be shown that the fixed points of the time $(0,T)$ advance map of (9) near $((0,0),0,0,0)$ comprise a set of the form

$$\{((\hat p_1(\delta,\mu_2,\mu_3),0),-\delta^2,\mu_2,\mu_3):(\delta,\mu_2,\mu_3)\text{ small}\},$$

with $\hat p_1$ of class C^{k-2},

$$\hat p_1(0,\mu_2,\mu_3)\equiv0\quad\text{and}\quad\frac{\partial p_1}{\partial\delta}(0,\mu_2,\mu_3)>0.$$

Let $(\hat p_1(\delta,\mu_2,\mu_3,t),0)$ denote the T-periodic solution of

$$\begin{aligned}\dot y_1&=a(y_1,-\delta^2,\mu_2)+\mu_3 b(y_1,-\delta^2,\mu_2,\mu_3,t),\\ \dot y_2&=y_2[c(y_1,y_2,-\delta^2,\mu_2)+\mu_3\, e(y_1,y_2,-\delta^2,\mu_2,\mu_3,t)]\end{aligned}\tag{10}$$

that has $\hat p_1(\delta,\mu_2,\mu_3,0)=\hat p_1(\delta,\mu_2,\mu_3)$.

Let $x=x(y,\mu_1,\mu_2,\mu_3,t)$ be the change of coordinates inverse to (8). Define

$$p(\delta,\mu_2,\mu_3)=x((\hat p_1(\delta,\mu_2,\mu_3),0),-\delta^2,\mu_2,\mu_3,0).$$

Then $p(\delta,\mu_2,\mu_3)$ is C^{k-2}, $p(0,0,0)=p$, and $p(\delta,\mu_2,\mu_3)$ is a fixed point of saddle-node type of $P_{(0,\mu_2,\mu_3)}$ if $\delta=0$; a fixed point of saddle type of $P_{(-\delta^2,\mu_2,\mu_3)}$ if $\delta>0$; a sink of $P_{(-\delta^2,\mu_2,\mu_3)}$ if $\delta<0$. Let $p(\delta,\mu_2,\mu_3,t)=x((\hat p_1(\delta,\mu_2,\mu_3,t),0),-\delta^2,\mu_2,\mu_3,t)$, the T-periodic solution of (4) with $p(\delta,\mu_2,\mu_3,0)=p(\delta,\mu_2,\mu_3)$.

System (10) has at the T-periodic solution $(\hat p_1(\delta,\mu_2,\mu_3,t),0)$ the invariant manifold $\{(y_1,y_2,t):y_1=\hat p_1(\delta,\mu_2,\mu_3,t)\}$. For $\delta=0$, this surface is the stable manifold of the saddle-node T-periodic solution $(\hat p_1(0,\mu_2,\mu_3,t),0)$ of (10); for $\delta>0$ it is the stable manifold of the saddle T-periodic solution $(\hat p_1(\delta,\mu_2,\mu_3,t),0)$ of (10); and for $\delta<0$, it is the strong stable manifold of the attracting T-periodic solution $(\hat p_1(\delta,\mu_2,\mu_3,t),0)$ of (10). Returning to x-coordinates, we find that (4) has at the T-periodic solution $p(\delta,\mu_2,\mu_3,t)$ the local invariant manifold $\{(x((y_1,y_2),-\delta^2,\mu_2,\mu_3,t),t):y_1=\hat p_1(\delta,\mu_2,\mu_3,t)\}$.

For each (δ,μ_2,μ_3) the local invariant manifold of (4) constructed above extends to a "global" invariant manifold $Q(\delta,\mu_2,\mu_3)$ by completing solution curves. $Q(\delta,\mu_2,\mu_3)$ is a two-dimensional submanifold of $\mathbb{R}^2\times\mathbb{R}$ that depends C^{k-2} on (δ,μ_2,μ_3). Notice that $Q(0,0,0)$ contains $\Gamma\times\mathbb{R}$. Now a construction similar to that of $q^c(\mu_1,\mu_2,\mu_3,t)$ yields a C^{k-2} family

$$q^s(\delta,\mu_2,\mu_3,t_0,t),\qquad(\delta,\mu_2,\mu_3)\text{ small},$$

each a solution of (4), such that $q^s(\delta,\mu_2,\mu_3,t_0,t_0)\in L$, $(q^s(\delta,\mu_2,\mu_3,t_0,t),t)\in Q(\delta,\mu_2,\mu_3)$, $q^s(0,0,0,t_0,t)=q(t-t_0)$, and $q^s(\delta,\mu_2,\mu_3,t_0+T,t+T)=q^s(\delta,\mu_2,\mu_3,t_0,t)$.

We can now discuss homoclinic solution of (1). For any vector $w=(w_1, w_2)\in R^2$, recall that $w^\perp=(-w_2, w_1)$. If $z=(z_1, z_2)$, let $w\wedge z=w_1z_2-w_2z_1=w^\perp\cdot z$. Define $d^c(\mu_1, \mu_2, \mu_3, t_0)$ and $d^s(\delta, \mu_2, \mu_3, t_0)$ by

$$q^c(\mu_1, \mu_2, \mu_3, t_0, t_0)=q(0)+[d^c(\mu_1, \mu_2, \mu_3, t_0)/\|f(q(0), 0, 0)\|^2]f^\perp(q(0), 0, 0),$$

$$q^s(\delta, \mu_2, \mu_3, t_0, t_0)=q(0)+[d^s(\delta, \mu_2, \mu_3, t_0)/\|f(q(0), 0, 0)\|^2]f^\perp(q(0), 0, 0).$$

Then d^c is C^k and d^s is C^{k-2}. We have

$$\begin{aligned}d^c(\mu_1, \mu_2, \mu_3, t_0)&=f^\perp(q(0), 0, 0)\cdot[q^c(\mu_1, \mu_2, \mu_3, t_0, t_0)-q(0)]\\&=f(q(0), 0, 0)\wedge[q^c(\mu_1, \mu_2, \mu_3, t_0, t_0)-q(0)].\end{aligned}$$

Similarly,

$$d^s(\delta, \mu_2, \mu_3, t_0)=f(q(0), 0, 0)\wedge[q^s(\delta, \mu_2, \mu_3, t_0, t_0)-q(0)].$$

There is a homoclinic orbit of (4) asymptotic to the T-periodic solution $p(\delta, \mu_2, \mu_3, t)$ if and only if $\delta\geqq 0$ and for some t_0,

$$d(\delta, \mu_2, \mu_3, t_0)\underset{\text{def}}{=}d^c(-\delta^2, \mu_2, \mu_3, t_0)-d^s(\delta, \mu_2, \mu_3, t_0)=0.$$

Here $d(\delta, \mu_2, \mu_3, t_0)$ is C^{k-2}. The homoclinic orbit is transverse if, in addition, $(\partial d/\partial t_0)(\delta, \mu_2, \mu_3, t_0)\neq 0$.

4. Proof of Theorem 1. Since for $\mu_3=0$ the perturbation is autonomous, the formulas for $\partial d/\partial\delta$ and $\partial d/\partial\mu_2$ follow immediately from [9]. We shall derive the formula for $\partial d/\partial\mu_3$. To simplify the notation, we set

$$\begin{aligned}&f(x, \mu_3)=f(x, 0, 0, \mu_3),\\&g(x, \mu_3, t)=g(x, 0, 0, \mu_3, t),\\&p(\mu_3)=p(0, 0, \mu_3),\\&p(\mu_3, t)=p(0, 0, \mu_3, t),\\&q^{c,s}(\mu_3, t_0, t)=q^{c,s}(0, 0, \mu_3, t_0, t),\\&d^{c,s}(\mu_3, t_0)=d^{c,s}(0, 0, \mu_3, t_0).\end{aligned}$$

For $q^{c,s}(\mu_3, t_0, t)$, we have the variational equations

$$\frac{d}{dt}\frac{\partial q^{c,s}}{\partial\mu_3}(0, t_0, t)=D_xf(q(t-t_0), 0)\frac{\partial q^{c,s}}{\partial\mu_3}(0, t_0, t)+\frac{\partial g}{\partial\mu_3}(q(t-t_0), 0, t).$$

Define

$$(11^{c,s})\qquad \Delta_{\mu_3}^{c,s}(t_0, t)=f(q(t-t_0), 0)\wedge\frac{\partial q^{c,s}}{\partial\mu_3}(0, t_0, t).$$

For $d^{c,s}(\mu_3, t_0)$ we have the derivative formulas

$$\frac{\partial d^{c,s}}{\partial\mu_3}(0, t_0)=\Delta_{\mu_3}^{c,s}(t_0, t_0).$$

Using the variational equations for q^c and q^s, we compute as in [3]:

$$\frac{d}{dt}\Delta_{\mu_3}^{c,s}(t_0, t)=\operatorname{div}f(q(t-t_0), 0)\,\Delta_{\mu_3}^{c,s}(t_0, t)+f(q(t-t_0), 0)\wedge\frac{\partial g}{\partial\mu_3}(q(t-t_0), 0, t).$$

Solving these linear differential equations, we obtain, for any t_1

$$\begin{aligned}\Delta^{c,s}_{\mu_3}(t_0, t_0) = \Delta^{c,s}_{\mu_3}(t_0, t_1) \exp\left[-\int_{t_0}^{t_1} \operatorname{div} f(q(t-t_0), 0)\, dt \right] \\ -\int_{t_0}^{t_1} \exp\left[-\int_{t_0}^{t} \operatorname{div} f(q(s-t_0), 0)\, ds \right] \\ \cdot f(q(t-t_0), 0) \wedge \frac{\partial g}{\partial \mu_3}(q(t-t_0), 0, t)\, dt.\end{aligned} \tag{$12^{c,s}$}$$

We define new C^k coordinates $y(x, \mu_3, t)$, T-periodic in t, on $\mathbb{R}^2$ near p, such that

$$y(p(\mu_3, t), \mu_3, t) \equiv 0, \tag{13}$$

the local stable manifold of the T-periodic solution $p(\mu_3, t)$ becomes the plane $y_1 = 0$ in yt-space, and the local center manifold of $p(\mu_3, t)$ becomes the plane $y_2 = 0$ in yt-space. (This coordinate change is a little different from that used in § 2.)

We shall first study (12^s) in the limit $t_1 \to \infty$. Define

$$\begin{aligned}\tilde{q}^s(\mu_3, t_0, t) &= y(q^s(\mu_3, t_0, t), \mu_3, t) \\ &= (0, y_2(\mu_3, t_0, t)).\end{aligned} \tag{14}$$

Since $q^s(\mu_3, t_0, t) \to p(\mu_3, t)$ as $t \to \infty$, for each μ_3 near 0, $\tilde{q}^s(\mu_3, t_0, t)$ and hence $y_2(\mu_3, t_0, t)$ are defined for sufficiently large t.

LEMMA 1. $(\partial y_2/\partial \mu_3)(0, t_0, t) \to 0$ *as* $t \to \infty$.

We shall postpone the proof of Lemma 1 to § 5. From Lemma 1 and (14) it follows immediately that

$$\frac{\partial q^s}{\partial \mu_3}(0, t_0, t) \to 0 \quad \text{as } t \to \infty. \tag{15}$$

By (14),

$$\frac{\partial q^s}{\partial \mu_3}(0, t_0, t) = D_x y(q(t-t_0), 0, t) \frac{\partial q^s}{\partial \mu_3}(0, t_0, t) + \frac{\partial y}{\partial \mu_3}(q(t-t_0), 0, t). \tag{16}$$

By (13),

$$D_x y(p, 0, t) \frac{\partial p}{\partial \mu_3}(0, t) + \frac{\partial y}{\partial \mu_3}(p, 0, t) = 0. \tag{17}$$

Let $t \to \infty$ in (16). Then $q(t-t_0) \to p$, so (15), (16), (17), and the invertibility of $D_x y$ imply

$$\lim_{t \to \infty} \left[\frac{\partial q^s}{\partial \mu_3}(0, t_0, t) - \frac{\partial p}{\partial \mu_3}(0, t) \right] = 0. \tag{18}$$

Since $(\partial p/\partial \mu_3)(0, t)$ is T-periodic in t, $(\partial q^s/\partial \mu_3)(0, t_0, t)$ is asymptotically T-periodic in t as $t \to \infty$.

Using (11^s) we rewrite (12^s) as

$$\begin{aligned}-\Delta^s_{\mu_3}(t_0, t) = \frac{\partial q^s}{\partial \mu_3}(0, t_0, t_1) \wedge f(q(t_1-t_0), 0) \exp\left[-\int_{t_0}^{t_1} \operatorname{div} f(q(t-t_0), 0)\, dt \right] \\ +\int_{t_0}^{t_1} \exp\left[-\int_{t_0}^{t} \operatorname{div} f(q(s-t_0), 0)\, ds \right] \\ \cdot f(q(t-t_0), 0) \wedge \frac{\partial g}{\partial \mu_3}(q(t-t_0), 0, t)\, dt.\end{aligned} \tag{19}$$

Now,

$$\lim_{t_1\to\infty} f(q(t_1-t_0),0)\exp\left[-\int_{t_0}^{t_1}\operatorname{div} f(q(t-t_0),0)\,dt\right]$$
$$=\lim_{t_1\to\infty} f(q(t_1-t_0),0)\exp\left[-\int_{0}^{t_1-t_0}\operatorname{div} f(q(s),0)\,ds\right]$$

is a negative multiple of v by [9]. Therefore (18) implies that the first summand of (19) is asymptotically periodic in t_1 as $t_1\to\infty$ with period T. Therefore the second summand of (19), the integral, is asymptotically periodic in t_1 as $t_1\to\infty$ with period T.

Next we shall study (12^c) in the limit $t_1\to-\infty$. Define

(20) $$\begin{aligned}\tilde{q}^c(\mu_3,t_0,t)&=y(q^c(\mu_3,t_0,t),\mu_3,t)\\&=(y_1(\mu_3,t_0,t),0).\end{aligned}$$

Since $q^c(\mu_3,t_0,t)\to p(\mu_3,t)$ as $t\to-\infty$, for each μ_3 near 0, $\tilde{q}^c(\mu_3,t_0,t)$ and hence $y_1(\mu_3,t_0,t)$ are defined for sufficiently negative t.

LEMMA 2. *$(\partial y_1/\partial\mu_3)(0,t_0,t)\to 0$ as $t\to-\infty$.*

We postpone the proof of Lemma 2 to § 6. Given Lemma 2, we prove as in [9] that the first summand of (12^c) approaches zero as $t_1\to-\infty$, so that the second summand of (12^c), the integral, approaches a limit as $t_1\to-\infty$. Therefore,

(21) $$\begin{aligned}\Delta^c_{\mu_3}(t_0,t_0)=\int_{-\infty}^{t_0}&\exp\left[-\int_{t_0}^{t}\operatorname{div} f(q(s-t_0),0)\,ds\right]f(q(t-t_0),0)\\&\wedge\frac{\partial g}{\partial\mu_3}(t-t_0,0,t)\,dt.\end{aligned}$$

The formula for $\partial d/\partial\mu_3$ in Theorem 1 now follows by adding (21) to (19) and using (18).

5. Proof of Lemma 1. To simplify notation, we fix t_0 and let $z(\mu_3,t)=y_2(\mu_3,t_0,t)$. Then $z(\mu_3,t)$ satisfies a differential equation of the form

(22) $$\dot z=-\lambda(\mu_3,t)z(1+zG(z,\mu_3,t)),$$

where $\lambda(0,t)\equiv\lambda>0$, $G(z,0,t)$ is independent of t, $\lambda(\mu_3,t)$ and $G(z,\mu_3,t)$ are T-periodic in t. When $\mu_3=0$ we can solve (22) by separation of variables as in [9] to obtain

(23) $$z(0,t)=\mathcal{O}(e^{-\lambda t})\quad\text{as } t\to\infty.$$

Using (23), we find that the variational equation for (22) along $z(0,t)$ is

$$\frac{d}{dt}\frac{\partial z}{\partial\mu_3}(0,t)=A(t)\frac{\partial z}{\partial\mu_3}(0,t)+B(t),$$

where $A(t)=-\lambda+\mathcal{O}(e^{-\lambda t})$ as $t\to\infty$, $B(t)=\mathcal{O}(e^{-\lambda t})$ as $t\to\infty$. Therefore

$$\begin{aligned}\frac{\partial z}{\partial\mu_3}(0,t)&=\int_0^t B(s)\exp\left(\int_s^t A(r)\,dr\right)ds\\&=\int_0^t\mathcal{O}(e^{-\lambda s})\exp\left(\int_s^t-\lambda+\mathcal{O}(e^{-\lambda r})\,dr\right)ds\\&=\int_0^t\mathcal{O}(e^{-\lambda s})e^{-\lambda(t-s)}\exp\left(\int_s^t\mathcal{O}(e^{-\lambda r})\,dr\right)ds\\&\leqq Kte^{-\lambda t}\to 0\quad\text{as } t\to\infty.\end{aligned}$$

6. Proof of Lemma 2. The function $y_1(\mu_3, t_0, t)$ satisfies a differential equation of the form

$$\dot{z} = h(z, \mu_3, t), \tag{24}$$

with h T-periodic in t and $h(0, \mu_3, t) \equiv 0$. We have

$$h(z, 0, t) = \eta z^2(1 + zH(z))$$

with $\eta > 0$. Moreover, for μ_3 small the time $(t_0, t_0 + T)$ advance map of (24) has the form

$$z \to z + A(\mu_3, t_0)z^2(1 + B(z, \mu_3, t_0)), \tag{25}$$

with $A(\mu_3, t_0) = A_1 + \mu_3 A_2(\mu_3, t_0)$, $B(z, \mu_3, t_0) = z[B_1(z) + \mu_3 B_2(z, \mu_3, t_0)]$, A_2 and B_2 T-periodic in t_0.

Let $\zeta(z, \mu_3, t_0, t)$ be the solution of (24) that has the value z at $t = t_0$. Then

$$\zeta(z, 0, t_0, t_0 + t) = \zeta(z, 0, 0, t) = z + \eta t z^2(1 + zC(z, t)). \tag{26}$$

Comparing (25) and (26), we see that

$$A_1 = \eta T \quad \text{and} \quad B_1(z) = C(z, t). \tag{27}$$

Choose numbers m and n such that $mT < A_2(\mu_3, t_0) < nT$ for μ_3 small and all t_0. Let $\eta_*(\mu_3) = \eta + m\mu_3$, $\eta^*(\mu_3) = \eta + n\mu_3$. Then for $\mu_3 > 0$ small and any t_0,

$$\eta_*(\mu_3)T < A(\mu_3, t_0) < \eta^*(\mu_3)T. \tag{28}$$

Consider, in addition to (25), the autonomous differential equations

$$\dot{z} = \eta_*(\mu_3)z^2(1 + zH(z)), \tag{29}$$

$$\dot{z} = \eta^*(\mu_3)z^2(1 + zH(z)). \tag{30}$$

The flows are

$$\zeta_*(z, \mu_3, t) = z + \eta_*(\mu_3)tz^2(1 + z[C(z, t) + \mu_3 C_*(z, \mu_3, t)]), \tag{31}$$

$$\zeta^*(z, \mu_3, t) = z + \eta^*(\mu_3)tz^2(1 + z[C(z, t) + \mu_3 C^*(z, \mu_3, t)]). \tag{32}$$

Therefore

$$\zeta^*(z, 0, t) = \zeta(z, 0, t_0, t_0 + t) = \zeta_*(z, 0, t) \tag{33}$$

for z small and any t_0, t. From (25), (27), (28), (31), (32), and $\eta > 0$, it follows that for $z > 0$ and $\mu_3 > 0$ small, any t_0, and any integer $N < 0$,

$$\zeta^*(z, \mu_3, NT) < \zeta(z, \mu_3, t_0, t_0 + NT) < \zeta_*(z, \mu_3, NT). \tag{34}$$

Now fix t_0 and consider $s \in [t_0 - T, t_0]$. Let $z(\mu_3, s) = y_1(\mu_3, t_0, s)$, so that

$$y_1(\mu_3, t_0, s + NT) = \zeta(z(\mu_3, s), \mu_3, s, s + NT). \tag{35}$$

Then (34) and (35) imply for $\mu_3 > 0$ small and any integer $N < 0$,

$$\zeta^*(z(\mu_3, s), \mu_3, NT) < y_1(\mu_3, t_0, s + NT) < \zeta_*(z(\mu_3, s), \mu_3, NT). \tag{36}$$

Also, (33) and (35) imply, for any integer N,

$$\zeta^*(z(0, s), 0, NT) = y_1(0, t_0, s + NT) = \zeta_*(z(0, s), 0, NT). \tag{37}$$

Let $\theta^*(\mu_3, s, t) = \zeta^*(z(\mu_3, s), \mu_3, t)$, $\theta_*(\mu_3, s, t) = \zeta_*(z(\mu_3, s), \mu_3, t)$. For each fixed (μ_3, s), $\theta^*(\mu_3, s, t)$ (resp. $\theta_*(\mu_3, s, t)$) is a solution of (30) (resp. (29)). Then (36) and (37) imply, for any integer $N < 0$,

$$(38)\qquad \frac{\partial \theta^*}{\partial \mu_3}(0, s, NT) \leqq \frac{\partial y_1}{\partial \mu_3}(0, t_0, s+NT) \leqq \frac{\partial \theta_*}{\partial \mu_3}(0, s, NT).$$

The proof of Lemma 3 in [9] can easily be adapted to show that

$$(39)\qquad \lim_{t\to-\infty} \frac{\partial \theta^*}{\partial \mu_3}(0, s, t) = \lim_{t\to-\infty} \frac{\partial \theta_*}{\partial \mu_3}(0, s, t) = 0,$$

and the limit is uniform in $s \in [t_0 - T, t_0]$. On the other hand, every $t < t_0$ can be expressed as $t = s + NT$, $s \in [t_0 - T, t_0]$, N a negative integer. Lemma 2 therefore follows from (38) and (39).

7. Proof of Theorem 2. Since $d(\delta, \mu_2(\delta), 0, t_0) \equiv 0$, we have

$$(40)\qquad \begin{aligned} d(\delta, \mu_2, \mu_3, t_0) = {} & \frac{\partial d}{\partial \mu_2}(\delta, \mu_2(\delta), 0, t_0)(\mu_2 - \mu_2(\delta)) \\ & + \frac{\partial d}{\partial \mu_3}(\delta, \mu_2(\delta), 0, t_0)\mu_3 \\ & + o(|\mu_2 - \mu_2(\delta)| + |\mu_3|). \end{aligned}$$

Since $(\partial d/\partial \mu_2)(0, 0, 0, t_0)$ is a nonzero constant, by the implicit function theorem there is a C^{k-2} function $\gamma(\delta, \mu_3, t_0)$, T-periodic in t_0, such that for δ, $\mu_2 - \mu_2(\delta)$, and μ_3 small, $d(\delta, \mu_2, \mu_3, t_0) = 0$ if and only if

$$(41)\qquad \mu_2 - \mu_2(\delta) = \gamma(\delta, \mu_3, t_0) = \gamma_1(\delta, t_0)\mu_3 + o(\mu_3).$$

For fixed (δ, μ_2, μ_3) there is a solution t_o of (41) if and only if

$$\min_{t_0} \gamma(\delta, \mu_3, t_0) \leqq \mu_2 - \mu_2(\delta) \leqq \max_{t_0} \gamma(\delta, \mu_2, t_0).$$

If $\mu_3 \neq 0$, an extremum with respect to t_0 of $\gamma(\delta, \mu_3, t_0)$ is also an extremum with respect to t_0 of

$$(42)\qquad \xi(\delta, \mu_3, t_0) = \frac{1}{\mu_3}\gamma(\delta, \mu_3, t_0) = \gamma_1(\delta, t_0) + \mathcal{O}(\mu_3),$$

ξ is T-periodic in t_0. At an extremum with respect to t_0 of ξ we have

$$\frac{\partial \xi}{\partial t_0}(\delta, \mu_3, t_0) = \frac{\partial \gamma_1}{\partial t_0}(\delta, t_0) + \mathcal{O}(\mu_3) = 0.$$

If we substitute (41) into (40), set $d = 0$, and solve to order μ_3, we obtain

$$(43)\qquad \gamma_1(\delta, t_0) = -\frac{\partial d}{\partial \mu_3}(\delta, \mu_2(\delta), 0, t_0) \Big/ \frac{\partial d}{\partial \mu_2}(\delta, \mu_2(\delta), 0, t_0).$$

(Of course this formula is also a consequence of the implicit function theorem.) Thus $\gamma_1(0, t_0) = M(t_0)$. By assumption, at $t_0 = t_0^{\min}$ we have $M'(t_0) = 0$ and $M''(t_0) > 0$. By the implicit function theorem there is a function $t_*(\delta, \mu_3)$, defined for (δ, μ_3) near $(0, 0)$, with $t_*(0, 0) = t_0^{\min}$, such that $(\partial \xi/\partial t_0)(\delta, \mu_3, t_*(\delta, \mu_3)) = 0$. Since $M(t_0)$ attains its minimum on $0 \leqq t < T$ uniquely at $t_0 = t_0^{\min}$, it follows easily from (42) that for (δ, μ_3) small, $\xi(\delta, \mu_3, t_0)$ attains its minimum on $0 \leqq t_0 < T$ uniquely at $t_*(\delta, \mu_3)$. Let $\gamma_*(\delta, \mu_3) = \gamma(\delta, \mu_3, t_*(\delta, \mu_3))$. A similar argument at $t = t_0^{\max}$ allows us to define $\gamma^*(\delta, \mu_3)$.

For a transverse homoclinic orbit we need $d(\delta, \mu_2, \mu_3, t_0) = 0$ and $(\partial d/\partial t_0)(\delta, \mu_2, \mu_3, t_0) \neq 0$. We shall now show that for δ, $\mu_2 - \mu_2(\delta)$, and $\mu_3 \neq 0$ all small, the only simultaneous solutions of $d = 0$ and $(\partial d/\partial t_0) = 0$ occur when $\mu_2 - \mu_2(\delta) = \gamma_*(\delta, \mu_3)$ or $\mu_2 - \mu_2(\delta) = \gamma^*(\delta, \mu_3)$. By (41) it suffices to consider

$$\begin{aligned} &\frac{\partial d}{\partial t_0}(\delta, \mu_2(\delta) + \gamma(\delta, \mu_3, t_0), \mu_3, t_0) \\ (44) \qquad &= \left[\frac{\partial^2 d}{\partial \mu_2 \partial t_0}(\delta, \mu_2(\delta), 0, t_0)\frac{\partial \gamma}{\partial \mu_3}(\delta, 0, t_0) + \frac{\partial^2 d}{\partial \mu_3 \partial t_0}(\delta, \mu_2(\delta), 0, t_0)\right]\mu_3 + o(\mu_3). \end{aligned}$$

But $(\partial\gamma/\partial\mu_3)(\delta, 0, t_0) = \gamma_1(\delta, t_0)$. We substitute (43) into (44) and rearrange to obtain

(45)

$$\begin{aligned} &\frac{\partial d}{\partial t_0}(\delta, \mu_2(\delta) + \gamma(\delta, \mu_3, t_0), \mu_3, t_0) \\ &\quad = \frac{\partial d}{\partial \mu_2}(\delta, \mu_2(\delta), 0, t_0) \cdot \frac{\partial}{\partial t_0}\left[\frac{\partial d}{\partial \mu_3}(\delta, \mu_2(\delta), 0, t_0) \Big/ \frac{\partial d}{\partial \mu_2}(\delta, \mu_2(\delta), 0, t_0)\right]\mu_3 + o(\mu_3) \\ &\quad = \left[\frac{\partial d}{\partial \mu_2}(0, 0, 0, t_0) + \mathcal{O}(\delta)\right] \cdot [M'(t_0) + \mathcal{O}(\delta)]\mu_3 + o(\mu_3). \end{aligned}$$

For $\mu_3 \neq 0$, (45) is zero if and only if $1/\mu_3 \cdot$ (45) is zero. Since $M'(t_0) = 0$ if and only if $t_0 = t_0^{\min}$ or $t_0 = t_0^{\max}$, and $M'' \neq 0$ at these points, we may use the implicit function theorem to show that for δ, $\mu_2 - \mu_2(\delta)$, and $\mu_3 \neq 0$ small, (45) is zero if and only if $t_0 = t_*(\delta, \mu_3)$ or $t_0 = t^*(\delta, \mu_3)$. This implies the result.

8. Dynamics of the Josephson junction. We shall study the family of differential equations (2) using Theorems 1 and 2. The variable ϕ is taken mod 2π, so that (2) is a family of differential equations on the cylinder. We first set $\varepsilon = 0$ in (2), yielding a family of autonomous equations. If $\rho = 1$, there is a saddle-node at $(\pi/2, 0)$. According to [5], there is a unique $\beta_0 > 0$ such that there is a saddle-node separatrix loop Γ at $\rho = 1$, $\beta = \beta_0$. We let $u = (1, 0)$, $v = (-\beta_0, 1)$; these are eigenvectors at $(\pi/2, 0)$ for the eigenvalues 0, $-1/\beta_0$, and they are tangent to Γ at $(\pi/2, 0)$. If we put $x = (\phi, y)$, $\nu_1 = \rho - 1$, $\nu_2 = \beta - \beta_0$, $\nu_3 = \varepsilon$, then assumptions (i)-(vi) are satisfied at $p = (\pi/2, 0)$ [9].

There is a C^∞ function $\rho = R(\beta, \varepsilon)$, defined near $(\beta_0, 0)$, with $R(\beta, 0) \equiv 1$, such that for $(\rho, \beta, \varepsilon)$ near $(1, \beta_0, 0)$, (2) has a period $2\pi/\omega$ solution of saddle-node type near the constant solution $(\phi, y) \equiv (\pi/2, 0)$ if and only if $\rho = R(\beta, \varepsilon)$. In order to use Theorem 1, we must calculate R to first order at $(\beta_0, 0)$.

LEMMA 3. $(\partial R/\partial\varepsilon)(\beta_0, 0) = 0$.

Let $\rho(\varepsilon) = R(\beta_0, \varepsilon)$. Then

$$(46) \qquad \beta_0\ddot{\phi} + \dot{\phi} + \sin\phi = \rho(\varepsilon) + \varepsilon\sin\omega t$$

has a smooth family of $2\pi/\omega$ periodic solutions $\phi(\varepsilon, t)$, one of whose Floquet multipliers is one. Of course, $\phi(0, t) \equiv \pi/2$. We shall need to calculate ϕ to order ε at $\varepsilon = 0$.

LEMMA 4. $\phi(\varepsilon, t) = \pi/2 - \varepsilon(\cos\omega t + \beta_0\omega\sin\omega t)/\omega(1 + \beta_0^2\omega^2) + \mathcal{O}(\varepsilon^2)$.

Proof of Lemmas 3 *and* 4. Write

$$\rho(\varepsilon) = 1 + \rho_1\varepsilon + \mathcal{O}(\varepsilon^2),$$

$$\phi(\varepsilon, t) = \frac{\pi}{2} + \left(a_0 + \sum_{n=1}^{\infty} a_n\cos n\omega t + \sum_{n=1}^{\infty} b_n\sin n\omega t\right)\varepsilon + \mathcal{O}(\varepsilon^2).$$

Then

$$\sin \phi(\varepsilon, t) = \cos\left[\left(a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t\right)\varepsilon + \mathcal{O}(\varepsilon^2)\right]$$
$$= 1 + \mathcal{O}(\varepsilon^2).$$

Substituting into (46) and collecting terms of order ε, we find that

$$-\beta_0\left(\sum_{n=1}^{\infty} n^2\omega^2 a_n \cos n\omega t + \sum_{n=1}^{\infty} n^2\omega^2 b_n \sin n\omega t\right)$$
$$-\sum_{n=1}^{\infty} n\omega a_n \sin n\omega t + \sum_{n=1}^{\infty} n\omega b_n \cos n\omega t = \rho_1 + \sin \omega t.$$

Therefore, $\rho_1 = 0$, which proves Lemma 3. Also,

$$(47) \qquad a_1 = \frac{-1}{\omega(1+\beta_0^2\omega^2)}, \qquad b_1 = \frac{-\beta_0}{(1+\beta_0^2\omega^2)},$$

and, for $n > 1$, $a_n = b_n = 0$. Thus

$$(48) \qquad \phi(\varepsilon, t) = \frac{\pi}{2} + (a_0 + a_1 \cos \omega t + b_1 \sin \omega t)\varepsilon + \mathcal{O}(\varepsilon^2)$$

where a_1 and b_1 are given by (47) and a_0 is yet to be determined. To complete the proof of Lemma 4, we must show that $a_0 = 0$.

Let $y(\varepsilon, t) = (\partial\phi/\partial t)(\varepsilon, t)$. We linearize (2) about $(\phi(\varepsilon, t), y(\varepsilon, t))$ to obtain the variational equation

$$(49) \qquad \frac{\partial}{\partial t}\begin{bmatrix} \psi(\varepsilon, t) \\ z(\varepsilon, t)\end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -(1/\beta_0)\cos\phi(\varepsilon, t) & -1/\beta_0\end{bmatrix}\begin{bmatrix} \psi(\varepsilon, t) \\ z(\varepsilon, t)\end{bmatrix}.$$

Since $(\phi(\varepsilon, t), y(\varepsilon, t))$ has exactly one Floquet multiplier that is one, there is a smooth family of solutions $(\psi(\varepsilon, t), z(\varepsilon, t))$ of (49) such that $(\psi(\varepsilon, t), z(\varepsilon, t))$ has period $2\pi/\omega$ in t. We may choose

$$(50) \qquad (\psi(0, t), z(0, t)) \equiv (1, 0),$$

since $u = (1, 0)$.

Differentiating (49) with respect to ε we obtain

$$\frac{\partial}{\partial t}\begin{bmatrix} \dfrac{\partial\psi}{\partial\varepsilon} \\[2mm] \dfrac{\partial z}{\partial\varepsilon}\end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\dfrac{1}{\beta_0}\cos\phi(\varepsilon, t) & -\dfrac{1}{\beta_0}\end{bmatrix}\begin{bmatrix} \dfrac{\partial\psi}{\partial\varepsilon} \\[2mm] \dfrac{\partial z}{\partial\varepsilon}\end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \dfrac{1}{\beta_0}\sin\phi(\varepsilon, t)\dfrac{\partial\phi}{\partial\varepsilon} & 0\end{bmatrix}\begin{bmatrix} \psi \\ z\end{bmatrix}.$$

We set $\varepsilon = 0$ and use $\phi(0, t) \equiv \pi/2$ and (50) to obtain

$$\frac{\partial}{\partial t}\begin{bmatrix} \dfrac{\partial\psi}{\partial\varepsilon}(0, t) \\[2mm] \dfrac{\partial z}{\partial\varepsilon}(0, t)\end{bmatrix} = \begin{bmatrix} \dfrac{\partial z}{\partial\varepsilon}(0, t) \\[2mm] -\dfrac{1}{\beta_0}\dfrac{\partial\psi}{\partial\varepsilon}(0, t)\end{bmatrix} + \begin{bmatrix} 0 \\[2mm] \dfrac{1}{\beta_0}\dfrac{\partial\phi}{\partial\varepsilon}(0, t)\end{bmatrix}.$$

This constant coefficient nonhomogeneous linear differential equation has the solution

$$(51) \qquad \frac{\partial\psi}{\partial\varepsilon}(0, t) = \frac{\partial\psi}{\partial\varepsilon}(0, 0) + \int_0^t \frac{\partial z}{\partial\varepsilon}(0, s)\, ds,$$

$$(52) \qquad \frac{\partial z}{\partial\varepsilon}(0, t) = e^{-t/\beta_0}\left[\frac{\partial z}{\partial\varepsilon}(0, 0) + \frac{1}{\beta_0}\int_0^t \frac{\partial\phi}{\partial\varepsilon}(0, s)\, e^{s/\beta_0}\, ds\right].$$

We substitute (48) into (52) and integrate to obtain

$$\begin{aligned}\frac{\partial z}{\partial \varepsilon}(0, t) &= e^{-t/\beta_0}\left[\frac{\partial z}{\partial \varepsilon}(0,0) - a_0 - (a_1 - \omega\beta_0 b_1)/(1+\omega^2\beta_0^2)\right] \\ &\quad + a_0 + [(a_1 - \omega\beta_0 b_1)\cos \omega t + (b_1 + \omega\beta_0 a_1)\sin \omega t]/(1+\omega^2\beta_0^2).\end{aligned} \tag{53}$$

Now we set $t = 2\pi/\omega$ in (53) and require $(\partial z/\partial\varepsilon)(0, 2\pi/\omega) = (\partial z/\partial \varepsilon)(0,0)$. This yields

$$\frac{\partial z}{\partial \varepsilon}(0,0) = a_0 + (a_1 - \omega\beta_0 b_1)/(1+\omega^2\beta_0^2).$$

Therefore,

$$\frac{\partial z}{\partial \varepsilon}(0, t) = a_0 + [(a_1 - \omega\beta_0 b_1)\cos \omega t + (b_1 + \omega\beta_0 a_1)\sin \omega t]/(1+\omega^2\beta_0^2). \tag{54}$$

We substitute (54) into (51) and integrate to obtain

(55)

$$\frac{\partial \psi}{\partial \varepsilon}(0, t) = \frac{\partial \psi}{\partial \varepsilon}(0,0) + a_0 t + [(a_1 - \omega\beta_0 b_1)\sin \omega t - (b_1 + \omega\beta_0 a_1)\cos \omega t]/\omega(1+\omega^2\beta_0^2).$$

Now we set $t = 2\pi/\omega$ in (55) and require $(\partial\varphi/\partial\varepsilon)\ (0, 2\pi/\omega) = (\partial\varphi/\partial\varepsilon)(0,0)$. We find that $a_0 = 0$. This completes the proof of Lemma 4. □

In order to use Theorem 1, we must change to the new parameters

$$\mu_1 = \rho - R(\beta, \varepsilon), \quad \mu_2 = \beta - \beta_{0'} \quad \mu_3 = \varepsilon.$$

Since $R(\beta, 0) \equiv 1$, and $(\partial R/\partial \varepsilon)(\beta_0, 0) = 0$ by Lemma 3, we have

$$\frac{\partial(\mu_1, \mu_2, \mu_3)}{\partial(\rho, \beta, \varepsilon)}(1, \beta_0, 0) = \text{Identity}.$$

Thus instead of calculating derivatives with respect to μ_1, μ_2, μ_3 at $(0,0,0)$, we may calculate derivatives with respect to ρ, β, ε, at $(1, \beta_0, 0)$.

Let

$$f(\phi, y, \rho, \beta) = \left(y, \frac{1}{\beta}(-y - \sin \phi + \rho)\right),$$

$$g(\phi, y, \rho, \beta, \varepsilon, t) = (0, \varepsilon \sin \omega t),$$

$q(t) = (\phi(t), y(t)) =$ a solution of $(\dot\phi, \dot y) = f(\phi, y, 1, \beta_0)$ that lies in Γ,

$$p(\varepsilon, t) = \begin{bmatrix}\phi(\varepsilon, t)\\ y(\varepsilon, t)\end{bmatrix} = \begin{bmatrix}\pi/2 - \varepsilon(\cos \omega t + \beta_0\omega \sin \omega t)/\omega(1+\beta_0^2\omega^2) + \mathcal{O}(\varepsilon^2)\\ \varepsilon(\omega \sin \omega t - \beta_0\omega^2 \cos \omega t)/\omega(1+\beta_0^2\omega^2) + \mathcal{O}(\varepsilon^2)\end{bmatrix}.$$

From [9], we know that $(\partial d/\partial \beta)(1, \beta_0, 0, t_0) > 0$. We now turn to $(\partial d/\partial\varepsilon)$ $(1, \beta_0, 0, t_0)$.

LEMMA 5. $(\partial d/\partial \varepsilon)(1, \beta_0, 0, t) = A(\omega)\cos \omega t + B(\omega) \sin \omega t$, *where*

$$A(\omega) = \lim_{s_1\to\infty}\left[\frac{k}{\omega}\cos \omega s_1 + \int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\sin \omega s\, ds\right],$$

$$B(\omega) = \lim_{s_1\to\infty}\left[-\frac{k}{\omega}\sin \omega s_1 + \int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\cos \omega s\, ds\right].$$

Proof. We have

$$\frac{\partial p}{\partial \varepsilon}(0,t) = -\left(\cos \omega t \begin{bmatrix} 1 \\ \omega^2 \beta_0 \end{bmatrix} + \sin \omega t \begin{bmatrix} \omega\beta_0 \\ -\omega \end{bmatrix}\right) \Big/ \omega(1+\beta_0^2\omega).$$

According to Theorem 1,

$$\lim_{t\to\infty} f(q(t-t_0),1,\beta_0)\exp\left[-\int_{t_0}^{t} \operatorname{div} f(q(s-t_0),1,\beta_0)\,ds\right] = -kv = -k\begin{bmatrix}\beta_0 \\ 1\end{bmatrix}, \tag{56}$$

where k is a positive constant. Therefore, as $t_1 \to \infty$,

$$\frac{\partial p}{\partial \varepsilon}(0,t_1)\wedge f(q(t_1-t_0),1,\beta_0)\exp\left[-\int_{t_0}^{t_1} \operatorname{div} f(q(s-t_0),1,\beta_0)\,ds\right] \to \frac{k}{\omega}\cos \omega t_1 \tag{57}$$

independent of t_0.

To study the second summand of the expression for $(\partial d/\partial\varepsilon)(1,\beta_0,0,t_0)$ given by Theorem 1, the integral, we note that

$$\operatorname{div} f(\phi, y, 1, \beta_0) = -\frac{1}{\beta_0},$$

$$\frac{\partial g}{\partial \varepsilon}(\phi, y, 1, \beta_0, 0, t_0) = \begin{bmatrix} 0 \\ 1/\beta_0 \sin \omega t \end{bmatrix},$$

$$f(q(t-t_0),1,\beta_0)\wedge \frac{\partial g}{\partial\varepsilon}(q(t-t_0),1,\beta_0,0,t) = \begin{bmatrix} \dot\phi(t-t_0) \\ \dot y(t-t_0)\end{bmatrix}$$

$$\wedge \begin{bmatrix} 0 \\ 1/\beta_0 \sin\omega t\end{bmatrix} = \frac{1}{\beta_0}\dot\phi(t-t_0)\sin\omega t = \frac{1}{\beta_0} y(t-t_0)\sin\omega t.$$

Therefore the integral becomes

$$\int_{-\infty}^{t_1} e^{(t-t_0)/\beta_0}\frac{1}{\beta_0} y(t-t_0)\sin\omega t\,dt. \tag{58}$$

Let $s = t - t_0$, $s_1 = t_1 - t_0$ in (58). We obtain

$$\left[\int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\sin\omega s\,ds\right]\cos\omega t_0 + \left[\int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\cos\omega s\,ds\right]\sin\omega t_0. \tag{54}$$

We now combine (57) and (59) and rewrite $\cos\ \omega t_1 = \cos \omega s_1\ \cos\omega t_0 - \sin\omega s_1 \sin\omega t_0$:

$$\frac{\partial d}{\partial\varepsilon}(1,\beta_0,0,t_0) = \lim_{s_1\to\infty}\left\{\left[\frac{k}{\omega}\cos\omega s_1 + \int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\sin\omega s\,ds\right]\cos\omega t_0\right.$$
$$\left.+\left[-\frac{k}{\omega}\sin\omega s_1 + \int_{-\infty}^{s_1} e^{s/\beta_0}\frac{y(s)}{\beta_0}\cos\omega s\,ds\right]\sin\omega t_0\right\},$$

where the limit exists by Theorem 1. Setting $t_0 = 0$ and $t_0 = \pi/2\omega$ we show that each term in square brackets approaches a limit as $s_1 \to \infty$. Let $A(\omega)$ and $B(\omega)$ equal these limits. □

LEMMA 6. *$A(\omega) \neq 0$ for all but a discrete set of ω's.*

Proof. Let

$$h(s) = \begin{cases} 0, & -\infty \leqq s < 0, \\ -k, & 0 \leqq s < \infty. \end{cases}$$

Then

$$\int_{-\infty}^{s_1}\left[e^{s/\beta_0}\frac{y(s)}{\beta_0}+h(s)\right]\cos\omega s\,ds\to B(\omega)\quad\text{as } s_1\to\infty,$$

$$\int_{-\infty}^{s_1}\left[e^{s/\beta_0}\frac{y(s)}{\beta_0}+h(s)\right]\sin\omega s\,ds\to A(\omega)-\frac{k}{\omega}\quad\text{as } s_1\to\infty.$$

We claim that

$$\left[e^{s/\beta_0}\frac{y(s)}{\beta_0}+h(s)\right]e^{s/\beta_0}\to 0\quad\text{as } s\to\infty.$$

From formula (29) in [9], for example, we have that

$$\begin{bmatrix}\phi(t)\\ y(t)\end{bmatrix}=c\begin{bmatrix}-\beta\\ 1\end{bmatrix}e^{-t/\beta_0}+\mathcal{O}(e^{-t/\beta_0})^2$$

for a positive constant c. Therefore,

$$e^{t/\beta_0}\frac{y(t)}{\beta_0}=\frac{c}{\beta_0}+\mathcal{O}(e^{-t/\beta_0}).\tag{60}$$

But (56) can be rewritten as

$$\lim_{t\to\infty}\begin{bmatrix}\dot\phi(t)\\ \dot y(t)\end{bmatrix}e^{t/\beta_0}=-k\begin{bmatrix}-\beta_0\\ 1\end{bmatrix}.\tag{61}$$

Since $\dot\phi(t)=y(t)$, (60) and (61) imply that $c/\beta_0=k$. Therefore (60) implies that as $t\to\infty$,

$$e^{t/\beta_0}\frac{y(t)}{\beta_0}+h(t)=\mathcal{O}(e^{-t/\beta_0}).$$

Therefore,

$$e^{t/\beta_0}\frac{y(t)}{\beta_0}+h(t)\in L^2(-\infty,\infty).$$

On the other hand, the Fourier transform of $e^{t/\beta_0}(y(t)/\beta_0)+h(t)$ is

$$F(\omega)=\int_{-\infty}^{\infty}\left[e^{s/\beta_0}\frac{y(s)}{\beta_0}+h(s)\right]e^{i\omega s}\,ds$$

$$=B(\omega)+i\left(A(\omega)-\frac{k}{\omega}\right).$$

If $A(\omega)\equiv 0$, then $F(\omega)\notin L^2\ (-\infty,\infty)$, which contradicts a well-known fact about the Fourier transform. Since $A(\omega)$ is analytic (because $e^{t/\beta_0}(y(t)/\beta_0)+h(t)$ is also absolutely integrable on $(-\infty,\infty)$), $A(\omega)\neq 0$ except on a discrete set of ω's. □

The idea of relating Melnikov functions to Fourier transforms comes from [4].

From Lemmas 5 and 6 and § 5 of [9] we have immediately the following theorem.

THEOREM 3. *If ω does not belong to the discrete set of Lemma* 6, *then the family* (2) *satisfies the hypotheses of Theorem* 2 *at* $(\phi,y,\rho,\beta,\varepsilon)=(\pi/2,0,1,\beta_0,0)$.

In fact, it only remains to remark that since $M(t_0)$ is a constant multiple of $A(\omega)\cos\omega t_0+B(\omega)\sin\omega t_0$, we have $M(t_0^{\max})>0$ and $M(t_0^{\min})<0$. It follows that for fixed small $\varepsilon\neq 0$, the set of (ρ,β), near $(1,\beta_0)$ for which (2) has a homoclinic orbit is given by Fig. 1. In interpreting Fig. 1, the reader should recall that $\mu_1=\rho-R(\beta,\varepsilon)$ and $\mu_2=\beta-\beta_0$. This result is consistent with [8], where it is shown that at every positive parameter pair (ρ,β) at which the autonomous equation $(\dot\phi,\dot y)=f(\phi,y,\rho,\beta)$ has a saddle separatrix loop, homoclinic orbits occur when one perturbs in the ε-direction.

REFERENCES

[1] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[2] S.-N. CHOW, J. K. HALE AND J. MALLET-PARET, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351-373.

[3] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[4] N. KOPELL AND R. B. WASHBURN, JR., *Chaotic motions in the two-degree-of-freedom swing equations*, IEEE Trans. Circuits and Systems, CAS-29 (1982), pp. 738-746.

[5] M. LEVI, F. C. HOPPENSTEADT AND W. L. MIRANKER, *Dynamics of the Josephson junction*, Quart. Appl. Math., 36 (1978), pp. 167-198.

[6] V. K. MELNIKOV, *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1-57.

[7] S. DE CARVALHO AND R. ROUSSARIE, *Some remarks about homoclinic points of second order differential equations*, in Geometric Dynamics, J. Palis, Jr., ed., Springer-Verlag, New York, 1983, pp. 88-95.

[8] F. M. A. SALAM AND S. S. SASTRY, *Dynamics of the forced Josephson junction circuit: the regions of chaos*, IEEE Trans. Circuits and Systems, CAS-32 (1985), pp. 784-796.

[9] S. SCHECTER, *The saddle-node separatrix-loop bifurcation*, this Journal, 18 (1987), pp. 1142-1157.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

016

ON STOCHASTIC FUNCTIONAL-DIFFERENTIAL EQUATIONS WITH UNBOUNDED DELAY*

CONSTANTIN TUDOR†

Abstract. In this paper we examine the existence and the pathwise uniqueness of solutions to infinite delay stochastic differential equations with general initial conditions and the Lipschitz conditions replaced by some less restrictive ones. Similar questions have been considered by Hale and Kato [3], Kappel and Schappacher [7], Schumacher [13] for deterministic equations and by Rodkina [11] for stochastic equations.

Key words. stochastic equations, strong and weak solutions

AMS(MOS) subject classification. 60H20

1. Introduction. Suppose we are given $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\geqq 0})$ a filtered probability space satisfying the usual assumptions, $Z=\{Z(t)\}_{t\geqq 0}$ a continuous R^m-valued semimartingale with $Z(0)=0$ and $Q=\{Q(t)\}_{t\geqq 0}$ a continuous control process for Z, i.e.,

$$E\left(\sup_{t\leqq\sigma}\left|\int_0^t f(u)\,dZ(u)\right|^2\right)\leqq E\left(Q_\sigma\int_0^\sigma |f(u)|^2\,dQ(u)\right)$$

for every stopping time σ and bounded predictable process f. $\tilde{Y}$ is the linear space of all R^d-valued processes $\{\phi(s)\}_{s\leqq 0}$ with $\mathcal{B}(\phi(s);\ s\leqq 0)\subset\mathcal{F}_0$ (*initial histories*). If $f: R\to R^d$ and $t\geqq 0$ *the history up to* t is the map $f_t: R_-\mapsto R^d$ defined by $f_t(s)=f(t+s)$.

We shall denote by R_∞ the set of all $(x_0, x_1, \cdots)$, $x_i\in R^d$, and by $\mathcal{B}_\infty$ the product Borel field on R_∞. We shall denote by C the set of all continuous functions defined on R_+ and with values in R^d (C is endowed with the compact convergence), by $\mathcal{C}$ the Borel field on C and by $(\mathcal{C}_t)_{t\geqq 0}$ the canonical filtration on C. Also, we shall denote $\bar{\Omega}=\Omega\times C$, $\bar{\mathcal{F}}=\mathcal{F}\otimes\mathcal{C}$, $\bar{\mathcal{F}}_t=\bigcap_{s>t}(\mathcal{F}_s\otimes\mathcal{C}_s)$ and by $\{\bar{x}(t)\}_{t\geqq 0}$ the canonical process defined by $\bar{x}(t,\omega,f)=f(t)$.

The topic of our study is the Cauchy problem (I) for the stochastic functional equation of Doleans-Dade-Protter type

$$\text{(I)}\qquad dx(t)=a(t,x_t)\,dZ(t),\qquad x_0=\phi\in\hat{Y}$$

where $a:\hat{Y}\subset\tilde{Y}\mapsto R^d\otimes R^m$.

Following Jacod and Memin [6] we consider two possible definitions of "a solution" to equation (I).

DEFINITION 1. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, (\tilde{\mathcal{F}}_t))$ be an extension of $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$, i.e.:

(1) $\tilde{\Omega}=\Omega\times\Omega'$, Ω' an auxiliary space;

(2) $\mathcal{F}\subset\tilde{\mathcal{F}}$, $\mathcal{F}_t\subset\tilde{\mathcal{F}}_t$ for every t;

(3) $\tilde{P}/\mathcal{F}=P$.

A process $\tilde{x}=\{\tilde{x}(t)\}_{t\geqq 0}$ is a solution process (or strong solution) of (I) if:

(i_1) Z is a semimartingale on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, (\tilde{\mathcal{F}}_t))$;

(i_2) $\tilde{x}$ is continuous and $\tilde{\mathcal{F}}_t$-adapted;

(i_3) $a(t,\tilde{x}_t)$ is $\tilde{\mathcal{F}}_t$-predictable and Z-integrable (here $\tilde{x}$ is extended onto R_- by $\tilde{x}(s)=\phi(s)$);

(i_4) $\tilde{P}$-a.s. and for all $t\geqq 0$

$$\tilde{x}(t)=\phi(0)+\int_0^t a(s,\tilde{x}_s)\,dZ(s).$$

* Received by the editors August 26, 1985; accepted for publication (in revised form) February 9, 1987.

† Faculty of Mathematics, University of Bucharest, Str. Academiei 14, 70109 Bucharest, Romania.

DEFINITION 2. A probability measure $\bar{P}$ on $\bar{\mathscr{F}}$ is a solution measure (or weak solution) of (I) if $(\bar{\Omega}, \bar{\mathscr{F}}, \bar{P}, (\bar{\mathscr{F}}_t))$ is an extension of $(\Omega, \mathscr{F}, P, (\mathscr{F}_t))$ on which the canonical process $\bar{x}$ is a solution process of (I).

DEFINITION 3. A solution process on $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P}, (\tilde{\mathscr{F}}_t))$ (resp. a solution measure $\bar{P}$) is very good (or regular) if every martingale on $(\Omega, \mathscr{F}, P, (\mathscr{F}_t))$ is also a martingale on $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P}, (\tilde{\mathscr{F}}_t))$ (resp. $(\bar{\Omega}, \bar{\mathscr{F}}, \bar{P}, (\bar{\mathscr{F}}_t))$).

We introduce the following two families of functions:

$$LS = \left\{ \alpha : R_+ \mapsto R_+;\ \alpha \text{ concave, nondecreasing and } \int_{0+}^{1} dt/\alpha(t) = \infty \right\},$$

and for a measurable function $h: R_+ \mapsto R_+$ and $a < b$,

$K(a, b, h) = \{Z(t, x): [a, b] \times R_+ \mapsto R_+;$ Z is measurable in (t, x), $Z(t, \cdot)$ is continuous, concave and nondecreasing such that for each $a < c \leqq b$ $u = 0$ is the unique nonnegative nondecreasing solution of $u(t) \leqq C \int_a^t Z(s, u(h(s)))\, ds$, $a \leqq t \leqq c$, for a constant C large enough$\}$.

In § 2, problem (I) is considered for initial histories which have paths in a general space of functions. The existence of weak (resp. strong) solutions of (I) is established for bounded (this assumption can be relaxed) coefficients $a(t, f)$ that are continuous in f (resp. a is $x|\log x|^{1-\varepsilon}$-Hölder-continuous in f). Also the convergence of successive approximations to the solution is obtained.

In § 3 we consider the equation (I) with $Z(t) = (t, w(t))'$ (w is a standard Wiener process) and with general histories. Existence and uniqueness of strong solutions are proved under less restrictive conditions (the Lipschitz functions are replaced by those of Osgood or Hölder).

Finally a stochastic version of the dangling spider equation is considered.

2. Stochastic functional-differential equations of Doleans–Dade and Protter type with unbounded delay. In this section the initial histories are taken with paths in a semi-normed linear space. More precisely, let X denote a linear real vector space of functions mapping R_- into R^d endowed with a semi-norm $|\cdot|_X$. We require that X satisfies the following general qualitative properties: If $f: R \mapsto R^d$ is continuous on $[\sigma, \infty)$ and $f_\sigma \in X$ then:

(j_1) $f_t \in X$ for every $t \geqq \sigma$;

(j_2) (*The fundamental inequality*): there exist K_1, K_2 locally bounded such that

$$|f_t|_X \leqq K_1(t-\sigma) \max_{\sigma \leqq s \leqq t} |f(s)| + K_2(t-\sigma)|f_\sigma|_X;$$

(j_3) The map $t \mapsto f_t$ is continuous.

Remark 2.1. Such a space X has been considered as phase space in the theory of retarded functional equations (see [1], [3], [7], [12]).

THEOREM 2.1. *Let $\phi \in \tilde{Y}$ be an initial history with the paths in X and $a(t, \omega, f): R_+ \times \Omega \times X \mapsto R^d \otimes R^m$ be a functional such that*

(1) *$a(t, \omega, \cdot)$ is continuous for every t, ω and a is bounded*;

(2) *The process $\{a(t, f_t)\}_t$ is $\mathscr{F}_t$-predictable for every process $\{f(t)\}_{t \in R}$ with $f_0 = \phi$ and $\{f(t)\}_{t \geqq 0}$ continuous and $\mathscr{F}_t$-adapted. Then there exists a regular weak solution of* (I).

THEOREM 2.2. *Assume the hypotheses of Theorem* 2.1 *are satisfied. Moreover suppose that there is $\alpha \in LS$ such that*

$$|a(t, \omega, f) - a(t, \omega, g)|^2 \leqq \gamma(t, \omega)\alpha(|f-g|_X^2)$$

for every $t \in R_+$, $\omega \in \Omega$, $f, g \in X$ and for a Q-integrable and predictable process γ. Then there exists a pathwise unique strong solution of (I).

Proof of Theorems 2.1 *and* 2.2. It follows from the following two lemmas and the corresponding results for the Doleans-Dade and Protter equation (see [6], [8], [10]) for weak solutions and [15] for strong solutions.

LEMMA 1. *Let a be as in Theorem* 2.1. *Define the functional* $a': R_+ \times \Omega \times C \mapsto R^d \otimes R^m$ *by* $a'(t, \omega, f) = a(t, \omega, \bar{f}_t)$, *where*

$$\bar{f}(t) = f(0)\lambda_{(-\infty,0)}(t) + f(t)\lambda_{R_+}(t).$$

Then

(k_1) $a'(t, \omega, \cdot)$ *is continuous with respect to the compact convergence and* a' *is bounded*,

(k_2) a' *is* $\mathscr{F}_t \otimes \mathscr{C}_t$*-predictable.*

Proof. (k_1). It is a consequence of the fundamental inequality and of hypothesis (1) of Theorem 2.1.

(k_2). By the monotone class theorem it is sufficient to consider the case $a'(t, \omega, f) = a_1(t, \omega)a_2(\bar{f}_t)$ with a_1 $\mathscr{F}_t$-predictable and a_2 X-continuous. Since $t \mapsto a_2(\bar{f}_t)$ is continuous it remains to prove that $a_2(\bar{f}_t)$ is $\mathscr{C}_t$-adapted. If $f, g \in C$ are such that $f(s) = g(s)$ for $s \leqq t$ then $\bar{f}_t = \bar{g}_t$, where from $a_2(\bar{f}_t) = a_2(\bar{g}_t)$.

LEMMA 2. *Assume the hypotheses of Theorem* 2.1 *are satisfied. Then, for every coefficient a, the following three assertions are equivalent*:

(k) *For every initial process* ϕ *there exists a strong* (*resp. a regular weak*) *solution of* (I).

(k') *There exists a strong* (*resp. a regular weak*) *solution of* (I) *with* $\phi = 0$.

(k'') *There exists a strong* (*resp. a regular weak*) *solution for the Doleans-Dade and Protter equation*

$$\text{(II)} \qquad dy(t) = a'(t, y)\, dZ(t), \qquad y(0) = 0$$

(*here* a' *is defined as in Lemma* 1).

Proof. The implications $(k) \Rightarrow (k') \Rightarrow (k'')$ are immediate. $(k'') \Rightarrow (k)$. Let y be a strong solution of (II) associated with the functional $a(t, \omega, f + (\bar{\phi})_0)$, where $\bar{\phi}(t) = \phi(t)\lambda_{(-\infty,0)}(t) + \phi(0)\lambda_{R_+}(t)$. It is easy to check that $x(t) = y(t) + \phi(0)$, $t \geqq 0$, is a strong solution of (I) with ϕ as the initial process.

We now consider the case of regular weak solutions. Define the mapping $\Psi: \bar{\Omega} \mapsto \bar{\Omega}$ by $\Psi(\omega, f) = (\omega, f - \phi(0))$. Ψ is bijective, bimeasurable and $\Psi(\bar{\mathscr{F}}_t) = \Psi^{-1}(\bar{\mathscr{F}}_t) = \bar{\mathscr{F}}_t$.

Denote $\hat{a} = a \circ \Psi^{-1}$ and let $\hat{P}$ be a regular solution of

$$y(t) = \int_0^t \hat{a}'(s, y)\, dZ(s)$$

or equivalently of

$$z(t) = \int_0^t \hat{a}(s, z_s)\, dZ(s), \qquad z_0 = 0.$$

Then the probability measure $\bar{P} = \hat{P} \circ \Psi$ is a regular weak solution of (I).

COROLLARY. *Let* $F(s, \omega, x): R_+ \times \Omega \times R_\infty \mapsto R^d \otimes R^m$ *be such that*

(1) *F is* $\mathscr{P} \otimes \mathscr{R}_\infty$*-measurable and bounded* ($\mathscr{P}$ *is the Borel field of predictable sets on* $R_+ \times \Omega$).

(2) *There exists* $\alpha \in LS$ *such that*

$$|F(t, \omega, x) - F(t, \omega, y)|^2 \leqq \gamma(t, \omega)\alpha\left(\sum_{i \geqq 0} c_i|x_i - y_i|^2\right)$$

for every $t \in R_+$, $\omega \in \Omega$, $x, y \in R$ *and for some Q-integrable and predictable process* γ, *where* $c_i \geqq 0$, $\sum_{i \geqq 0} c_i = 1$. *Then the stochastic equation*

$$dx(t) = F(t, \tilde{x}_t)\, dZ(t), \qquad x_0 = \phi$$

has a pathwise unique strong solution for every initial process ϕ *with paths in* X, *where* $\tilde{x}_t = (x(t), x(t-s_1), \cdots)$, $s_i \geqq 0$.

THEOREM 2.3. *Suppose the hypotheses of Theorem* 2.2 *are satisfied and let* x *be the strong solution of* (I). *Define the successive approximations*

$$\begin{aligned} x^0(t) &= \phi(t), \qquad t<0, \\ &= \phi(0), \qquad t \geqq 0, \\ x^{n+1}(t) &= \phi(t), \qquad t<0, \\ &= \phi(0) + \int_0^t a(s, x_s^n)\, dZ(s), \qquad t \geqq 0. \end{aligned}$$

Then x^n *converges to* x *with respect to the compact convergence in probability.*

Proof. It is sufficient to prove the assertion for the Doleans-Dade and Protter equation (II) associated by Lemma 2.

Denote y^n the successive approximations associated to (II). Choose a stopping time θ such that $Q(\theta)+(\gamma \cdot Q)(\theta) \leqq c$. All we need to show is that

$$E\left(\sup_{t \leqq \theta} |y^n(t)-y(t)|^2\right) \mapsto 0.$$

By standard arguments we see that there exists a constant K such as

$$\sup_n E\left(\sup_{t \leqq \theta} |y^n(t)|^2\right) + E\left(\sup_{t \leqq \theta} |y(t)|^2\right) \leqq K.$$

By using the time change theorem with $\theta_s = \inf(t \leqq \theta; (\gamma \cdot Q)(t) > s)$ and the Jensen inequality we get

$$\overline{\lim_{n\to\infty}}\, E\left(\sup_{t \leqq \theta_s} |y^n(t)-y(t)|^2\right) \leqq \int_0^s \alpha\left(\overline{\lim_{n\to\infty}}\, E\left[\sup_{u \leqq \theta_t} |y^n(u)-y(u)|^2\right]\right) dt$$

where from

$$\lim_{n\to\infty} E\left(\sup_{t \leqq \theta_s} |y^n(t)-y(t)|^2\right) = 0 \quad \text{for every } s \leqq c.$$

Taking $s = c$, we obtain the conclusion.

Remark 2.2. The result of Theorem 2.3 for Ito equations and for non-Lipschitz conditions has been proved in [16].

Remark 2.3. If the semi-norm $|\cdot|_X$ satisfies $|f(0)| \leqq c|f|_X$ for all f and for some constant c and the initial process ϕ is such that $E(|\phi|_X^2) < \infty$, then the boundedness hypothesis on a can be replaced by the following: There is a Q-integrable and predictable process γ such that

$$|a(t, \omega, f)|^2 \leqq \gamma(t, \omega)(1+|f|_X^2)$$

for every $t \geqq 0$, $\omega \in \Omega$, $f \in X$.

3. Stochastic functional-differential equations of Ito type with unbounded delay. For every $t>0$ let $(\tilde{C}_t, |\cdot|_t)$ be a Banach space of continuous adapted R^d-valued processes $\{\eta(s)\}_{0 \leqq s \leqq t}$ with $\eta(0)=0$ such that if $\eta \in \tilde{C}_t$ and $s<t$ then $\eta \in \tilde{C}_s$ and $|\eta|_s \leqq |\eta|_t$.

Let w be an R^m-valued Wiener process. For $\psi \in \tilde{Y}$ and a process $\{\eta(t)\}_{t \geqq 0}$ we define the process $\psi \vee \eta$ by

$$(\psi \vee \eta)(s) = \psi(s)\lambda_{(-\infty,0)}(s) + [\psi(0)+\eta(s)]\lambda_{R_+}(s).$$

We consider the operators $I_i(t): \tilde{C}_t \mapsto L^2([0, t] \times \Omega, ds \times dP)$, $i = 1, 2$, defined by

$$(I_1(t)\eta)(s) = \int_0^s \eta(u)\, dw(u); \qquad (I_2(t)\eta)(s) = \int_0^s \eta(u)\, du.$$

We introduce the following assumptions.

ASSUMPTION 1. A subset $\hat{Y}$ of $\tilde{Y}$ is given such that for $\psi \in \hat{Y}$, $\eta \in \tilde{C}_t$ we have $(\psi \vee \eta)_t \in \hat{Y}$.

ASSUMPTION 2. There exist $B_1(t)$, $B_2(t)$ locally bounded such that for every $\eta \in \tilde{C}_t$ and $0 < s < t$ we have

$$|\eta|_t^2 \leqq B_1(t-s)|\eta|_s^2 + B_2(t-s)E\left(\max_{s \leqq u \leqq t} |\eta(u)|^2\right).$$

ASSUMPTION 3. There exist $C_1(t)$, $C_2(t)$ locally bounded such that

$$|I_i(t)\eta|_t^2 \leqq C_i(t) \| |\eta|.\|_{L^2([0,t]\times\Omega)} \qquad i = 1, 2.$$

Remark 3.1. Concrete Banach spaces $\tilde{C}_t$ can be found in [11] and typical examples of spaces $\hat{Y}$ are given by $L^2(\Omega, \mathscr{F}, P, X)$, where X is as shown in the beginning of § 2.

Consider the stochastic functional-differential equation of neutral type:

$$\text{(III)} \qquad d[x(t) - f(t, x_t)] = F(t, x_t)\, dt + G(t, x_t)\, dw(t), \qquad x_0 = \phi \in \hat{Y}$$

where

$$F(t, \psi), f(t, \psi): R_+ x \hat{Y} \mapsto L^2(\Omega, \mathscr{F}, P, R^d),\ G(t, \psi): R_+ \times \hat{Y} \mapsto L^2(\Omega, \mathscr{F}, P, R^d \otimes R^m).$$

THEOREM 3.1. *Suppose Assumptions* 1 *and* 2 *are satisfied. Moreover assume that*

(i) *There exists* $H \in L^1_{\text{loc}}(R_+)$ *such that* $|F(t, \psi)|^2 + |G(t, \psi)|^2 \leqq H(t)$ *for all* $\psi \in \hat{Y}$.

(ii) *For every* $\psi \in \hat{Y}$ *and process* $\eta = \{\eta(t)\}_{t \geqq 0}$ *such that* $\eta \in \tilde{C}_t$ *for every* t, *the processes* $s \to F(s, (\psi \vee \eta)_s)$, $s \to G(s, (\psi \vee \eta)_s)$ *are measurable and adapted.*

(iii) *For every* $\psi \in \hat{Y}$ *and* $t > 0$ *the operators*

$$\eta \to \int_0^t F(s, (\psi \vee \eta)_s)\, ds, \qquad \eta \to \int_0^t G(s, (\psi \vee \eta)_s)\, dw(s): \tilde{C}_t \mapsto L^2(\Omega, \mathscr{F}_t, P, R^d),$$

$$\eta \to \int_0^{\cdot} F(s, (\psi \vee \eta)_s)\, ds, \qquad \eta \to \int_0^{\cdot} G(s, (\psi \vee \eta)_s)\, dw(s): \tilde{C}_t \mapsto \tilde{C}_t$$

are continuous.

(iv) *There exists a measurable function* $h: R_+ \to R_+$ *with* $h(s) \leqq s$ *for all* s *and for every* $t_0 \geqq 0$ *there exist* $J = [t_0, t_0 + \varepsilon]$ *and* $Z \in K(J, h)$ *such that for all* $t \in J$, $\psi \in \hat{Y}$, η, $\eta' \in \tilde{C}_t$

$$\text{(3.1)} \qquad \int_{t_0}^t E[|F(s, (\psi \vee \eta)_s) - F(s, (\psi \vee \eta')_s)|^2]\, ds \leqq \int_{t_0}^t Z(s, |\eta - \eta'|_{h(s)}^2)\, ds,$$

$$\text{(3.2)} \qquad \int_{t_0}^t E[|G(s, (\psi \vee \eta)_s) - G(s, (\psi \vee \eta')_s)|^2]\, ds \leqq \int_{t_0}^t Z(s, |\eta - \eta'|_{h(s)}^2)\, ds.$$

(v) *For every* $\psi \in \hat{Y}$ *and* $t > 0$ *the operator* $\hat{F}: \tilde{C}_t \mapsto \tilde{C}_t$ *defined by*

$$(\hat{F}\eta)(s) = f(s, (\psi \vee \eta)_s) - f(0, \psi)$$

satisfies

$$|\hat{F}\eta - \hat{F}\eta'|_t \leqq k|\eta - \eta'|_t, \qquad k < 1.$$

For $\phi \in \hat{Y}$ define the successive approximations by

$$(y^{n+1})_0 = \phi,$$

$$y^{n+1}(t) = (\hat{F}y^{n+1})(t) + \int_0^t F(s, (\phi \vee y^n)_s)\, ds + \int_0^t G(s, (\phi \vee y^n)_s)\, dw(s),$$

for $t \geqq 0$, where the initial approximation is $y^0 = \phi \vee \tilde{y}$, $\tilde{y} \in \tilde{C}_T$ for all T. Then there exists a continuous process $\{y(t)\}_{t \geqq 0}$ such that

$$|y^n - y|_T \mapsto 0 \quad \textit{for every } T \geqq 0.$$

In particular the process $x_0 = \phi$, $x(t) = y(t) + \phi(0)$, $t \geqq 0$, is a solution of (III), *i.e., x is continuous for $t \geqq 0$ and satisfies*

$$x(t) = f(t, x_t) - f(0, \phi) + \phi(0) + \int_0^t F(s, x_s)\, ds + \int_0^t G(s, x_s)\, dw(s).$$

Proof. Step 1. Assume that $\lim_{m,n\to\infty} |y^m - y^n|_t = 0$ and choose $y = \{y(s)\}_{s \leqq t} \in \tilde{C}_t$ such that $|y^n - y|_t \mapsto 0$.

Denote

$$I_n(t) = \int_0^t [F(s, (\phi \vee y^n)_s) - F(s, (\phi \vee y)_s)]\, ds,$$

$$J_n(t) = \int_0^t [G(s, (\phi \vee y^n)_s) - G(s, (\phi \vee y)_s)]\, dw(s),$$

$$\theta_n(t) = \|I_n(t)\|^2_{L^2(\Omega, \mathcal{F}, P)} + \|J_n(t)\|^2_{L^2(\Omega, \mathcal{F}, P)} + |I_n|^2_t + |J_n|^2_t.$$

By hypothesis (iii) we have $\lim_{m\to\infty} \theta_m(t) = 0$. By Assumption 1, (3.1), (3.2), and Doob and Schwartz's inequalities we get for $T - t$ small enough and for a constant $C(T)$:

$$|y^{m+1} - y^{n+1}|^2_T \leqq C(T)\left[\theta_m(t) + \theta_n(t) + \int_t^T H(s)\, ds\right] \bigg/ (1-k)^2, \tag{3.3}$$

$$|y^{m+1} - y^{n+1}|^2_T \leqq C(T)\left[\theta_m(t) + \theta_n(t) + \int_t^T Z(s, |y^m - y^n|^2_s)\, ds\right] \bigg/ (1-k)^2. \tag{3.4}$$

Step 2. Define $T = \sup\,(t \geqq 0;\ \lim_{m,n\to\infty} |y^m - y^n|_t = 0)$. We have $T \geqq 0$ and we want to prove that $T = \infty$. Assume $T < \infty$. From (3.3) we have that

$$\lim_{m,n\to\infty} |y^m - y^n|_T = 0. \tag{3.5}$$

It is easily seen that

$$\sup_n \sup_{s \leqq t} |y^n|^2_s \leqq C(t) < \infty.$$

Denote $\delta_p(t) = \sup_{m,n \geqq p} |y^m - y^n|^2_t$; $\delta(t) = \lim_{p\to\infty} \delta_p(t)$ and let $J = [T, T + \varepsilon]$ and $Z \in K(J, H)$ be given by hypothesis (iv).

From (3.4) and for $t \in J$ we deduce

$$\delta_{p-1}(t) \leqq C(T+\varepsilon)\left[\sup_{m \geqq p} \theta_m(T) + \sup_{n \geqq p} \theta_n(T) + \int_T^t Z(s, \delta_p(s))\, ds\right] \bigg/ (1-k)^2;$$

hence

$$\delta(t) \leqq C \int_T^t Z(s, \delta(s))\, ds$$

where from $\delta(t)=0$. In particular we obtain $\lim_{m,n\to\infty}|y^m-y^n|_{T+\varepsilon}=0$ which contradicts the definition of T.

THEOREM 3.2. *Suppose Assumptions* 1 *and* 3 *and hypothesis* (v) *of Theorem* 3.1 *are satisfied. Moreover assume that*

(1) *For every* $\psi\in\hat{Y}$ *and* $t>0$ *the operators*

$$F(\cdot,(\psi\vee\cdot).),\ G_i(\cdot,(\psi\vee\cdot).):\tilde{C}_t\mapsto\tilde{C}_t,\ i=1,\cdots,d$$

are continuous.

(2) *There is a measurable function* $h:R_+\mapsto R_+$ *with* $h(s)\leqq s$ *for each* s *and for every* $t_0\geqq 0$ *there exists* $J=[t_0,t_0+\varepsilon]$ *and* $Z\in K(J,h)$ *such that for each* $t\in J$, $\psi\in\hat{Y}$, η, $\eta'\in\tilde{C}_t$

$$|F(\cdot,(\psi\vee\eta).)-F(\cdot,(\psi\vee\eta')).)|_t^2+|G(\cdot,(\psi\vee\eta).)-G(\cdot,(\psi\vee\eta')).)|_t^2\leqq Z(t,|\eta-\eta'|_{h(t)}^2).$$

Then we have the same conclusion as in Theorem 3.1.

Proof. It is easily seen that

$$|y^{n+1}|_s^2\leqq C(t,k,\phi)\left(1+\int_0^s|y^n|_u^2\,du\right)$$

where from by induction we obtain

$$\sup_n\sup_{s\leqq t}|y_s^n|^2\leqq C(t)<\infty. \tag{3.6}$$

Now we fix a positive constant T and assume that $\lim_{m,n\to\infty}|y^m-y^n|_t=0$ for every $t<T$.

Let $\{y(t)\}_{0\leqq t<T}$ be a continuous adapted process such that $|y^n-y|_t\mapsto 0$ for every $t<T$. Then we have

$$\int_0^T|F(\cdot,(\phi\vee y^m).)-F(\cdot,(\phi\vee y^n).)|_t^2\,dt\leqq 2\int_0^T|F(\cdot,(\phi\vee y^m).)$$
$$-F(\cdot,(\phi\vee y).)|_t^2\,dt+2\int_0^T|F(\cdot,(\phi\vee y^n).)-F(\cdot,(\phi\vee y).)|_t^2\,dt\to 0$$

as $m,n\to\infty$ by the continuity of F, (3.6) and the dominated convergence theorem.

An analogue computation holds for G. Therefore if we put

$$\delta_{m,n}(T)=\int_0^T[|F(\cdot,(\phi\vee y^m).)-F(\cdot,(\phi\vee y^n).)|_t^2$$
$$+|G(\cdot,(\phi\vee y^m).)-G(\cdot,(\phi\vee y^n).)|_t^2]\,dt$$

then we have

$$\lim_{m,n\to\infty}\delta_{m,n}(T)=0. \tag{3.7}$$

Let T be as in step 2 of Theorem 2.1. If we assume $T<\infty$ then a little computation yields

$$|y^{m+1}-y^{n+1}|_t^2\leqq C(k,T,\phi)\left[\delta_{m,n}(T)+\int_T^t Z(s,|y^m-y^n|_{h(s)}^2)\,ds\right]$$

Now the proof continues as in Theorem 2.1.

COROLLARY. *Suppose Assumptions* 1 *and* 2 *and hypotheses* (i), (ii) *and* (v) *of Theorem* 3.1 *are satisfied. Moreover assume there exist a measurable function* $h:R_+\mapsto R_+$ *with* $h(s)\leqq s$ *for each* s *and* $Z\in K(R_+,h)$ *such that* (3.1) *and* (3.2) *hold for every* $t\geqq 0$, $\psi\in\hat{Y}$, η, $\eta'\in\tilde{C}_t$. *Then the equation* (III) *has a pathwise unique strong solution.*

Remark 3.2. The results of Theorems 3.1 and 3.2 and of the corollary cover those from [5], [11], [14].

Remark 3.3. An abstract deterministic case closely related to our setting has been considered in [13].

Remark 3.4. Another general approach for Ito integrodifferential equations can be found in [9].

THEOREM 3.3. *For* $j=1,2,\cdots$ *let* $A_j : R_+ \times R^d \mapsto R^d$, $B_j : R_+ \times R^d \mapsto R^d \otimes R^m$, $a_j, b_j : R_+ \mapsto R_+$ *be measurable functions and let* $Z : R_+^2 \mapsto R_+$ *be such that* Z *is measurable,* $Z(t,\cdot)$ *is continuous, concave and nondecreasing. Assume the following*:

(i) *There exists a sequence* $\theta_{-1} = -\theta < 0 = \theta_0 < \theta_1 < \cdots$, $\theta_n \nearrow \infty$ *such as*:

(i_1) *For each* j, $k \geqq 1$, a_j, b_j map $[\theta_{k-1}, \theta_k]$ *into* $[\theta_{r-1}, \theta_r]$ *with* $r \leqq k$;

(i_2) *If* a_j, b_j map $[\theta_{k-1}, \theta_k]$ *into* $[\theta_{k-1}, \theta_k]$, *then* $a_j(t) \leqq t$, $b_j(t) \leqq t$ *for* $\theta_{k-1} \leqq t \leqq \theta_k$,

(i_3) *If we define for* $k \geqq 0$

$$J_1(k) = \{j;\ a_j([\theta_{k-1}, \theta_k]) \subset [\theta_{k-1}, \theta_k]\},$$

$$J_2(k) = \{j;\ b_j([\theta_{k-1}, \theta_k]) \subset [\theta_{k-1}, \theta_k]\},$$

then $u=0$ *is the unique nondecreasing solution of*

$$\begin{aligned} u &= 0, \qquad t \leqq \theta_{k-1}, \\ &\leqq C \int_{\theta_{k-1}}^{t} \left[\sum_{j \in J_1(k)} Z(s, u(a_j(s))) + \sum_{j \in J_2(k)} Z(s, u(b_j(s))) \right] ds, \qquad t \geqq \theta_{k-1}, \end{aligned}$$

for a constant C *large enough.*

(ii) *There is a locally integrable function* γ such that

$$\sum_{j \geqq 1} |A_j(t,x)|^2 + \sum_{j \geqq 1} |B_j(t,x)|^2 \leqq \gamma(t) \quad \textit{for all } t \geqq 0, \quad x \in R^d.$$

(iii) *For every* $j \geqq 1$, $t \geqq 0$, $x, y \in R^d$

$$|A_j(t,x) - A_j(t,y)|^2 + |B_j(t,x) - B_j(t,y)|^2 \leqq Z(t, |x-y|^2).$$

Let $(w_j)_{j \geqq 1}$ *be an infinite number of independent d-dimensional Wiener processes.*

Then the stochastic equation

$$x_0 = \phi \in \hat{Y},$$

$$\text{(IV)} \qquad \begin{aligned} x(t) &= \phi(0) + \sum_{j \geqq 1} \int_0^t A_j(s, x(a_j(s)))\, ds \\ &\quad + \sum_{j \geqq 1} \int_0^t B_j(s, x(b_j(s)))\, dw_j(s), \qquad t \geqq 0 \end{aligned}$$

has a pathwise unique strong solution (*for definition and properties of stochastic integral* $\sum_{j \geqq 1} \int_0^t B_j(\cdot)\, dw_j(\cdot)$; *see* [4]).

Proof. We proceed by induction on the intervals $[\theta_{k-1}, \theta_k]$. Assume that a unique solution exists on $[0, \theta_k]$. We shall prove the existence of a unique solution on $[\theta_k, \theta_{k+1}]$. On the interval $[\theta_k, \theta_{k+1}]$ the equation (IV) becomes

$$x(t) = z(t) + Sx(t)$$

where

$$z(t) = x(\theta_k) + \sum_{j \notin J_1(k)} \int_{\theta_k}^{t} A_j(s, x(a_j(s)))\, ds + \sum_{j \notin J_2(k)} \int_{\theta_k}^{t} B_j(s, x(b_j(s)))\, dw_j(s),$$

$$Sx(t) = \sum_{j \in J_1(k)} \int_{\theta_k}^{t} A_j(s, x(a_j(s)))\, ds + \sum_{j \in J_2(k)} \int_{\theta_k}^{t} B_j(s, x(b_j(s)))\, dw_j(s).$$

Remark that $z(t)$ depends only on the known values of x on $[-\theta, \theta_k]$. Now the conclusion of the theorem is a consequence of Theorem 3.1 (the fact that Sx is an infinite sum of integrals does not change the proof and the validity of Theorem 3.1).

Remark 3.5. Theorem 3.3 represents a stochastic version of a result of Datko [2, Thm. 2.6].

4. An example. We conclude by briefly outlining a stochastic version of the dangling spider equation (see [11], [12] for the deterministic case). The initial histories have the paths in $X = \{f = (f_1, f_2): R_- \mapsto R^2;\ f \text{ continuous}\}$. The semi-norm in X is given by

$$|f|_X = |f_1(0)| + |f_2(0)| + \left[\int_{-\infty}^{0} k(s)|f_1(s)|^p\, ds \right]^{1/p}$$

where $p \geqq 1$, $k \in L_{\text{loc}}(R_-, R_+)$ and $\bar{k}(t) = \operatorname{ess\,sup} k(s+t)/k(s) < \infty$. The stochastic functional equation which we consider is

$$\text{(V)} \qquad \begin{aligned} dv(t) &= [F(t, x(t), v(t)) - \tau(t, x_t)]\, dt + \sigma(t, x(t), v(t))\, dM(t), \\ dx(t) &= v(t)\, dt \end{aligned}$$

where $F: R_+^2 \times R \mapsto R_+$, $\tau: R_+^2 \times X \mapsto R_+$, $\sigma: R_+^2 \times R \mapsto R$ and $\{M(t)\}_{t \geqq 0}$ is a continuous square integrable martingale.

The equation (V) can describe the evolution in time of the length of an extensible, massless, viscoelastic filament which has one end fixed while the other supports a ball (or spider) of unit mass. It is assumed that the ball moves "up and down" with velocity $v(t)$ and $x(t)$ denotes the length of the filament at t. The number $F(t, x(t), v(t))$ is an applied force pulling the ball downward, $\tau(t, x_t)$ is the tension in the filament and $\sigma(t, x(t), v(t))\, dM(t)$ is a random perturbation.

The equation (V) can be written in the form (I) with

$$a(t,f) = \begin{pmatrix} F(t, f_1(0), f_2(0)) - \tau(t, f_1) & \sigma(t, f_1(0), f_2(0)) \\ f_2(0) & 0 \end{pmatrix},$$

$$Z(t) = (t, M(t))'.$$

By using Theorem 2.2 we can state Theorem 4.1:

THEOREM 4.1. *Suppose the following*:

(1) $a(t, 0)$ *is locally bounded.*

(2) *There exist* $\alpha \in LS$ *and* $\gamma(t)$ *locally bounded such that*

$$|F(t, x, y) - F(t, x', y')|^2 + |\sigma(t, x, y) - \sigma(t, x', y')|^2 \leqq \gamma(t)\alpha(|x - x'|^2 + |y - y'|^2),$$

$$|\tau(t, f) - \tau(t, g)|^2 \leqq \gamma(t)\alpha(|f - g|_X^2)$$

for every $t \geqq 0, f, g \in X, x, x', y, y' \in R$. *Then* (V) *has a pathwise unique strong solution for all initial history* ϕ *which is continuous and* $E(|\phi|_X^2) < \infty$.

REFERENCES

[1] B. D. Coleman and D. R. Owen, *On the initial value problem for a class of functional-differential equations*, Arch. Rational Mech. Anal., 55 (1974), pp. 275-299.

[2] R. Datko, *Representation of solutions and stability of linear differential-difference equations in a Banach space*, J. Differential Equations, 29 (1978), pp. 105-166.

[3] J. K. Hale and J. Kato, *Phase space for retarded equations with infinite delay*, Funkcial. Ekvac., 21 (1978), pp. 11-41.

[4] M. Hitsuda and H. Watanabe, *On stochastic integrals with respect to an infinite number of Brownian motions and its applications*, Proc. Internat. Symposium SDE, Kyoto, Japan, 1976, pp. 57-74; John Wiley, New York, 1978.

[5] K. Ito and M. Nisio, *On stationary solutions of stochastic differential equations*, J. Math. Kyoto Univ., 4,1 (1964), pp. 1-79.

[6] J. Jacod and J. Memin, *Weak and strong solutions of stochastic differential equations: Existence and stability*, in Stochastic Integrals, Lecture Notes in Mathematics, 851, Springer-Verlag, New York-Berlin-Heidelberg, 1981, pp. 169-212.

[7] F. Kappel and W. Schappacher, *Some considerations to the fundamental theory of infinite delay equations*, J. Differential Equations, 37 (1980), pp. 141-183.

[8] V. A. Lebedev, *On the existence of weak solutions for stochastic differential equations with driving martingales and random measures*, Stochastics, 9 (1983), pp. 37-76.

[9] B. G. Pachpatte, *On Ito type stochastic integrodifferential equations*, Tamkang J. Math., 10 (1979), pp. 1-18.

[10] J. Pellaumail, *Solutions faibles et semimartingales*, in Sém. Probabilités XV, Lecture Notes in Mathematics 850, Springer-Verlag, New York-Berlin-Heidelberg, 1981, pp. 561-586.

[11] A. E. Rodkina, *On existence and uniqueness of solution of stochastic differential equations with heredity*, Stochastics, 12 (1984), pp. 187-200.

[12] K. Schumacher, *Existence and continuous dependence for functional-differential equations with unbounded delay*, Arch. Rational Mech. Anal., 67 (1978), pp. 315-335.

[13] ———, *Remarks on semilinear partial functional-differential equations with infinite delay*, J. Math. Anal. Appl., 80 (1981), pp. 261-290.

[14] C. Tudor, *Successive approximations for solutions of stochastic integral equations of Volterra type*, J. Math. Anal. Appl., 104 (1984), pp. 27-37.

[15] ———, *Sur les solutions fortes des équations différentielles stochastiques*, C.R. Acad. Sci., Paris, 299 (1984), pp. 117-120.

[16] T. Yamada, *On the successive approximation of solutions of stochastic differential equations*, J. Math. Kyoto Univ., 21 (1981), pp. 501-515.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

017

STABILITY OF SINGULARLY PERTURBED SOLUTIONS TO SYSTEMS OF REACTION-DIFFUSION EQUATIONS*

YASUMASA NISHIURA† AND HIROSHI FUJII†

Abstract. Stability theorem is presented for large amplitude singularly perturbed solutions (SPS) of reaction-diffusion systems on a finite interval. Spectral analysis shows that there exists a unique real critical eigenvalue $\lambda_c(\varepsilon)$ which behaves like $\lambda_c(\varepsilon) \simeq \tau\varepsilon$ as $\varepsilon \downarrow 0$, where ε is a small parameter contained in the system. All the other noncritical eigenvalues have strictly negative real parts independent of ε. The singular limit eigenvalue problem in § 2 plays a key role to judge the sign of τ, which determines the stability of SPS for small ε. Under a natural framework of nonlinearities, τ becomes negative, namely, SPS is asymptotically stable. Instability result is also shown in § 4.

Key words. stability, singularly perturbed solutions, reaction-diffusion equations, asymptotic behaviors

AMS(MOS) subject classifications. 35B25, 35B40, 35K57

Introduction. In this paper, we present a stability theorem of singularly perturbed (and *large amplitude*) stationary solutions with an interior transition layer (SPS1) to systems of reaction-diffusion equations of the form:

$$\text{(P)} \qquad \begin{aligned} &u_t = \varepsilon^2 u_{xx} + f(u, v), \\ &v_t = D v_{xx} + g(u, v), \end{aligned} \qquad (t, x) \in (0, \infty) \times I, \quad I = (0, 1),$$

$$u_x = 0 = v_x, \qquad (t, x) \in (0, \infty) \times \partial I, \quad \partial I = \{0, 1\},$$

where ε is a small parameter and $D > 0$.

The system (P) appears in a number of fields such as eco-systems, morphogenesis in developing biology, chemical reactions, and so on. For some classes of f and g, (P) exhibits as its stationary solutions large amplitude patterns with interior transition layers when one of the diffusion coefficients ε is small. Singular perturbation approaches have been one of the most established methods to construct such spatially inhomogeneous large amplitude patterns. See Fife [4] and the survey by Conway [3]. There have been many works concerning the construction of such singularly perturbed solutions (SPS), e.g., Fife [4], Mimura, Tabata and Hosono [15], Ito [11], Mimura, Nishiura, Tesei and Tsujikawa [13], Fujii and Hosono [5] and so on, in which not only solutions of interior transition layer type but also boundary layer type and mixture of both types are treated. On the other hand, concerning the stability properties of SPS, very few works have been known at least to the authors' best knowledge. (See the survey [3].) The difficulty may lie in the largeness of amplitude of SPS and subtle behaviors of the spectra of the linearized operators as ε tends to zero. The work for the degenerate case $\varepsilon = 0$ of a simple density-dependent diffusion system, by Aronson, Tesei and Weinberger [1], appears to be an exception. However, there is a gap between the degenerate and nondegenerate cases, as will become clear in § 2. Recent works of the authors [17], [19] show the stability of SPS1, i.e., SPS of mode 1, for the nondegenerate case ($\varepsilon > 0$) with large $D > 0$. The basic method there is a perturbation from the limit of $D \uparrow +\infty$. Nevertheless, for a general D, the stability problem of SPS has remained open up to the present time.

* Received by the editors January 29, 1986; accepted for publication October 15, 1986.

† Institute of Computer Sciences, Kyoto Sangyo University, Kyoto 603, Japan.

The goal of this paper is to give a stability theorem to SPS1 for a general D, which provides a framework of global assumptions to be imposed on the nonlinearities to obtain a *stable* SPS1. A violation of our stability assumptions may cause instability. In fact, we shall show an instability theorem at the final section.

A key role in our theory will be played by the *singular limit eigenvalue problem* introduced in § 2, which preserves information on the interior transition layer in the form of *Dirac's δ-function*, even though it is a limit equation of $\varepsilon \downarrow 0$.

In order to treat the limiting case as $D \uparrow \infty$ equally, we set $D = 1/\sigma$. The stationary problem of (P) becomes

$$\text{(SP)} \qquad \varepsilon^2 u_{xx} + f(u, v) = 0, \quad \frac{1}{\sigma} v_{xx} + g(u, v) = 0, \quad u_x = 0 = v_x \quad \text{on } \partial I.$$

When $\sigma \downarrow 0$, (P) is replaced by

$$\text{(P)}_0 \qquad u_t = \varepsilon^2 u_{xx} + f(u, \xi), \quad \xi_t = \int_I g(u, \xi)\, dx, \quad u_x = 0,$$

where $v = \xi$ is a constant function of x and the associated stationary problem is given by the following:

$$\text{(SP)}_0 \qquad \begin{aligned} &\varepsilon^2 u_{xx} + f(u, \xi) = 0, \quad \int_I g(u, \xi)\, dx = 0 \quad \text{in } I, \\ &u_x = 0 \quad \text{on } \partial I. \end{aligned}$$

(SP)$_0$ is called the *shadow system* (see Nishiura [16] for details).

Since (P) is a system of semilinear parabolic equations with diagonal diffusion matrix, the stability of SPS1 is determined by the spectra of the linearized eigenvalue problem:

$$\text{(LP)} \qquad \begin{gathered} \mathcal{L}^{\varepsilon,\sigma}\begin{pmatrix} w \\ z \end{pmatrix} \overset{\text{def}}{=} \begin{bmatrix} \varepsilon^2 \dfrac{d^2}{dx^2} + f_u^{\varepsilon,\sigma} & f_v^{\varepsilon,\sigma} \\ g_u^{\varepsilon,\sigma} & \dfrac{1}{\sigma}\dfrac{d^2}{dx^2} + g_v^{\varepsilon,\sigma} \end{bmatrix} \begin{pmatrix} w \\ z \end{pmatrix} = \lambda \begin{pmatrix} w \\ z \end{pmatrix}, \\ w_x = 0 = z_x \quad \text{on } \partial I, \end{gathered}$$

where all partial derivatives are evaluated at SPS1, $U^{\varepsilon,\sigma}$ (see Theorem 1.1 in § 1), namely, $f_u^{\varepsilon,\sigma} = f_u(u(x; \varepsilon, \sigma), v(x; \varepsilon, \sigma))$ and so on. For the limiting case as $\sigma \downarrow 0$, (LP) is replaced by

$$\text{(LP)}_0 \qquad \begin{gathered} \mathcal{L}^{\varepsilon,0}\begin{pmatrix} w \\ \eta \end{pmatrix} \overset{\text{def}}{=} \begin{bmatrix} \varepsilon^2 \dfrac{d^2}{dx^2} + f_u^{\varepsilon,0} & f_v^{\varepsilon,0} \\ \displaystyle\int_I g_u^{\varepsilon,0} & \displaystyle\int_I g_v^{\varepsilon,0} \end{bmatrix} \begin{pmatrix} w \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} w \\ \eta \end{pmatrix}, \\ w_x = 0 \quad \text{on } \partial I, \end{gathered}$$

where $z = \eta$ is a constant function. Hereafter, (LP) automatically means (LP)$_0$ when $\sigma = 0$. If $\operatorname{Re} \lambda < 0$ for all eigenvalues of (LP), then, $U^{\varepsilon,\sigma}$ is an asymptotically stable solution of (P). It will be convenient to divide the spectra into two classes: one is the class of *critical eigenvalues* which tend to zero as $\varepsilon \downarrow 0$, and the other *noncritical* ones which are bounded away from zero for small ε. We will see in § 2 that noncritical eigenvalues are not dangerous to the stability of SPS1 (Proposition 2.1). Therefore, the stability wholely depends on the asymptotic behavior of critical eigenvalues as $\varepsilon \downarrow 0$. Our conclusion is the following.

MAIN THEOREM. *Under* (A.5), *and* (A.0)-(A.4) *as well, there exists only one critical eigenvalue* $\lambda = \lambda_c(\varepsilon, \sigma)$ *for* $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$ (*see Theorem* 1.1), *and which is real and simple. Furthermore, the principal part of* $\lambda_c(\varepsilon, \sigma)$ *as* $\varepsilon \downarrow 0$ *is given by*

$$\lambda_c(\varepsilon, \sigma) \simeq \tau_N^{*,\sigma} \varepsilon (\tau_N^{*,\sigma} < 0), \tag{0.1}$$

namely, λ_c *approaches to zero from the negative side with* $O(\varepsilon)$ *when* $\varepsilon \downarrow 0$.

Apparently, this theorem implies the asymptotic stability of SPS1 for any small ε. The Main Theorem has been announced in Nishiura and Fujii [18].

Now we state the assumptions for f and g (see Fig. 1).

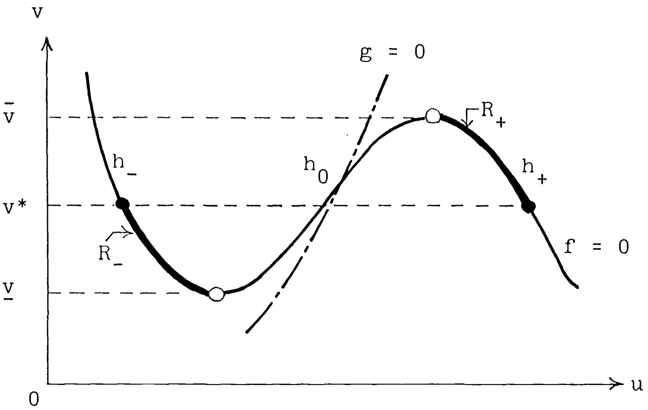


FIG. 1. *Functional forms of f and g.*

Existence Assumptions.

(A.0) f and g are smooth functions of u and v defined on some open set $\mathcal{O}$ in $\mathbb{R}^2$.

(A.1) The nullcline of f is sigmoidal and consists of three continuous curves $u = h_-(v)$, $h_0(v)$ and $h_+(v)$ defined on the intervals I_-, I_0 and I_+, respectively. Let $\min I_- = \underline{v}$ and $\max I_+ = \bar{v}$, then the inequality $h_-(v) < h_0(v) < h_+(v)$ holds for $v \in I^* \overset{\text{def}}{=} (\underline{v}, \bar{v})$, and $h_+(v)(h_-(v))$ coincides with $h_0(v)$ at only one point $v = \bar{v}(\underline{v})$, respectively.

(A.2) $J(v)$ has an isolated zero at $v = v^* \in I^*$ such that $dJ/dv < 0$ at $v = v^*$, where $J(v) = \int_{h_-(v)}^{h_+(v)} f(s, v)\, ds$.

(A.3) $f_u < 0$ on $R_+ \cup R_-$, where $R_+(R_-)$ denotes the part of the curve $u = h_+(v)$ $(h_-(v))$ defined by $R_+(R_-) = \{(u, v) \mid u = h_+(v)\ (h_-(v)) \text{ for } v^* \leqq v < \bar{v}(\underline{v} < v \leqq v^*)\}$ (solid parts of $f = 0$ in Fig. 1), respectively.

(A.4) (a) $g|_{R_-} < 0 < g|_{R_+}$.

(b) $\det(\partial(f, g)/\partial(u, v))|_{R_+ \cup R_-} > 0$.

Stability Assumption.

(A.5) $g_v|_{R_+ \cup R_-} \leqq 0$.

Remark 0.1. Let $G_\pm(v) \overset{\text{def}}{=} g(h_\pm(v), v)$ for $v \in I_\pm$. Then, the assumption (A.4b) is equivalent to

$$\frac{d}{dv} G_\pm(v)|_{R_\pm} < 0, \quad \text{respectively,} \tag{0.2}$$

since it follows from $f(h_\pm(v), v) = 0$ and (A.3) that

$$\frac{d}{dv}G_\pm(v)\bigg|_{R_\pm} = \frac{f_u g_v - f_v g_u}{f_u}\bigg|_{R_\pm}.$$

Note that in [15] the condition (0.2) is assumed in place of (A.4b).

Remark 0.2. It holds that $f_u = 0$ at $(h_+(\bar{v}), \bar{v})$ and $(h_-(\underline{v}), \underline{v})$. Moreover, in general, g and $\det(\partial(f, g)/\partial(u, v))$ in (A.4) may become zero at these end points.

The outline of this paper is the following. In § 1, first, the existence theorem of SPS1 and the asymptotic form of the stretched SPS1 are presented. Second, the spectral behavior of the Sturm-Liouville operator $L^{\varepsilon,\sigma}$ with respect to the parameters (ε, σ) is studied in detail, which is a basic result for later spectral analysis. In § 2, we introduce the singular limit eigenvalue problem (SLEP), which characterizes the asymptotic behavior of the critical eigenvalues of (LP) as $\varepsilon \downarrow 0$. A geometric consideration of the solution of SLEP implies that the critical eigenvalue really behaves as in Main Theorem. It is also proved that all the other noncritical eigenvalues have strictly negative real parts independent of ε. A justification of SLEP is given in § 3, namely, it is shown that the critical eigenvalue is real, unique and simple, and the asymptotic behavior of it as $\varepsilon \downarrow 0$ is just the same as is given by SLEP in § 2. Finally, in § 4, the instability of SPS1 is discussed when some of the assumptions for nonlinearities are violated.

We show several examples which fall into our framework.

Example 1 (*Diffusive prey-predator model* [14]. See Fig. 2).

$$u_t = \varepsilon^2 u_{xx} + f_0(u)u - kuv, \qquad v_t = Dv_{xx} + g_0(v)v + kuv,$$

where k is a positive constant; $f_0(u)$ is a smooth function such that

$$f_0(0) \geqq 0, \quad \frac{d}{du}f_0(u)\begin{cases} >0, & 0 \leqq u < c, \\ =0, & u = c, \\ <0, & u > c, \end{cases}$$

for some positive constant c, and $g_0(v) = c_0 + c_1 v^m$ $(c_0, c_1, m > 0)$. We assume that

$$g_0(v^*) + \left(\frac{d}{dv}g_0(v^*)\right)v^* > kh_+(v^*),$$

which guarantees the stability assumption (A.5), where $u = h_+(v)$ is the right branch of the nullcline $f(u, v) = 0$.

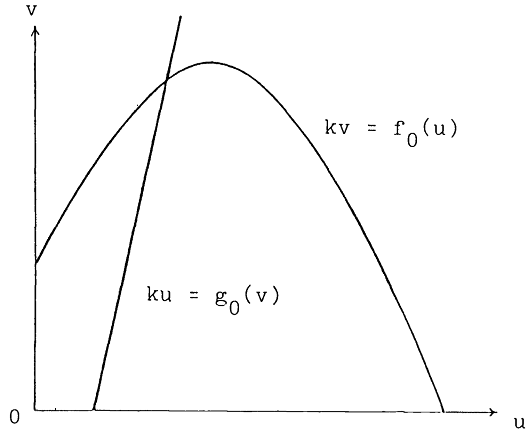


FIG. 2. *Functional forms of prey-predator model.*

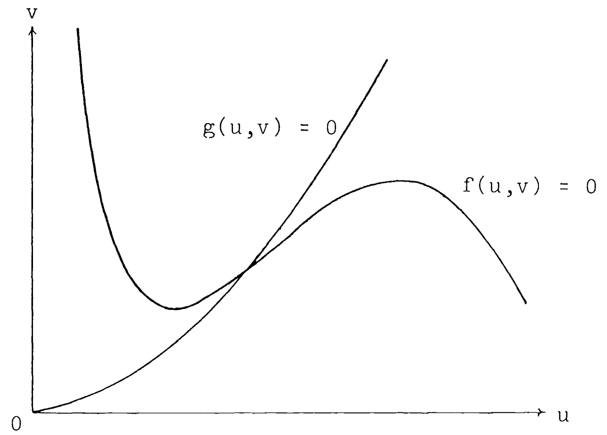


FIG. 3. *Functional forms of Gierer-Meinhardt model with saturation.*

Example 2 (*Gierer-Meinhardt model with saturation* [9]. See Fig. 3).

$$u_t = \varepsilon^2 u_{xx} + \rho\rho_0 + \frac{c\rho u^2}{v(1+\kappa u^2)} - \mu u,$$

$$v_t = Dv_{xx} + c'\rho' u^2 - \nu v,$$

where ρ, ρ', ρ_0, c, c', κ, μ and ν are all positive constants.

Example 3 (*Seelig's model with diffusion* [12]. See Fig. 4).

$$u_t = \varepsilon^2 u_{xx} + j_1 - u - \beta r(u, v), \qquad v_t = Dv_{xx} + j_2 - \gamma r(u, v),$$

where $r(u, v) = uv/(1+u+v+Ku^2)$ and j_1, j_2, β, γ and K are all positive constants.

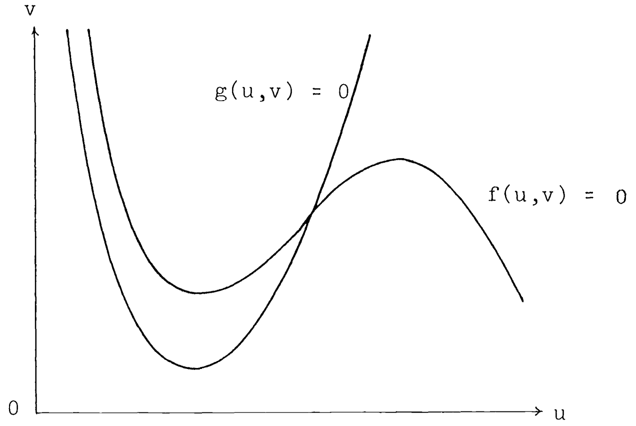


FIG. 4. *Functional forms of Seelig's model.*

We use the following notation and acronym (with page reference) throughout the paper:

$C^p(\bar{I})$ = the space of p-times continuous differentiable functions on $\bar{I}$ with usual norm,

$C^p_\varepsilon(\bar{I})$ = the space of p-times continuous differentiable functions on $\bar{I}$ with the norm

$$\|u\|_{C^p_\varepsilon} = \sum_{k=0}^{p} \max \left| \left(\varepsilon \frac{d}{dx} \right)^k u(x) \right|,$$

$H^p(I)=$ the usual Sobolev space,

$H^p_N(I)=$ the space of closure of $\{\cos(n\pi x)\}_{n=0}^{\infty}$ in $H^p(I)$.

$\langle\cdot,\cdot\rangle=$ the inner product in $L^2(I)$-space.

$C^k_{\text{c.u.}}(\Omega)=$ the compact uniform convergence in C^k-sense in Ω, namely, the uniform convergence on any compact subset of Ω in C^k-sense.

Display designator	*Page number*
(P)	1726
(SP)	1727
$(\mathrm{P})_0$	1727
$(\mathrm{SP})_0$	1727
(LP)	1727
$(\mathrm{LP})_0$	1727
(RP.1)	1731
(RP.2)	1731
(RSP)	1732
(SL)	1736
$(\widetilde{\mathrm{SL}})$	1736
(SL)*	1737
$(\mathrm{SL})_\alpha$	1739
$(\widetilde{\mathrm{SP}})$	1740
$(\mathrm{P2})_{\mathrm{a}}$	1743
$(\mathrm{P2})_{\mathrm{b}}$	1743
(REP)	1743
(AI)	1751
$(\mathrm{SLEP})_{\mathrm{a}}$	1754
$(\mathrm{SLEP\text{-}1})_{\mathrm{b}}$	1755
$(\mathrm{SLEP\text{-}2})_{\mathrm{b}}$	1755

1. Existence theorem and preliminaries. In this section, we show the existence theorem of SPS1 and prepare several lemmas which will be used in subsequent sections.

First, let us begin with the study of reduced solutions and their σ-dependency to understand the form of SPS1. *The reduced problem* is given by putting $\varepsilon=0$ in (SP),

$$\text{(RP.1)} \qquad f(u,v)=0,$$

$$\text{(RP.2)} \qquad \frac{1}{\sigma}v_{xx}+g(u,v)=0,$$

subject to zero flux boundary conditions and $(u,v)\in L^2(I)\times\{H^2(I)\cap H^1_N(I)\}$. The amount of solutions expand to a great extent, when we switch the problem from (SP) to (RP). However, we are interested in the solutions of (RP) which are the limits of those of (SP) as $\varepsilon\downarrow 0$. One of the important such classes is the following. We take

$$\text{(1.1)} \qquad u=h^*(v)=\begin{cases} h_-(v) & \text{for } v\leqq v^*,\ v\in I_-, \\ h_+(v) & \text{for } v\geqq v^*,\ v\in I_+, \end{cases}$$

as a special solution of (RP.1). Substituting this into (RP.2), we have a scalar equation for v, i.e., the *reduced equation,*

$$\text{(RSP)} \qquad \frac{1}{\sigma} v_{xx} + G^*(v) = 0, \qquad v \in H^2(I) \cap H^1_N(I),$$

where $G^*(v) = g(h^*(v), v)$. It follows from (0.2) that $G^*(v)$ is strictly decreasing in each of the subintervals $\underline{v} \leqq v \leqq v^*$ and $v^* \leqq v \leqq \bar{v}$ and that it has a discontinuity of the first kind at $v = v^*$ as in Fig. 5. We only consider monotone increasing C^1-matching solutions of (RSP) at $v = v^*$. Let Num $(v(x))$ denotes the numerical range of a monotone increasing solution $v(x)$ of (RSP), i.e.,

$$\text{Num}\,(v(x)) = [\min_{x \in I} v(x), \max_{x \in I} v(x)].$$

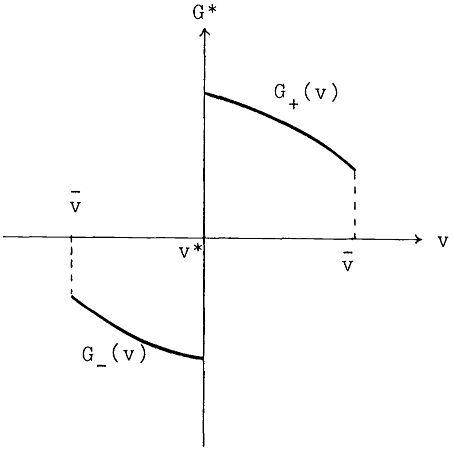


FIG. 5. *Functional form of* $G^*(v)$.

PROPOSITION 1.1. *There exists a uniquely determined positive constant* σ_1^* *such that*

(a) *Monotone increasing* C^1*-matching solutions* $V^{*,\sigma}(x)$ *of* (RSP) *exist for* $0 < \sigma \leqq \sigma_1^*$,

(b) Num $(V^{*,\sigma}(x))$ *is a monotone increasing sequence of intervals of* σ *with respect to inclusion relation and satisfies* $\lim_{\sigma \downarrow 0} \text{Num}\,(V^{*,\sigma}(x)) = \{v^*\}$. *Moreover,* $\lim_{\sigma \downarrow 0} V^{*,\sigma}(x) = v^*$ *in* $C^1(\bar{I})$*-sense*;

(c) Num $(V^{*,\sigma_1^*}(x))$ *contains at least one of the end points* $\underline{v}$ *and* $\bar{v}$, *namely, the reduced solution* (*see* (1.2)) *covers completely at least one of the curves* R_+ *and* R_- *defined in* (A.3).

Proof. See Appendix 1 and Proposition 3.2 in [6] (see also § 3 in [15]).

Using (1.1), we obtain the σ-family of reduced solutions (see Fig. 6),

$$(1.2) \qquad (U^{*,\sigma}(x), V^{*,\sigma}(x)), \qquad 0 \leqq \sigma \leqq \sigma_1^*,$$

where $U^{*,\sigma}(x) = h^*(V^{*,\sigma}(x))$.

Remark 1.1. The value $\sigma = \sigma_1^*$ is an interesting point from a global bifurcation point of view. In fact, it is a taking off point of the D_1 (i.e., one-mode solution)-sheet from the singular wall. Namely, SPS1 constructed by our method ceases to exist for $\sigma > \sigma_1^*$. See [6] and [7] for more detailed discussions.

PROPOSITION 1.2. *The matching point* $x_1^*(\sigma)$ *is well defined by*

$$(1.3) \qquad V^{*,\sigma}(x_1^*(\sigma)) = v^*$$

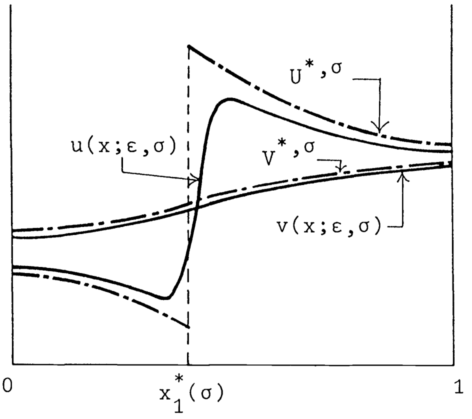


FIG. 6. *Reduced solution and* SPS.

due to the monotonicity of $V^{*,\sigma}(x)$. Then, $x_1^*(\sigma)$ is a continuous function of σ for $0<\sigma\leqq\sigma_1^*$ and uniquely extendable to $\sigma=0$. Here $x_1^*(0)=\lim_{\sigma\downarrow 0}x_1^*(\sigma)$ is determined by

$$\int_I g(U^{*,0}(x),v^*)\,dx=0,$$

where

$$U^{*,0}(x)=\begin{cases}h_-(v^*) & \mathit{for}\, x\in[0,x_1^*(0)),\\ h_+(v^*) & \mathit{for}\, x\in(x_1^*(0),1].\end{cases}\tag{1.4}$$

Proof. See Appendix 1, [15] and [16].

THEOREM 1.1 (*Existence Theorem of* SPS1). *For any σ_0 with $0<\sigma_0<\sigma_1^*$, there is an $\varepsilon_0>0$ such that* (SP) *has an (ε,σ)-family of solutions $U^{\varepsilon,\sigma}=(u(x;\varepsilon,\sigma),v(x;\varepsilon,\sigma))\in C_\varepsilon^2(\bar I)\times C^2(\bar I)$ for $(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}=\{(\varepsilon,\sigma)\,|\,0<\varepsilon<\varepsilon_0,\ 0\leqq\sigma<\sigma_0\}$. $U^{\varepsilon,\sigma}$ are uniformly bounded in $C_\varepsilon^2(\bar I)\times C^2(\bar I)$, and satisfy*

$$\lim_{\varepsilon\downarrow 0}u(x;\varepsilon,\sigma)=U^{*,\sigma}(x)\quad \mathit{uniformly\ on}\ I\backslash I_\kappa\ \mathit{for\ any}\ \kappa>0\tag{1.5}$$

and

$$\lim_{\varepsilon\downarrow 0}v(x;\varepsilon,\sigma)=V^{*,\sigma}(x)\quad \mathit{uniformly\ on}\ I,\tag{1.6}$$

where $I_\kappa=(x_1^*(\sigma)-\kappa,\ x_1^*(\sigma)+\kappa)$, and σ_1^*, $U^{*,\sigma}(x)$ and *$V^{*,\sigma}(x)$ are defined in Proposition* 1.2 *and* (1.2). *See Fig.* 6.

Moreover, when $\sigma\downarrow 0$, $U^{\varepsilon,\sigma}$ converges to the shadow SPS1 *$U^{\varepsilon,0}=(u(x;\varepsilon,0),\ \xi(\varepsilon))$ of* $(\mathrm{SP})_0$ *satisfying*

$$\lim_{\varepsilon\downarrow 0}(x;\varepsilon,0)=U^{*,0}(x),\tag{1.7}$$

$$\lim_{\varepsilon\downarrow 0}\xi(\varepsilon)=v^*,\tag{1.8}$$

where $U^{*,0}(x)$ is defined in Proposition 1.2, *and the convergent manner is the same as in* (1.5) *and* (1.6) *replacing $x_1^*(\sigma)$ by $x_1^*(0)$.*

Finally, $U^{\varepsilon,\sigma}$ depends continuously on $(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}$ in $C_\varepsilon^2\times C^2$-topology, and continuously on $(\varepsilon,\sigma)\in\bar\Omega_{\varepsilon_0,\sigma_0}$ in $L^2\times C^1$-topology, where $\bar\Omega_{\varepsilon_0,\sigma_0}=\{(\varepsilon,\sigma)\,|\,0\leqq\varepsilon<\varepsilon_0,\ 0\leqq\sigma<\sigma_0\}$.

Proof. See Appendix 1. Also, see Mimura, Tabata and Hosono [15], Hosono and Mimura [10] and Ito [11].

In the following, we use the same notation ε_0 for the bound of ε, though we may replace it by a smaller one at need.

The following lemmas concern the asymptotic form of the stretched solution and the spectral behavior of the Sturm-Liouville operator. Here the *stretching* means the change of variables from x to $y=(x-x_1(\varepsilon,\sigma))/\varepsilon$, where $x_1(\varepsilon,\sigma)$ is the C^1-matching point of SPS1 with $\lim_{\varepsilon\downarrow 0} x_1(\varepsilon,\sigma)=x_1^*(\sigma)$ and $u(x_1(\varepsilon,\sigma);\varepsilon,\sigma)=h_0(v^*)$ (see (17) in Appendix 1). For a given function $w(x)$ on I, we denote the *stretched function* by $\tilde{w}(y)$, i.e.,

$$\tilde{w}(y)=w(x_1(\varepsilon,\sigma)+\varepsilon y). \tag{1.9}$$

We also use the following notation for the stretched intervals:

$$\tilde{I}=(-l/\varepsilon, r/\varepsilon),\quad \tilde{I}_-=(-l/\varepsilon,0)\quad\text{and}\quad \tilde{I}_+=(0,r/\varepsilon), \tag{1.10}$$

where $l=x_1(\varepsilon,\sigma)$ and $r=1-x_1(\varepsilon,\sigma)$.

The first lemma is on the convergence of the stretched SPS1 to the static front solution as $\varepsilon\downarrow 0$.

LEMMA 1.1. *Let $\tilde{U}^{\varepsilon,\sigma}=(\tilde{u}(y;\varepsilon,\sigma),\tilde{v}(y;\varepsilon,\sigma))$ be the stretched solution of* SPS1 $U^{\varepsilon,\sigma}$, *and let $\tilde{u}^*$ be a unique monotone increasing solution on $\mathbb{R}$ of the following problem*

$$\frac{d^2}{dy^2}u+f(u,v^*)=0, \tag{1.11a}$$

$$u(\pm\infty)=h_\pm(v^*), \tag{1.11b}$$

$$u(0)=h_0(v^*). \tag{1.11c}$$

Then we have

$$\lim_{\varepsilon\downarrow 0}\tilde{U}^{\varepsilon,\sigma}=\tilde{U}^*\quad in\ C^2_{\text{c.u.}}(\mathbb{R})\text{-}sense, \tag{1.12}$$

where $\tilde{U}^*$ is defined by $\tilde{U}^*=(\tilde{u}^*,v^*)$.

Recall that $C^2_{\text{c.u.}}(\mathbb{R})$ is the abbreviation of *compact uniform convergence in C^2-sense on* $\mathbb{R}$. Here, we note that the limiting function $\tilde{U}^*$ is independent of σ. The convergence (1.12) is uniform with respect to σ for $0\leqq\sigma<\sigma_0$.

Remark 1.2. Though $\tilde{U}^{\varepsilon,\sigma}$ is not defined for all $x\in\mathbb{R}$, (1.12) makes sense, because for any fixed compact interval K, there exists an ε_K such that $\tilde{U}^{\varepsilon,\sigma}$ is defined on K for $0<\varepsilon<\varepsilon_K$.

Proof of Lemma 1.1. We can assume without loss of generality that $u(x;\varepsilon,\sigma)$ takes the value $h_0(v^*)$ at the matching point $x=x_1(\varepsilon,\sigma)$. We prove this lemma according to the construction of SPS1 in Appendix 1.

We divide the solution $U^{\varepsilon,\sigma}$ into two parts at $x=x_1(\varepsilon,\sigma)$, i.e.,

$$U^{\varepsilon,\sigma}=\begin{cases}U_-^{\varepsilon,\sigma} & \text{for } x\in\bar{I}_-=[0,x_1(\varepsilon,\sigma)],\\ U_+^{\varepsilon,\sigma} & \text{for } x\in\bar{I}_+=[x_1(\varepsilon,\sigma),1].\end{cases}$$

We shall show that $\tilde{U}_+^{\varepsilon,\sigma}(\tilde{U}_-^{\varepsilon,\sigma})$ converges to the right(left)-half of $\tilde{U}^*$ in $C^2_{\text{c.u.}}$-sense on $\mathbb{R}^+(\mathbb{R}^-)$ as $\varepsilon\downarrow 0$, where $\mathbb{R}^+=[0,+\infty)$ ($\mathbb{R}^-=(-\infty,0]$). First, note the following two remarks: For any interval $I_\#\subseteqq I$; (i) Let $w(x;\varepsilon)$ be uniformly bounded in $C^2_\varepsilon(I_\#)$-sense for small $\varepsilon(I_\#\subseteqq I)$, then after stretching, $\tilde{w}(y;\varepsilon)$ is uniformly bounded in $C^2(\tilde{I}_\#)$; (ii) Let $w(x;\varepsilon)$ be uniformly bounded in $C^2(I_\#)$-sense for small ε, and $w(x_1(\varepsilon,\sigma);\varepsilon)$ converges to w^* as $\varepsilon\downarrow 0$, then $\tilde{w}(y;\varepsilon)$ converges to w^* in $C^2_{\text{c.u.}}(\tilde{I}_\#)$-sense as $\varepsilon\downarrow 0$.

In view of (8) and (10) in Appendix 1, $U_{\pm}^{\varepsilon,\sigma} = (u_{\pm}(x;\, \varepsilon, \sigma), v_{\pm}(x;\, \varepsilon, \sigma))$ has the form

$$\begin{aligned} u_{\pm}(x;\, \varepsilon, \sigma) &= h_{\pm}(\hat{V}_{\pm}(x;\, \delta, \omega, \sigma)) + z_{\pm}(x;\, \delta, \omega, \varepsilon, \sigma) + r_{\pm}(\varepsilon, \sigma) \\ &\quad + \sigma \frac{d}{dv} h_{\pm}(\hat{V}_{\pm}(x;\, \delta, \omega, \sigma)) \cdot s_{\pm}(\varepsilon, \sigma), \end{aligned} \tag{1.13}$$

$$v_{\pm}(x;\, \varepsilon, \sigma) = v^* + \omega + \sigma\{W_{\pm}(x;\, \delta, \omega, \sigma) + \varepsilon^2 Y_{\pm} + s_{\pm}(\varepsilon, \sigma)\}. \tag{1.14}$$

Since $v_{\pm}(x;\, \varepsilon, \sigma)$ remains uniformly bounded in $C^2(I_{\pm})$ for small ε, it follows from the above remark, Lemma A.2, Remark A.2, Theorem A.1, and (16) in Appendix 1 that

$$\lim_{\varepsilon \downarrow 0} \tilde{v}_{\pm}(y;\, \varepsilon, \sigma) = v^* \quad \text{in } C^2_{\text{c.u.}}(\mathbb{R}^{\pm})\text{-sense.} \tag{1.15}$$

For the u-component, first note that $\lim_{\varepsilon \downarrow 0} \hat{V}_{\pm} = v^*$ in $C^2_{\text{c.u.}}(\mathbb{R}^{\pm})$-sense (see (7) in Appendix 1). Therefore, we have

$$\lim_{\varepsilon \downarrow 0} h_{\pm}(\hat{V}_{\pm}) = h_{\pm}(v^*) \quad \text{in } C^2_{\text{c.u.}}(\mathbb{R}^{\pm})\text{-sense.} \tag{1.16}$$

On the other hand, it follows from the definition of $z_{\pm}$ (see § A.2 in Appendix 1) and the continuous dependence of solutions on initial data and parameters on any fixed compact interval that

$$\lim_{\varepsilon \downarrow 0} \tilde{z}_{\pm} = \tilde{z}_{\pm}^* \quad \text{in } C^2_{\text{c.u.}}\ (\mathbb{R}^{\pm})\text{-sense,} \tag{1.17}$$

where $\tilde{z}_{\pm}^*$ is the unique solution of

$$\begin{gathered} \frac{d^2}{dy^2} \tilde{z}_{\pm}^* + f(h_{\pm}(v^*) + \tilde{z}_{\pm}^*, v^*) = 0, \\ \tilde{z}_{\pm}^*(0) = h_0(v^*) - h_{\pm}(v^*), \\ \tilde{z}_{+}^*(+\infty) = 0 \quad \text{or} \quad \tilde{z}_{-}^*(-\infty) = 0. \end{gathered} \tag{1.18}$$

Using the above results and Theorem A.1 in Appendix 1 and recalling that $\tilde{u}_{-}(y;\, \varepsilon, \sigma)$ and $\tilde{u}_{+}(y;\, \varepsilon, \sigma)$ are matched in C^2-sense at $y = 0$, we can see that

$$\lim_{\varepsilon \downarrow 0} \tilde{u}(y;\, \varepsilon, \sigma) = \tilde{u}^*(y) \quad \text{in } C^2_{\text{c.u.}}(\mathbb{R})\text{-sense.} \tag{1.19}$$

Since all the above convergence results have a uniformity with respect to σ, the convergence (1.12) is uniform for $0 \leqq \sigma < \sigma_0$, which completes the proof of Lemma 1.1.

Remark 1.3. Differentiating (1.11a) with respect to y, we see that $(d/dy)\tilde{u}^*(y)$ is a constant multiple of the positive normalized principal eigenfunction ϕ_0^* of the Sturm–Liouville operator $(d^2/dy^2) + f_u(\tilde{u}^*(y), v^*)$ on $\mathbb{R}$. Therefore, it follows from Lemma 1.1 that

$$\frac{d}{dy} \tilde{u}(y;\, \varepsilon, \sigma) \xrightarrow[\varepsilon \downarrow 0]{} \kappa^{*^{-1}} \hat{\phi}_0^* \left(= \frac{d}{dy} \tilde{u}^*(y) \right) \tag{1.20}$$

compact uniformly in C^2-sense, where $\kappa^{*^{-1}} = \|(d/dy)\tilde{u}^*(y)\|_{L^2(\mathbb{R})}$.

COROLLARY 1.1. *Let $F(u, v)$ be a smooth function of u and v. Then, the composite function $F(\tilde{u}(y;\, \varepsilon, \sigma), \tilde{v}(y;\, \varepsilon, \sigma))$ converges to $F(\tilde{u}^*, v^*)$ compact uniformly in C^2-sense.*

In view of the construction of $U^{\varepsilon,\sigma}$ and the properties of $z_{\pm}$, we have the following.

COROLLARY 1.2. *For any $\gamma > 0$, there exists $M > 0$ such that*

$$|\tilde{U}^{\varepsilon,\sigma}(y;\, \varepsilon, \sigma) - \tilde{U}^{*,\sigma}| < \gamma \quad \textit{for } |y| > M \textit{ and all small } \varepsilon,$$

where $\tilde{U}^{*,\sigma}$ is the stretched function of the reduced solution, i.e., $\tilde{U}^{*,\sigma} = U^{*,\sigma}(x_1(\varepsilon, \sigma) + \varepsilon y)$.

Let us introduce the following Sturm-Liouville problem at $U^{\varepsilon,\sigma}$,

$$\text{(SL)}\qquad \begin{aligned} &L^{\varepsilon,\sigma}\phi \overset{\text{def}}{=} \left(\varepsilon^2\frac{d^2}{dx^2}+f_u^{\varepsilon,\sigma}\right)\phi=\zeta\phi \quad \text{in } I,\\ &\phi_x=0 \quad \text{on } \partial I,\end{aligned}$$

where $f_u^{\varepsilon,\sigma}=f_u(u(x;\,\varepsilon,\sigma),\ v(x;\,\varepsilon,\sigma))$ (see Fig. 7(a)). Let $\{\phi_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ be the complete orthonormal set (CONS), and $\{\zeta_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ the associated eigenvalues of (SL), which are all real and simple. It is convenient to define the stretched Sturm-Liouville problem for (SL):

$$(\widetilde{\text{SL}})\qquad \begin{aligned} &\tilde{L}^{\varepsilon,\sigma}\hat{\phi} \overset{\text{def}}{=} \left(\frac{d^2}{dy^2}+\tilde{f}_u^{\varepsilon,\sigma}\right)\hat{\phi}=\zeta\hat{\phi} \quad \text{in } \tilde{I},\\ &\hat{\phi}_y=0 \quad \text{on } \partial\tilde{I},\end{aligned}$$

where $\tilde{f}_u^{\varepsilon,\sigma}$ is the stretched potential of $f_u^{\varepsilon,\sigma}$ (see Fig. 7(b)). Similarly as above, let $\{\hat{\phi}_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ be the CONS and $\{\zeta_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ the associated real simple eigenvalues of $(\widetilde{\text{SL}})$. Note that the set of eigenvalues $\{\zeta_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ remain the same after stretching. On the other hand, we need $\sqrt{\varepsilon}$-factor for the eigenfunctions

$$(1.21)\qquad \hat{\phi}_n^{\varepsilon,\sigma}=\sqrt{\varepsilon}\,\tilde{\phi}_n^{\varepsilon,\sigma},$$

where $\tilde{\phi}_n^{\varepsilon,\sigma}$ is the stretched function of $\phi_n^{\varepsilon,\sigma}$, i.e.,

$$\tilde{\phi}_n^{\varepsilon,\sigma}(y)=\phi_n^{\varepsilon,\sigma}(x_1(\varepsilon,\sigma)+\varepsilon y).$$

The relation (1.21) comes from the normalization

$$\int_I(\phi_n^{\varepsilon,\sigma})^2\,dx=1=\int_{\tilde{I}}(\sqrt{\varepsilon}\,\tilde{\phi}_n^{\varepsilon,\sigma})^2\,dy.$$

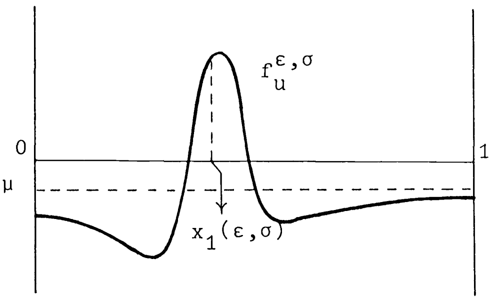


FIG. 7(a). *Functional form of the potential* $f_u^{\varepsilon,\sigma}$.

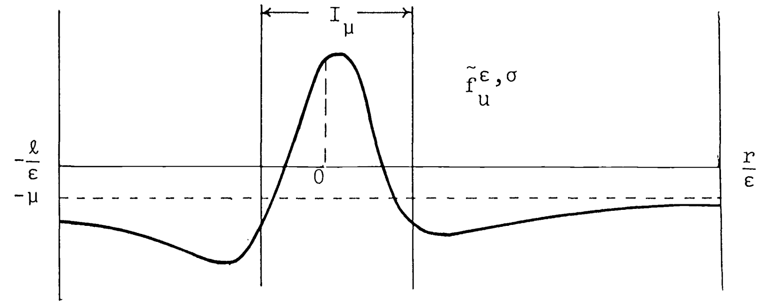


FIG. 7(b). *The stretched potential* $\tilde{f}_u^{\varepsilon,\sigma}$.

Motivated by

$$\tilde{f}_u^{\varepsilon,\sigma} \underset{\varepsilon\downarrow 0}{\longrightarrow} \tilde{f}_u^* \quad \text{in } C^2_{\text{c.u.}}\text{-sense,}$$

which follows from Corollary 1.1, where $\tilde{f}_u^* = f_u(\tilde{u}^*, v^*)$, we introduce a Sturm-Liouville problem on $\mathbb{R}$:

$$\text{(SL)}^* \qquad \tilde{L}^*\hat{\phi} = \left(\frac{d^2}{dy^2} + \tilde{f}_u^*\right)\hat{\phi} = \zeta\hat{\phi} \quad \text{in } \mathbb{R},$$

which is a natural limiting problem of $(\widetilde{\text{SL}})$ as $\varepsilon \downarrow 0$. In view of (A.3) and Lemma A.3 in Appendix 1, we see that the potential $\tilde{f}_u^*$ approaches to negative constants $f_u(h_\pm(v^*), v^*)$ as $y \to \pm\infty$ in the following way (see Fig. 7(c));

$$(1.22) \qquad |f_u(\tilde{u}^*(y), v^*) - f_u(h_\pm(v^*), v^*)| \leqq C \exp(-\gamma|y|), \qquad y \to \pm\infty,$$

where C and γ are positive constants.

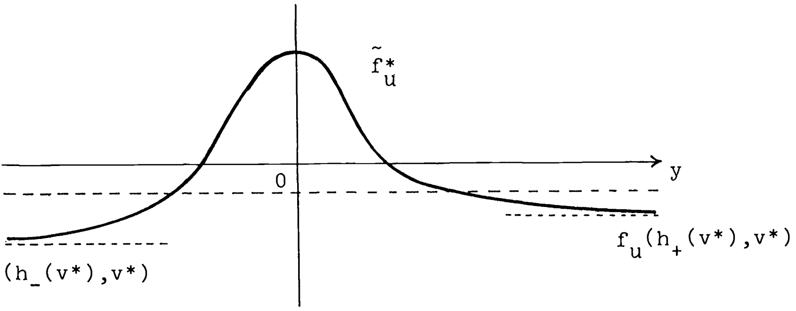


FIG. 7(c). *Limiting form $\tilde{f}_u^*$ of the stretched potential.*

The following observation for $(\text{SL})^*$ is useful.

LEMMA 1.2. *The value 0 is the principal eigenvalue of* $(\text{SL})^*$, *which is simple in* $L^2(\mathbb{R})$, *and the associated normalized positive eigenfunction $\hat{\phi}_0^*$ is given by*

$$\hat{\phi}_0^* = \kappa^* \frac{d}{dy}\tilde{u}^*$$

with

$$\kappa^* = \left\| \frac{d}{dy}\tilde{u}^* \right\|^{-1}_{L^2(\mathbb{R})}.$$

Proof. Noting that the potential $\tilde{f}_u^*$ of $(\text{SL})^*$ is of well-type, which satisfies (1.22), this lemma is a direct consequence of Remark 1.3 and the fact that $(\text{SL})^*$ is of limit point type (see Coddington and Levinson [2, Chap. 9]).

The principal eigenfunction $\hat{\phi}_0^{\varepsilon,\sigma}$ of $(\widetilde{\text{SL}})$ converges to that of $(\text{SL})^*$ in the following sense.

LEMMA 1.3. *It follows that*

$$\hat{\phi}_0^{\varepsilon,\sigma}(=\sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}) \underset{\varepsilon\downarrow 0}{\longrightarrow} \hat{\phi}_0^* \quad \textit{in } C^2_{\text{c.u.}}(\mathbb{R})\textit{-sense,}$$

$\hat{\phi}_0^{\varepsilon,\sigma}(\hat{\phi}_0^*)$ *is the normalized positive principal eigenfunction of* $(\widetilde{\text{SL}})$ $((\text{SL})^*)$, *respectively. This convergence is uniform for* $0 \leqq \sigma < \sigma_0$.

Proof. First, since $\tilde{u}^*$ approaches to $h_+(v^*)$ (or, $h_-(v^*)$) with an exponential order as $y \to +\infty$ (or, $-\infty$) (Lemma A.3 in Appendix 1), $\tilde{f}_u^{\varepsilon,\sigma}$ converges to $\tilde{f}_u^* = f_u(\tilde{u}^*, v^*)$ in $C^2_{\text{c.u.}}$-sense from Corollary 1.1. Second, from Corollary 1.2 and (A.3),

$$(1.23) \qquad \tilde{f}_u^{\varepsilon,\sigma} < -\mu < 0$$

holds uniformly for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0,\sigma_0}$ outside of an appropriate fixed finite interval I_μ, μ being independent of ε and σ. Therefore, it is easy to verify that $\{\hat{\phi}_0^{\varepsilon,\sigma}\}_{0<\varepsilon<\varepsilon_0}(\|\hat{\phi}_0^{\varepsilon,\sigma}\|_{L^2(\tilde{I})} = 1)$ is uniformly bounded, with the decaying estimate

$$\hat{\phi}_0^{\varepsilon,\sigma}(y) \leqq \hat{C} \exp(-\hat{\gamma}|y|) \quad \text{for } y \in \tilde{I} \backslash I_\mu, \tag{1.24}$$

where $\hat{C}$ and $\hat{\gamma}$ are positive constants which do not depend on (ε, σ). Let K be an arbitrary compact interval. Then, using the above fact, and that $\hat{\phi}_0^{\varepsilon,\sigma}$ is a solution of $(\widetilde{\mathrm{SL}})$, we can see that $\{\hat{\phi}_0^{\varepsilon,\sigma}\}_{0<\varepsilon<\varepsilon_0}$ forms an Ascoli-Arzela set on K in C^2-sense (i.e., precompact set in C^2-topology on K). Let us choose a subsequence $\{\hat{\phi}_0^{\varepsilon_n,\sigma}\}_{n\geqq 1}$ which constitutes a Cauchy sequence on K in C^2-sense. Using a diagonal argument on an expanding sequence of compact intervals which eventually covers the whole line $\mathbb{R}$, we can select a subsequence $\{\hat{\phi}_0^{\varepsilon_n,\sigma}\}_{n\geqq 1}$ (we use the same symbol as before) which converges uniformly on any compact interval. Namely, it forms a convergent sequence on $\mathbb{R}$ in $C^2_{\text{c.u.}}$-topology. Let us denote this limit by Φ^*. Let $\{\zeta_0^{\varepsilon_n,\sigma}\}$ be the set of corresponding eigenvalues. We can then assume without loss of generality that $\zeta_0^{\varepsilon_n,\sigma} \to \zeta^*$ as $n \to +\infty$, since it must be in the bounded interval

$$\left[\min_{y\in\bar{I},(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}} f_u^{\varepsilon,\sigma}, \max_{\text{the same region}} f_u^{\varepsilon,\sigma}\right].$$

Consequently, Φ^* satisfies

$$\begin{gathered} \frac{d^2}{dy^2}\Phi^* + \tilde{f}_u^*\Phi^* = \zeta^*\Phi^*, \\ \Phi^* > 0, \qquad \|\Phi^*\|_{L^2(\mathbb{R})} = 1. \end{gathered} \tag{1.25}$$

The strict positivity comes from $\hat{\phi}_0^{\varepsilon,\sigma} > 0$ and that Φ^* satisfies the above equation. On the other hand, 0 is the principal eigenvalue of (SL)* from Lemma 1.2, and the associated normalized positive eigenfunction is given by $\hat{\phi}_0^*(= \kappa^*(d/dy)\tilde{u}^*)$. By virtue of the uniqueness of the normalized positive principal eigenfunction and the orthogonal property of eigenfunctions, we can conclude that $\Phi^* = \hat{\phi}_0^*$ and $\zeta^* = 0$. Thus, the limit Φ^* does not depend on the choice of subsequences. Hence, $\hat{\phi}_0^{\varepsilon,\sigma}$ itself converges to $\hat{\phi}_0^*$ in $C^2_{\text{c.u.}}$-sense. The uniformity with respect to σ is easily shown by contradiction with the aid of the uniqueness of $\hat{\phi}_0^*$.

COROLLARY 1.3.

$$\int_I \phi_0^{\varepsilon,\sigma}\, dx = L(\varepsilon, \sigma)\sqrt{\varepsilon},$$

where $L(\varepsilon, \sigma)$ *is a positive continuous function of* $(\varepsilon, \sigma) \in \bar{\Omega}_{\varepsilon_0\sigma_0}$. *Moreover,*

$$L^* \overset{\text{def}}{=} L(0, \sigma) = \kappa^*(h_+(v^*) - h_-(v^*)) > 0. \tag{1.26}$$

Note that the limit L^* *is independent of* σ.

Proof. Using a stretched variable $y = (x - x_1(\varepsilon, \sigma))/\varepsilon$, the integral is rewritten as

$$\int_I \phi_0^{\varepsilon,\sigma}\, dx = \sqrt{\varepsilon} \int_{\tilde{I}} \sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}\, dy.$$

Define $L(\varepsilon, \sigma)$ by

$$L(\varepsilon, \sigma) = \int_{\tilde{I}} \sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}\, dy. \tag{1.27}$$

It is clear from the smooth parametric dependency of SPS1 $U^{\varepsilon,\sigma}$ in $\Omega_{\varepsilon_0,\sigma_0}$ that $L(\varepsilon,\sigma)$ is a continuous function in $\Omega_{\varepsilon_0,\sigma_0}$. It follows from Lemmas 1.2 and 1.3 and (1.24) that

$$\begin{aligned}\lim_{\varepsilon\downarrow 0} L(\varepsilon,\sigma) &= \int_{\mathbb{R}} \hat{\phi}_0^* \, dy \\ &= \kappa^* \int_{\mathbb{R}} \frac{d}{dy}\tilde{u}^* \, dy \\ &= \kappa^*(h_+(v^*) - h_-(v^*)) \quad \text{uniformly for } 0 \leqq \sigma < \sigma_0.\end{aligned}$$

Note that the right-hand side does not depend on σ. Consequently, we see that $L(\varepsilon,\sigma)$ is uniquely extended to be a continuous function on $\bar{\Omega}_{\varepsilon_0,\sigma_0}$ with the limit condition (1.26).

Remark 1.4. Let $\phi_{0,\alpha}^{\varepsilon,\sigma}$ be the normalized positive principal eigenfunction of

$$(\mathrm{SL})_\alpha \qquad \begin{aligned} &L^{\varepsilon,\sigma} = \left(\varepsilon^2 \frac{d^2}{dx^2} + f_u^{\varepsilon,\sigma}\right)\phi = \zeta\phi \quad \text{in } I, \\ &\alpha_i \frac{\partial u}{\partial n} + (1-\alpha_i)u = 0 \quad \text{at } x = i \ (i=0,1), \end{aligned}$$

where $0 \leqq \alpha_i \leqq 1$ $(i=0,1)$ and $\partial/\partial n$ denotes the outer normal derivative at the boundary point. Then, Lemma 1.3 and Corollary 1.3 also hold for $\phi_{0,\alpha}^{\varepsilon,\sigma}$ without any change. In particular, one can take Dirichlet boundary conditions at both ends or Dirichlet boundary condition at one end and Neumann boundary condition on the other end.

LEMMA 1.4. *Let* $\{\zeta_n^{\varepsilon,\sigma}\}_{n\geqq 0}$ *be the complete set of real simple eigenvalues of* (SL). *Then, it holds that for* $(\varepsilon,\sigma) \in \Omega_{\varepsilon_0,\sigma_0}$ *(see Fig.* 8),

$$(1.28) \qquad \begin{aligned} &\zeta_0^{\varepsilon,\sigma} > 0 > \zeta_1^{\varepsilon,\sigma} > \cdots > \zeta_n^{\varepsilon,\sigma} > \cdots, \\ &\zeta_0^{\varepsilon,\sigma} = \hat{\zeta}_0(\varepsilon,\sigma)\varepsilon\sigma + \mathrm{Exp}(\varepsilon,\sigma), \end{aligned}$$

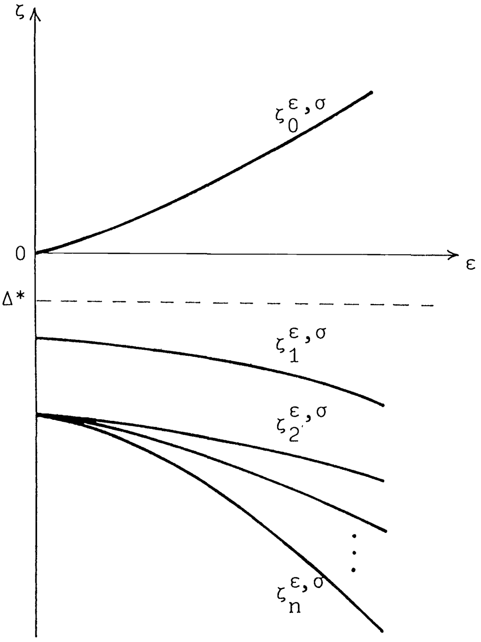


FIG. 8. *Asymptotic behaviors of the eigenvalues of* (SL).

and

$$\zeta_1^{\varepsilon,\sigma} < -\Delta^* < 0, \tag{1.29}$$

where $\hat{\zeta}_0(\varepsilon, \sigma)$ *is a positive continuous function in* $\Omega_{\varepsilon_0,\sigma_0}$ *uniquely extendable to* $\varepsilon = 0$ *as*

$$\hat{\zeta}_0^{*,0} \overset{\text{def}}{=} \lim_{\varepsilon \downarrow 0} \hat{\zeta}_0(\varepsilon, \sigma) = (\kappa^*)^2 \frac{dJ}{dv}(v^*) \int_0^{x_1^*(\sigma)} g(U^{*,\sigma}, V^{*,\sigma})\, dx, \tag{1.30}$$

and Exp *is a continuous function of* $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0,\sigma_0}$ *satisfying*

$$|\text{Exp}\,(\varepsilon, \sigma)| \leqq C \exp(-\gamma/\varepsilon). \tag{1.31}$$

Here, Δ^*, C, *and* γ *are positive constants independent of* $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0,\sigma_0}$.

Proof. For simplicity, we use the symbols $u^{\varepsilon,\sigma}$, $\tilde{u}^{\varepsilon,\sigma}$ and $\tilde{u}_y^{\varepsilon,\sigma}$ instead of $u(x; \varepsilon, \sigma)$, $\tilde{u}(y; \varepsilon, \sigma)$, and $(d/dy)\tilde{u}(y; \varepsilon, \sigma)$ in the following proof.

First we show the asymptotic formula (1.28). The principal eigenfunction $\hat{\phi}_0^{\varepsilon,\sigma}$ of $(\widetilde{\text{SL}})$ satisfies

$$\frac{d^2}{dy^2}\hat{\phi}_0^{\varepsilon,\sigma} + \tilde{f}_u^{\varepsilon,\sigma}\hat{\phi}_0^{\varepsilon,\sigma} = \zeta_0^{\varepsilon,\sigma}\hat{\phi}_0^{\varepsilon,\sigma}. \tag{1.32}$$

The stretched equations for (SP) becomes

$$(\widetilde{\text{SP}}) \qquad \frac{d^2}{dy^2}\tilde{u}^{\varepsilon,\sigma} + f(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma}) = 0, \qquad \frac{1}{\sigma}\frac{d^2}{dy^2}\tilde{v}^{\varepsilon,\sigma} + \varepsilon^2 g(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma}) = 0.$$

Differentiating the first equation of $(\widetilde{\text{SP}})$ with respect to y, we obtain

$$\frac{d^2}{dy^2}\tilde{u}_y^{\varepsilon,\sigma} = \tilde{f}_u^{\varepsilon,\sigma}\tilde{u}_y^{\varepsilon,\sigma} = -\tilde{f}_v^{\varepsilon,\sigma}\tilde{v}_y^{\varepsilon,\sigma}. \tag{1.33}$$

Multiplying $\tilde{u}_y^{\varepsilon,\sigma}$ to (1.32), we have

$$\left\langle \frac{d^2}{dy^2}\hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \right\rangle + \langle \tilde{f}_u^{\varepsilon,\sigma}\hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle = \zeta_0^{\varepsilon,\sigma}\langle \hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle.$$

Applying the integration by parts twice, and using $(\widetilde{\text{SP}})$ and (1.33), we have

$$\hat{\phi}_0^{\varepsilon,\sigma} f(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma})|_{-l/\varepsilon}^{r/\varepsilon} + \langle \hat{\phi}_0^{\varepsilon,\sigma}, -\tilde{f}_v^{\varepsilon,\sigma}\tilde{v}_y^{\varepsilon,\sigma} \rangle = \zeta_0^{\varepsilon,\sigma}\langle \hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle. \tag{1.34}$$

Owing to Remark 1.3, Lemma 1.3, and (1.24), we see that

$$\langle \hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle \xrightarrow[\varepsilon \downarrow 0]{} \kappa^{*^{-1}}\langle \hat{\phi}_0^*, \hat{\phi}_0^* \rangle = \kappa^{*^{-1}} \tag{1.35}$$

holds uniformly for $0 \leqq \sigma < \sigma_0$. This implies $\langle \hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle$ is positive and uniformly bounded away from zero for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0,\sigma_0}$. Integrating the second equation of $(\widetilde{\text{SP}})$, we have

$$\tilde{v}_y^{\varepsilon,\sigma}(y) = \left(-\varepsilon \int_{-l/\varepsilon}^{y} g(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma})\, dy\right)\varepsilon\sigma \overset{\text{def}}{=} \tilde{\Theta}(y; \varepsilon, \sigma)\varepsilon\sigma. \tag{1.36}$$

Substituting (1.36) into (1.34), we obtain the following

$$\zeta_0^{\varepsilon,\sigma} = \hat{\zeta}_0(\varepsilon, \sigma)\varepsilon\sigma + \text{Exp}\,(\varepsilon, \sigma),$$

where

$$\hat{\zeta}_0(\varepsilon, \sigma) \overset{\text{def}}{=} \left\langle \hat{\phi}_0^{\varepsilon,\sigma}, -\tilde{f}_v^{\varepsilon,\sigma}\left(-\varepsilon \int_{-l/\varepsilon}^{y} g(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma})\, dy\right)\right\rangle \Big/ P(\varepsilon, \sigma), \tag{1.37}$$

$$\text{Exp}\,(\varepsilon, \sigma) \overset{\text{def}}{=} [\hat{\phi}_0^{\varepsilon,\sigma} f(\tilde{u}^{\varepsilon,\sigma}, \tilde{v}^{\varepsilon,\sigma})]|_{l/\varepsilon}^{r/\varepsilon} / P(\varepsilon, \sigma) \tag{1.38}$$

and

$$P(\varepsilon, \sigma) = \langle \hat{\phi}_0^{\varepsilon,\sigma}, \tilde{u}_y^{\varepsilon,\sigma} \rangle. \tag{1.39}$$

It is easily seen from Theorem 1.1, (A.0) and the continuous dependence of solutions on coefficients that $\hat{\zeta}_0(\varepsilon, \sigma)$ and Exp (ε, σ) are continuous functions of $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$. In view of the definition (1.38), the estimate (1.31) easily follows from (1.24) and (1.35). In order to study the behavior of $\hat{\zeta}_0(\varepsilon, \sigma)$, we first consider the term $\tilde{\Theta}(y; \varepsilon, \sigma)$. It is apparent from the definition (1.36) that $\tilde{\Theta}(y; \varepsilon, \sigma)$ is the stretched function of

$$\Theta(x; \varepsilon, \sigma) \overset{\text{def}}{=} -\int_0^x g(u^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})\, dx$$

by using $y = (x - x_1(\varepsilon, \sigma))/\varepsilon$. $\Theta(x; \varepsilon, \sigma)$ is uniformly bounded in $C^1(I)$-sense for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$, and converges to the following value as $\varepsilon \downarrow 0$ at $x = x_1(\varepsilon, \sigma)$,

$$\Theta(x_1(\varepsilon, \sigma); \varepsilon, \sigma) \xrightarrow[\varepsilon \downarrow 0]{} -\int_0^{x_1^*(\sigma)} g(U^{*,\sigma}, V^{*,\sigma})\, dx.$$

Consequently, we can see that the stretched function $\tilde{\Theta}(y; \varepsilon, \sigma)$ satisfies

$$\tilde{\Theta}(y; \varepsilon, \sigma) \xrightarrow[\varepsilon \downarrow 0]{} -\int_0^{x_1^*(\sigma)} g(U^{*,\sigma}, V^{*,\sigma})\, dx, \tag{1.40}$$

compact uniformly in C^1-sense. On the other hand, it follows from Corollary 1.1, Lemma 1.3 and (1.24) that

$$\begin{aligned} \lim_{\varepsilon \downarrow 0} \langle \hat{\phi}_0^{\varepsilon,\sigma}, -\tilde{f}_v^{\varepsilon,\sigma} \rangle &= -\kappa^* \int_{-\infty}^{\infty} \tilde{f}_v^* \tilde{u}_y^*\, dy \\ &= -\kappa^* \int_{h_-(v^*)}^{h_+(v^*)} \tilde{f}_v^*\, d\tilde{u}^* \\ &= -\kappa^* \frac{dJ}{dv}(v^*) > 0 \quad (\text{see (A.2)}). \end{aligned} \tag{1.41}$$

Thus, combining the above results (1.35), (1.40), (1.41), and recalling (1.24) again, we can see that

$$\hat{\zeta}_0^{*,\sigma} \overset{\text{def}}{=} \lim_{\varepsilon \downarrow 0} \hat{\zeta}_0(\varepsilon, \sigma) = (\kappa^*)^2 \frac{dJ}{dv}(v^*) \int_0^{x_1^*(\sigma)} g(U^{*,\sigma}, V^{*,\sigma})\, dx. \tag{1.30}$$

Since $x_1^*(\sigma)$ is a continuous function of σ (see Proposition 1.2), so is $\hat{\zeta}_0^{*,\sigma}$. Here, it holds from (A.2) and (A.3) that

$$\hat{\zeta}_0^{*,\sigma} > 0 \quad \text{for } 0 \leqq \sigma < \sigma_0. \tag{1.42}$$

Since the convergence (1.30) is uniform for $0 \leqq \sigma < \sigma_0$, we can conclude from (1.42) that $\hat{\zeta}_0(\varepsilon, \sigma)$ is uniquely extendable to be a strictly positive continuous function on $\bar{\Omega}_{\varepsilon_0, \sigma_0}$.

Finally, we shall show (1.29), namely, $\zeta_1^{\varepsilon,\sigma}$ is bounded away from zero and strictly negative for small ε. As we have seen in (1.23),

$$\tilde{f}_u^{\varepsilon,\sigma} < -\mu < 0, \tag{1.43}$$

holds outside of a fixed finite interval I_μ, where μ and I_μ do not depend on $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$. Therefore, we can assume without loss of generality that

$$\tilde{f}_u^{\varepsilon,\sigma} - \zeta_1^{\varepsilon,\sigma} < -\hat{\mu} < 0 \tag{1.44}$$

holds outside of a fixed finite interval $I_{\hat{\mu}}$, where $\hat{\mu}$ and $I_{\hat{\mu}}$ do not depend on

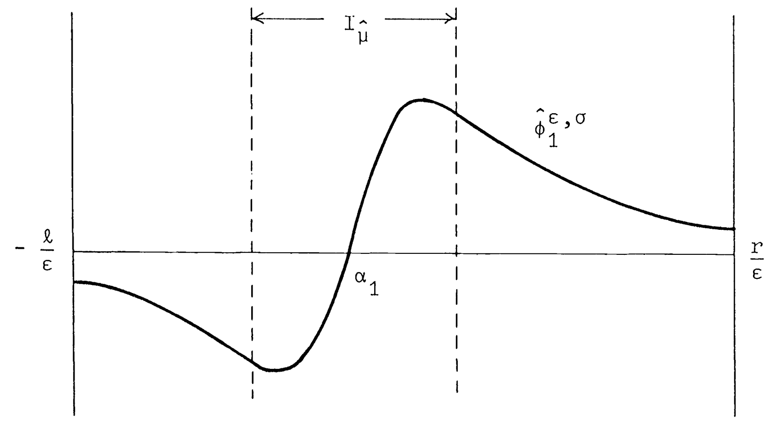


FIG. 9. *Functional form of* $\hat\phi_1^{\varepsilon,\sigma}$.

$(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}$; otherwise $\zeta_1^{\varepsilon,\sigma}$ already satisfies (1.29). From the nodal property, the associated eigenfunction $\hat\phi_1^{\varepsilon,\sigma}$ has only one zero. This zero, denoted by α_1, is located inside of $I_{\hat\mu}$, since $\hat\phi_1^{\varepsilon,\sigma}$and its y-derivative behave, without changing signs, monotonically in the region where $\hat f_u^{\varepsilon,\sigma}-\zeta_1^{\varepsilon,\sigma}<-\hat\mu<0$. See Fig. 9.

Then, we may assume that in $(\alpha_1, r/\varepsilon]$, $\hat\phi_1^{\varepsilon,\sigma}>0$. $\hat\phi_1^{\varepsilon,\sigma}$ satisfies the equation

$$\frac{d^2}{dy^2}\hat\phi_1^{\varepsilon,\sigma}+\tilde f_u^{\varepsilon,\sigma}\hat\phi_1^{\varepsilon,\sigma}=\zeta_1^{\varepsilon,\sigma}\hat\phi_1^{\varepsilon,\sigma}. \tag{1.45}$$

By using (1.44) and (1.45), we see that $\hat\phi_1^{\varepsilon,\sigma}$ satisfies the same type of estimate as (1.24). Making $(1.32)\times\hat\phi_1^{\varepsilon,\sigma}-(1.46)\times\hat\phi_0^{\varepsilon,\sigma}$ and integrating over $(\alpha_1, r/\varepsilon)$, we have

$$\left(\frac{d}{dy}\hat\phi_0^{\varepsilon,\sigma}\right)\hat\phi_1^{\varepsilon,\sigma}\Big|_{\alpha_1}^{r/\varepsilon}-\left(\frac{d}{dy}\hat\phi_1^{\varepsilon,\sigma}\right)\hat\phi_0^{\varepsilon,\sigma}\Big|_{\alpha_1}^{r/\varepsilon}=(\zeta_0^{\varepsilon,\sigma}-\zeta_1^{\varepsilon,\sigma})\int_{\alpha_1}^{r/\varepsilon}\hat\phi_0^{\varepsilon,\sigma}\hat\phi_1^{\varepsilon,\sigma}\,dy. \tag{1.46}$$

Noting that $\hat\phi_1^{\varepsilon,\sigma}(\alpha_1)=0$ and the exponentially decaying property of eigenfunction in the outer region, (1.46) becomes

$$\left(\frac{d}{dy}\hat\phi_1^{\varepsilon,\sigma}\right)(\alpha_1)\cdot\hat\phi_0^{\varepsilon,\sigma}(\alpha_1)+O\{\exp(-\bar\gamma/\varepsilon)\}=(\zeta_0^{\varepsilon,\sigma}-\zeta_1^{\varepsilon,\sigma})\int_{\alpha_1}^{r/\varepsilon}\hat\phi_0^{\varepsilon,\sigma}\hat\phi_1^{\varepsilon,\sigma}\,dy, \tag{1.47}$$

where $\bar\gamma$ is a positive constant which does not depend on (ε,σ). Since both $\hat\phi_0^{\varepsilon,\sigma}$ and $\hat\phi_1^{\varepsilon,\sigma}$ are L^2-normalized eigenfunctions on $\tilde I$ and decay monotonically with exponential order outside of $I_{\hat\mu}$, it is easy to verify by contradiction that

$$\left(\frac{d}{dy}\hat\phi_1^{\varepsilon,\sigma}\right)(\alpha_1)\cdot\hat\phi_0^{\varepsilon,\sigma}(\alpha_1)\quad\text{and}\quad\int_{\alpha_1}^{r/\varepsilon}\hat\phi_0^{\varepsilon,\sigma}\hat\phi_1^{\varepsilon,\sigma}\,dy \tag{1.48}$$

are positive and uniformly bounded away from zero for $(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}$. Consequently, using (1.47) and (1.28), we can conclude (1.29).

Remark 1.5. Lemma 1.4 also holds for $(\mathrm{SL})_\alpha$ without an essential change, since the asymptotic limit $\lim_{\varepsilon\downarrow 0}\hat\phi_{0,\alpha}^{\varepsilon,\sigma}(=\hat\phi_0^*)$ does not depend on α (see Remark 1.4), and the exponentially decaying estimate (1.24) also holds for the L^2-normalized eigenfunctions of the stretched problem $(\widetilde{\mathrm{SL}})_\alpha$. Especially, using the similar identify as (1.46), we can show that the difference of principal eigenvalues $|\zeta_{0,\alpha}^{\varepsilon,\sigma}-\zeta_{0,\alpha'}^{\varepsilon,\sigma}|$ $(0\leqq\alpha,\alpha'\leqq 1)$ is at most of exponentially decaying order. Therefore, the asymptotic limit of the coefficient $\hat\zeta_{0,\alpha}(\varepsilon,\sigma)$ of $\varepsilon\sigma$ (see (1.28)) does not depend on α, namely,

$$\lim_{\varepsilon\downarrow 0}\hat\zeta_{0,\alpha}(\varepsilon,\sigma)=\hat\zeta_0^*(\sigma). \tag{1.49}$$

2. Singular limit eigenvalue problem and the behavior of critical eigenvalues. In order to obtain the stability of SPS1, we have to show the following:

(P2)$_a$ All noncritical eigenvalue of (LP) have strictly negative real parts.

(P2)$_b$ If there is a critical eigenvalues of (LP), it has to approach to zero from the left-half plane ($\mathrm{Re}\,\lambda < 0$) as $\varepsilon \downarrow 0$.

Note that when $\varepsilon \downarrow 0$, (LP) becomes singular in two ways, namely, the degeneracy of the second-order term due to ε, and the discontinuity of coefficients $f_u^{\varepsilon,\sigma}$, $f_v^{\varepsilon,\sigma}$, $g_u^{\varepsilon,\sigma}$ and $g_v^{\varepsilon,\sigma}$ at the layer position.

A natural and standard approach to deal with the above problems may be to look for an appropriate limit problem of (LP) as $\varepsilon \downarrow 0$, and then extract from it information on the spectral behavior for $\varepsilon > 0$. In fact, the most optimistic way may be to put $\varepsilon = 0$ in (LP) to have

$$\text{(REP)} \qquad f_u^{*,\sigma} w + f_v^{*,\sigma} z = \lambda w, \qquad \frac{1}{\sigma} z_{xx} + g_u^{*,\sigma} w + g_v^{*,\sigma} z = \lambda z,$$

subject to zero flux boundary conditions for z, where all the partial derivatives are evaluated at the reduced solution (1.2), i.e., $f_u^{*,\sigma} = f_u(U^{*,\sigma}(x), V^{*,\sigma}(x))$ and so on. This is called the *reduced eigenvalue problem* (REP) of (LP), since it is formally obtained by linearizing (RP) in § 1. Since $f_u^{*,\sigma} < 0$ (see (A.3)), we can solve the first equation of (REP) as $w = -f_v^{*,\sigma} z/(f_u^{*,\sigma} - \lambda)$, assuming that $\mathrm{Re}\,\lambda > -\mu$ (see (1.23)). Substituting this into the second equation of (REP), we obtain after some computation

$$\text{(2.1)} \qquad \frac{1}{\sigma} z_{xx} + \frac{\det^{*,\sigma}}{f_u^{*,\sigma} - \lambda} z = \lambda \left(1 + \frac{g_v^{*,\sigma}}{f_u^{*,\sigma} - \lambda} \right) z,$$

where $\det^{*,\sigma} = f_u^{*,\sigma} g_v^{*,\sigma} - f_v^{*,\sigma} g_u^{*,\sigma}$. Recalling (A.4) and (A.5), we can show as in Proposition 2.1 that there are no spectra of (2.1) in the region $\mathrm{Re}\,\lambda > -\mu^*$ for an appropriate positive number μ^*. Thus, one may think that the stability of SPS1 is obtained in a straightforward way from (REP). However, if we look at (REP) carefully, we realize that it includes no terms which come from the interior transition layer. The reason for this is clear. (REP) is the formal L^2-limit of (LP). Therefore, the contribution from the layer part, the width of which is of $O(\varepsilon)$, becomes negligible as $\varepsilon \downarrow 0$. Recalling that not all solutions of (RP) are extended to be solutions of (SP) for $\varepsilon > 0$, (REP) may be inadequate for our purpose in the sense that it is too rough to solve our whole problem. In fact, it loses information on the behavior of critical eigenvalues. Nevertheless, (REP) is useful to solve the first problem (P2)$_a$ (see Proposition 2.1), and Lemma 2.2 plays an important role to justify this step.

Now the problem is that "What is the exact limit of (LP) of $\varepsilon \downarrow 0$?", which inherits necessary information from the layer part, governing the asymptotic behavior of critical eigenvalues. Lemma 2.3 becomes a keystone to answer this question. It shows that when $\varepsilon \downarrow 0$, the information on layer is condensed into *the Dirac's δ-function* at $x = x_1^*(\sigma)$, and its magnitude determines the asymptotic behavior of the critical eigenvalue. Using Lemma 2.3 as well as Lemma 2.2, we obtain a new limiting problem of (LP) in H^{-1}-sense as $\varepsilon \downarrow 0$ called the *singular limit eigenvalue problem* (SLEP), which essentially enables us to solve both problems (PS)$_a$ and (P2)$_b$.SLEP is simple and much more tractable than the original linearized problem (LP). The justification of SLEP as well as the uniqueness and simplicity of the critical eigenvalue will be given in the next section.

First we note the following lemma.

LEMMA 2.1. *It holds that $\zeta_0^{\varepsilon,\sigma} \notin \sigma(\mathrm{LP})$ for small $\varepsilon > 0$, i.e., the principal eigenvalue of $L^{\varepsilon,\sigma}$ in Lemma* 1.4 *never becomes an eigenvalue of* (LP) *for small positive ε, where $\sigma(\mathrm{LP})$ denotes the set of spectra of* (LP).

Proof. See Appendix 2.

Remark 2.1. Though $\zeta_0^{\varepsilon,\sigma} \notin \sigma(\mathrm{LP})$ for small $\varepsilon > 0$, $\lim_{\varepsilon\downarrow 0} \zeta_0^{\varepsilon,\sigma}(=0)$ is contained in the limit set of $\sigma(\mathrm{LP})$ as $\varepsilon \downarrow 0$ (see Lemma 1.4 and Main Theorem).

Now we start to solve (LP). Hereafter, we only consider the spectra of (LP) which lie in the region Λ_1 defined by

$$\text{(2.2)} \qquad \Lambda_1 = \{\lambda \mid \operatorname{Re}\lambda > -\mu_1 > \max\{-\Delta^*, -\mu\}\} \quad \text{for some fixed } \mu_1 > 0,$$

where $-\Delta^*$ and $-\mu$ are the constants which appeared in (1.29) and (1.23), respectively. Apparently, this restriction does not lose any generality. Noting Lemmas 2.1, 1.4 and (2.2), the first equation of (LP) can be solved with respect to w as

$$\text{(2.3)} \qquad w = (L^{\varepsilon,\sigma} - \lambda)^{-1}(-f_v^{\varepsilon,\sigma} z).$$

Applying the eigenfunction expansion to (2.3), we have

$$\text{(2.4)} \qquad w = \frac{\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma}\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda}\phi_0^{\varepsilon,\sigma} + \sum_{n\geqq 1} \frac{\langle -f_v^{\varepsilon,\sigma} z, \phi_n^{\varepsilon,\sigma}\rangle}{\zeta_n^{\varepsilon,\sigma} - \lambda}\phi_n^{\varepsilon,\sigma}.$$

Substituting (2.4) into the second equation of (LP), we obtain the scalar eigenvalue problem for z:

$$\text{(2.5)} \qquad \frac{1}{\sigma} z_{xx} + \frac{\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma}\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda} g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma} + g_u^{\varepsilon,\sigma}(L^{\varepsilon,\sigma} - \lambda)^{\dagger}(-f_v^{\varepsilon,\sigma} z) + g_v^{\varepsilon,\sigma} z = \lambda z,$$

$$z \in H^2(I) \cap H_N^1(I),$$

where $(L^{\varepsilon,\sigma} - \lambda)^{\dagger}$ is defined by

$$\text{(2.6)} \qquad (L^{\varepsilon,\sigma} - \lambda)^{\dagger}(\cdot) = \sum_{n\geqq 1} \frac{\langle \cdot, \phi_n^{\varepsilon,\sigma}\rangle}{\zeta_n^{\varepsilon,\sigma} - \lambda}\phi_n^{\varepsilon,\sigma}, \qquad L^2(I) \to L^2(I) \cap \{\phi_0^{\varepsilon,\sigma}\}^{\perp}.$$

It follows from (1.29) and (2.2) that $\zeta_n^{\varepsilon,\sigma} - \lambda$ $(n \geqq 1)$ is uniformly bounded away from zero. Therefore, in view of the definition (2.6), $(L^{\varepsilon,\sigma} - \lambda)^{\dagger}$ is a uniformly L^2-bounded operator for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0,\sigma_0}$ and $\lambda \in \Lambda_1$. More precisely, we have the estimate,

$$\text{(2.7)} \qquad \|(L^{\varepsilon,\sigma} - \lambda)^{\dagger}\| \leqq \frac{1}{|\zeta_1^{\varepsilon,\sigma} - \lambda|}.$$

Remark 2.2. It follows from the definition (2.6) that the generalized inverse $(L^{\varepsilon,\sigma} - \lambda)^{\dagger}$ depends analytically on λ for $\lambda \in \Lambda_1$ as an L^2-bounded operator-valued function, and satisfies the resolvent formula

$$(L^{\varepsilon,\sigma} - \lambda_1)^{\dagger} - (L^{\varepsilon,\sigma} - \lambda_2)^{\dagger} = (\lambda_1 - \lambda_2)(L^{\varepsilon,\sigma} - \lambda_1)^{\dagger}(L^{\varepsilon,\sigma} - \lambda_2)^{\dagger}.$$

In order to study the asymptotic behavior of the spectra of (LP), we need explicit characterizations of the asymptotic form of the second and the third terms of (2.5) as $\varepsilon \downarrow 0$.

The first key lemma is about the behavior of the operator $(L^{\varepsilon,\sigma} - \lambda)^{\dagger}$ as $\varepsilon \downarrow 0$.

LEMMA 2.2. (*The first key lemma.*) *Let $F(u, v)$ be a smooth function of u and v. Then,*

$$\text{(2.8)} \qquad (L^{\varepsilon,\sigma} - \lambda)^{\dagger}(F^{\varepsilon,\sigma} h) \xrightarrow[\varepsilon\downarrow 0]{} \frac{F^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \quad \textit{strongly in } L^2\textit{-sense},$$

for any function $h \in L^2(I) \cap L^\infty(I)$ *and* $\lambda \in \Lambda_1$, *where* $F^{\varepsilon,\sigma} = F(u(x;\varepsilon,\sigma),\ v(x;\varepsilon,\sigma))$ *and* $F^{*,\sigma} = F(U^{*,\sigma}(x),\ V^{*,\sigma}(x))$ (*see* (1.2)). *The convergence* (2.8) *is uniform for* $0 \leqq \sigma < \sigma_0$ *and* $\lambda \in \Lambda_1$. *Furthermore, if h belongs to* $H^1(I)$, *the above convergence is also uniform on a bounded set in* $H^1(I)$.

Proof. Let $s_k^*(x)$ be a smooth cut-off function at $x = x_1^*(\sigma)$ satisfying

$$s_k^*(x) = \begin{cases} 1 & \text{if } |x - x_1^*(\sigma)| \geqq k/2, \\ 0 & \text{if } |x - x_1^*(\sigma)| \leqq k/4 \end{cases}$$

with

$$(2.9) \qquad 0 \leqq s_k^*(x) \leqq 1, \quad \sup_{x \in I} \left| \frac{d}{dx} s_k^*(x) \right| \leqq M_k, \quad \sup_{x \in I} \left| \frac{d^2}{dx^2} s_k^*(x) \right| \leqq M_k.$$

Here, the positive constant M_k tends to $+\infty$ as $k \downarrow 0$. Define $F_k^{\varepsilon,\sigma}$ by $F_k^{\varepsilon,\sigma} = s_k^* F^{\varepsilon,\sigma}$. Then, it holds that

$$(2.10) \qquad \lim_{k \downarrow 0} \| F_k^{\varepsilon,\sigma} - F^{\varepsilon,\sigma} \|_{L^2} = 0 \quad \text{uniformly for } (\varepsilon, \sigma) \in \Omega_{\varepsilon_0 \cdot \sigma_0}$$

because of the uniform boundedness of $F^{\varepsilon,\sigma}$ with respect to ε and σ. Let $H^{\varepsilon,\sigma}$ and $H_k^{\varepsilon,\sigma}$ be defined by

$$H^{\varepsilon,\sigma} = (L^{\varepsilon,\sigma} - \lambda)^\dagger F^{\varepsilon,\sigma} h \quad \text{and} \quad H_k^{\varepsilon,\sigma} = (L^{\varepsilon,\sigma} - \lambda)^\dagger F_k^{\varepsilon,\sigma} h,$$

respectively.

It suffices to prove Lemma 2.2 where $F^{\varepsilon,\sigma}$ and $F^{*,\sigma}$ are replaced by $F_k^{\varepsilon,\sigma}$ and $F_k^{*,\sigma}$, respectively, for any $k > 0$. In fact, for an arbitrary $\rho > 0$, there exists a $k_0 > 0$ such that

$$\| H^{\varepsilon,\sigma} - H_k^{\varepsilon,\sigma} \|_{L^2} \leqq \| (L^{\varepsilon,\sigma} - \lambda)^\dagger \| \, \| F^{\varepsilon,\sigma} - F_k^{\varepsilon,\sigma} \|_{L^2} \| h \|_{L^\infty} < \frac{\rho}{3},$$

and

$$\left\| \frac{F^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} - \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \right\|_{L^2} < \frac{\rho}{3},$$

hold for $0 < k < k_0$ uniformly with respect to ε and σ. On the other hand, if this lemma is true for the cut-off case, then, for any fixed k_1 with $0 < k_1 < k_0$, we can find an ε_1 such that

$$\left\| H_{k_1}^{\varepsilon,\sigma} - \frac{F_{k_1}^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \right\|_{L_2} < \rho/3 \quad \text{for } 0 < \varepsilon < \varepsilon_1$$

holds. Combining these three inequalities, we have

$$(2.11) \qquad \begin{aligned} \left\| H^{\varepsilon,\sigma} - \frac{F^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \right\|_{L_2} &< \| H^{\varepsilon,\sigma} - H_{k_1}^{\varepsilon,\sigma} \|_{L^2} + \left\| H_{k_1}^{\varepsilon,\sigma} - \frac{F_{k_1}^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \right\|_{L_2} \\ &\quad + \left\| \frac{F_{k_1}^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} - \frac{F^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \right\|_{L^2} < \rho \end{aligned}$$

for $0 < \varepsilon < \varepsilon_1$. Since ρ is arbitrary, (2.11) implies (2.8). The uniformity of convergence for $\lambda \in \Lambda_1$ follows from that of the above three inequalities.

Now, let us prove Lemma 2.2 for the cut-off case,

$$(2.12) \qquad (L^{\varepsilon,\sigma} - \lambda)^\dagger (F_k^{\varepsilon,\sigma} h) \xrightarrow[\varepsilon \downarrow 0]{} \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda} \quad \text{in } L^2\text{-strong sense.}$$

Here, we can assume without loss of generality that h is smooth and $h_x = 0$ on ∂I, since any bounded L^2-function can be approximated arbitrarily close by such a function in L^2-sense. We divide $H_k^{\varepsilon,\sigma}$ into two parts,

$$(2.13)\qquad H_k^{\varepsilon,\sigma} = H_{1,k}^{\varepsilon,\sigma} + H_{2,k}^{\varepsilon,\sigma}, \quad \text{where } H_{1,k}^{\varepsilon,\sigma} = F_k^{*,\sigma} h/(f_u^{*,\sigma} - \lambda).$$

We have to show that $H_{2,k}^{\varepsilon,\sigma}$ goes to zero in L^2-sense uniformly for $\lambda \in \Lambda_1$, when $\varepsilon \downarrow 0$. First note that $H_{1,k}^{\varepsilon,\sigma}$ becomes a smooth function due to $F_k^{*,\sigma}$, though $f_u^{*,\sigma}$ has a discontinuity at $x = x_1^*(\sigma)$. The equation $H_k^{\varepsilon,\sigma} = (L^{\varepsilon,\sigma} - \lambda)^\dagger (F_k^{\varepsilon,\sigma} h)$ is equivalent to

$$(2.14)\qquad (L^{\varepsilon,\sigma} - \lambda) H_k^{\varepsilon,\sigma} = F_k^{\varepsilon,\sigma} h - P(F_k^{\varepsilon,\sigma} h),\ H_k^{\varepsilon,\sigma} \in \{\phi_0^{\varepsilon,\sigma}\}^\perp,$$

where P is an orthogonal projection onto $\phi_0^{\varepsilon,\sigma}$, i.e., $P(\cdot) = \langle \cdot, \phi_0^{\varepsilon,\sigma}\rangle \phi_0^{\varepsilon,\sigma}$. Substituting $H_{1,k}^{\varepsilon,\sigma}$ of (2.13) into (2.14), we have the following equation for $H_{2,k}^{\varepsilon,\sigma}$,

$$(2.15)\quad (L^{\varepsilon,\sigma} - \lambda) H_{2,k}^{\varepsilon,\sigma} = F_k^{\varepsilon,\sigma} h - P(F_k^{\varepsilon,\sigma} h) - \frac{f_u^{\varepsilon,\sigma} - \lambda}{f_u^{*,\sigma} - \lambda} F_k^{*,\sigma} h - \varepsilon^2 \frac{d^2}{dx^2}\left(\frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}\right).$$

We denote the right-hand side of (2.15) by $R_k^{\varepsilon,\sigma}$. Since both $F_k^{*,\sigma}/(f_u^{*,\sigma} - \lambda)$ and h are smooth and satisfy Neumann boundary conditions, we have the following by using integration by parts;

$$\begin{aligned}
\langle R_k^{\varepsilon,\sigma}, \phi_0^{\varepsilon,\sigma}\rangle &= -\left\langle \frac{f_u^{\varepsilon,\sigma} - \lambda}{f_u^{*,\sigma} - \lambda} F_k^{*,\sigma} h, \phi_0^{\varepsilon,\sigma}\right\rangle - \left\langle \varepsilon^2 \frac{d^2}{dx^2}\left(\frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}\right), \phi_0^{\varepsilon,\sigma}\right\rangle \\
&= -\left\langle \frac{f_u^{\varepsilon,\sigma} - \lambda}{f_u^{*,\sigma} - \lambda} F_k^{*,\sigma} h, \phi_0^{\varepsilon,\sigma}\right\rangle - \left\langle \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}, \varepsilon^2 \frac{d^2}{dx^2}\phi_0^{\varepsilon,\sigma}\right\rangle \\
&= -\left\langle \frac{f_u^{\varepsilon,\sigma} - \lambda}{f_u^{*,\sigma} - \lambda} F_k^{*,\sigma} h, \phi_0^{\varepsilon,\sigma}\right\rangle - \left\langle \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}, (-f_u^{\varepsilon,\sigma} + \zeta_0^{\varepsilon,\sigma})\phi_0^{\varepsilon,\sigma}\right\rangle \\
&= (\lambda - \zeta_0^{\varepsilon,\sigma})\left\langle \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}, \phi_0^{\varepsilon,\sigma}\right\rangle.
\end{aligned}$$

Thus, we have

$$(2.16)\qquad P(R_k^{\varepsilon,\sigma}) = (\lambda - \zeta_0^{\varepsilon,\sigma})\left\langle \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}, \phi_0^{\varepsilon,\sigma}\right\rangle \phi_0^{\varepsilon,\sigma}.$$

In view of (2.15) and (2.16), we can see that $H_{2,k}^{\varepsilon,\sigma}$ is given by

$$(2.17)\qquad H_{2,k}^{\varepsilon,\sigma} = (L^{\varepsilon,\sigma} - \lambda)^\dagger (R_k^{\varepsilon,\sigma}) - \left\langle \frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}, \phi_0^{\varepsilon,\sigma}\right\rangle \phi_0^{\varepsilon,\sigma}.$$

It is apparent from (2.13) and (2.17) that

$$P(H_k^{\varepsilon,\sigma}) = P(H_{1,k}^{\varepsilon,\sigma} + H_{2,k}^{\varepsilon,\sigma}) = 0.$$

Using the expression (2.17) for $H_{2,k}^{\varepsilon,\sigma}$, we shall show that $H_{2,k}^{\varepsilon,\sigma} \to 0$ in L^2-sense as $\varepsilon \downarrow 0$. Since $F_k^{*,\sigma} h/(f_u^{*,\sigma} - \lambda)$ is uniformly bounded for $\lambda \in \Lambda_1$, we can see from Corollary 1.3 that when $\varepsilon \downarrow 0$, the second term of (2.17) converges to zero with $O(\sqrt{\varepsilon})$ in L^2-sense uniformly for $\lambda \in \Lambda_1$. For the first term of (2.17), it suffices to prove $\|R_k^{\varepsilon,\sigma}\|_{L^2} \downarrow 0$ as $\varepsilon \downarrow 0$, because of the uniform L^2-boundedness of $(L^{\varepsilon,\sigma} - \lambda)^\dagger$ for $\lambda \in \Lambda_1$ (see (2.7)). Recalling that the C^2-norm of the outer part of the reduced solution is uniformly bounded for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$, and that (2.9), we can see that

$$(2.18)\qquad \lim_{\varepsilon \downarrow 0} \left\| \varepsilon^2 \frac{d^2}{dx^2}\left(\frac{F_k^{*,\sigma} h}{f_u^{*,\sigma} - \lambda}\right)\right\|_{L^2} = 0 \quad \text{uniformly for } \lambda \in \Lambda_1.$$

It is easily seen that

$$\lim_{\varepsilon\downarrow 0}\left\|F_k^{\varepsilon,\sigma}h-\frac{f_u^{\varepsilon,\sigma}-\lambda}{f_u^{*,\sigma}-\lambda}F_k^{*,\sigma}h\right\|_{L^2}=0 \quad \text{uniformly for } \lambda\in\Lambda_1, \tag{2.19}$$

$$\lim_{\varepsilon\downarrow 0}\|P(F_k^{\varepsilon,\sigma}h)\|_{L^2}=0. \tag{2.20}$$

Using (2.18)-(2.20), we obtain

$$\lim_{\varepsilon\downarrow 0}\|R_k^{\varepsilon,\sigma}\|_{L^2}=0 \quad \text{uniformly for } \lambda\in\Lambda_1,$$

which leads to the conclusion

$$H_{2,k}^{\varepsilon,\sigma}\xrightarrow[\varepsilon\downarrow 0]{} 0 \quad \text{in } L^2\text{-sense uniformly for } \lambda\in\Lambda_1. \tag{2.21}$$

In view of (2.13), we can see that (2.12) and therefore (2.8) are proved. It also follows from the above arguments that the convergence (2.8) is uniform for $0\leqq\sigma<\sigma_0$.

Finally, we show, by contradiction, the last part of Lemma 2.2. Suppose that the convergence (2.8) is not uniform on $H_M^1 \overset{\text{def}}{=} \{h\in H^1(I)|\ \|h\|_{H^1}\leqq M\}$, then we can find a positive constant δ and a sequence $\{h_{\varepsilon_n}\}(\varepsilon_n\downarrow 0$, as $n\uparrow\infty)$ in H_M^1 such that

$$\left\|(L^{\varepsilon_n,\sigma}-\lambda)^{\dagger}(F^{\varepsilon_n,\sigma}h_{\varepsilon_n})-\frac{F^{*,\sigma}h_{\varepsilon_n}}{f_u^{*,\sigma}-\lambda}\right\|_{L^2}\geqq\delta>0 \quad \text{for } n\geqq 1, \tag{2.22}$$

where δ does not depend on n. Since H_M^1 is compact in $L^2(I)$, we can choose a subsequence $\{h_{\varepsilon_n}\}$ (we use the same notation as before), which converges to $\hat{h}$ strongly in L^2-sense. Then, we have the inequality,

$$\begin{aligned}\left\|(L^{\varepsilon_n,\sigma}-\lambda)^{\dagger}(F^{\varepsilon_n,\sigma}h_{\varepsilon_n})-\frac{F^{*,\sigma}h_{\varepsilon_n}}{f_u^{*,\sigma}-\lambda}\right\|_{L^2}&\leqq\|(L^{\varepsilon_n,\sigma}-\lambda)^{\dagger}\{F^{\varepsilon_n,\sigma}(h_{\varepsilon_n}-\hat{h})\}\|_{L^2}\\&\quad+\left\|(L^{\varepsilon_n,\sigma}-\lambda)^{\dagger}(F^{\varepsilon_n,\sigma}\hat{h})-\frac{F^{*,\sigma}\hat{h}}{f_u^{*,\sigma}-\lambda}\right\|_{L^2}\\&\quad+\left\|\frac{F^{*,\sigma}}{f_u^{*,\sigma}-\lambda}(\hat{h}-h_{\varepsilon_n})\right\|_{L^2}.\end{aligned} \tag{2.23}$$

The right-hand side of (2.23) converges to zero as $n\uparrow\infty$, which contradicts (2.22). Thus, we have finished the proof of Lemma 2.2.

Remark 2.3. Lemma 2.2 also holds for the general boundary conditions $(\mathrm{SL})_\alpha$ in Remark 1.4.

Making use of Lemma 2.2, the next proposition shows that noncritical eigenvalues are not dangerous to the stability of SPS1.

PROPOSITION 2.1 (*a priori bound for noncritical eigenvalues*). *Let B_δ be a closed ball with center at the origin and radius δ in the complex plane $\mathbb{C}$. Suppose that λ is an arbitrary noncritical eigenvalue of* (LP) *which stays outside of B_δ for small ε.*

Then, there exist positive constants μ^* and ε_δ such that

$$\operatorname{Re}\lambda<-\mu^*<0 \quad \textit{for } 0<\varepsilon<\varepsilon_\delta, \tag{2.24}$$

where μ^* does not depend on δ and $(\varepsilon,\sigma)\in\Omega_{\varepsilon_\delta,\sigma_0}$.

Proof. Let $\Lambda_{1,\delta}(\subset\Lambda_1)$ be defined by

$$\Lambda_{1,\delta}=\{\lambda\in\mathbb{C}|\lambda\in\Lambda_1 \text{ and } \lambda\in C\backslash B_\delta\}. \tag{2.25}$$

We can assume without loss of generality that $\lambda \in \Lambda_{1,\delta}$ in the following. Let us start from the eigenvalue problem (2.5) for $\sigma > 0$.

$$\frac{1}{\sigma} z_{xx} + \frac{\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma} \rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda} g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma} + g_u^{\varepsilon,0} (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} z) + g_v^{\varepsilon,\sigma} z = \lambda z, \tag{2.5}$$
$$z \in H^2(I) \cap H_N^1(I).$$

We introduce the following bilinear form on $H_N^1(I)$ associated with (2.5):

$$\begin{aligned} B^{\varepsilon,\sigma,\lambda}(z^1, z^2) = {} & \frac{1}{\sigma} \langle z_x^1, z_x^2 \rangle - \frac{1}{\zeta_0^{\varepsilon,\sigma} - \lambda} \langle -f_v^{\varepsilon,\sigma} z^1, \phi_0^{\varepsilon,\sigma} \rangle \langle g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}, z^2 \rangle \\ & - \langle (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} z^1), g_u^{\varepsilon,\sigma} z^2 \rangle - \langle g_v^{\varepsilon,\sigma} z^1, z^2 \rangle \\ & + \lambda \langle z^1, z^2 \rangle \quad \text{for } z^1, z^2 \in H_N^1(I). \end{aligned} \tag{2.26}$$

We shall show the boundedness and positivity of (2.26). Since $\lambda \in \Lambda_{1,\delta}$, $\lim_{\varepsilon \downarrow 0} \zeta_0^{\varepsilon,\sigma} = 0$ (Lemma 1.4) and $\int \phi_0^{\varepsilon,\sigma} \, dx = O(\sqrt{\varepsilon})$ (Corollary 1.3), we can see that there exists a $\varepsilon_\delta > 0$ such that the estimate

$$|\text{the second term of (2.26)}| \leqq \frac{c_1 \varepsilon}{\delta} \|z^1\|_{H^1} \|z^2\|_{H^1} \tag{2.27}$$

holds for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_\delta, \sigma_0}$ and $\lambda \in \Lambda_{1,\delta}$, where $c_1(>0)$ does not depend on δ, (ε, σ), and λ. Therefore, using (2.27) and (2.7), we obtain the following:

$$\begin{aligned} |B^{\varepsilon,\sigma,\lambda}(z^1, z^2)| \leqq {} & \frac{1}{\sigma} \|z_x^1\|_{L^2} \|z_x^2\|_{L^2} + \frac{c_1 \varepsilon}{\delta} \|z^1\|_{H^1} \|z^2\|_{H^1} \\ & + c_2 \|z^1\|_{L^2} \|z^2\|_{L^2} + |\lambda| \, \|z^1\|_{L^2} \|z^2\|_{L^2}, \end{aligned} \tag{2.28}$$

holds for $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_\delta, \sigma_0}$ and $\lambda \in \Lambda_{1,\delta}$, where $c_2(>0)$ is independent of $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$ and $\lambda \in \Lambda_1$. Here $\mathring{\Omega}$ denotes the interior of the set. The inequality (2.28) shows the boundedness of $B^{\varepsilon,\sigma,\lambda}$.

We first consider real noncritical eigenvalues. We shall show that there exists a positive constant μ^* such that $B^{\varepsilon,\sigma,\lambda}$ satisfies the positivity property

$$|B^{\varepsilon,\sigma,\lambda}(z, z)| \geqq c_3 \|z\|_{H^1}^2 \tag{2.29}$$

for $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_\delta, \sigma_0}$ and $\lambda > -\mu^*$ with $\lambda \in \Lambda_{1,\delta}$, where the positive constants c_3 and μ^* do not depend on (ε, σ) and λ. The inequality (2.29) implies from the Lax-Milgram theorem that the set $\{\lambda \in \mathbb{R} \mid \lambda \in \Lambda_{1,\delta} \text{ and } \lambda > -\mu^*\}$ belongs to the resolvent of (LP) for $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_d, \sigma_0}$. By virtue of Lemma 2.2, we see that

$$\langle (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} z), g_u^{\varepsilon,\sigma} z \rangle \xrightarrow[\varepsilon \downarrow 0]{} \left\langle \frac{-f_v^{*,\sigma} z}{f_u^{*,\sigma} - \lambda}, g_u^{*,\sigma} z \right\rangle \tag{2.30}$$

uniformly for $\|z\|_{H^1} \leqq M$, $0 \leqq \sigma < \sigma_0$, and $\lambda \in \Lambda_1$. Similarly, we have

$$\langle g_v^{\varepsilon,\sigma} z, z \rangle \xrightarrow[\varepsilon \downarrow 0]{} \langle g_v^{*,\sigma} z, z \rangle, \tag{2.31}$$

uniformly for $\|z\|_{H^1} \leqq M$ and $0 \leqq \sigma < \sigma_0$. Therefore, using the uniformity of the convergence of (2.30) and (2.31), and the estimate (2.27), we can see that if the limiting bilinear form $B^{*,\sigma,\lambda}$ defined by

$$B^{*,\sigma,\lambda}(z^1, z^2) = \frac{1}{\sigma} \langle z_x^1, z_x^2 \rangle - \left\langle \frac{-f_v^{*,\sigma} z^1}{f_u^{*,\sigma} - \lambda}, g_u^{*,\sigma} z^2 \right\rangle - \langle g_v^{*,\sigma} z^1, z^2 \rangle + \lambda \langle z^1, z^2 \rangle \tag{2.32}$$

satisfies the positivity property (2.29), then $B^{\varepsilon,\sigma,\lambda}$ also satisfy the same inequality with the appropriate change of constants c_3 and ε_δ.

After a simple computation, we obtain

$$B^{*,\sigma,\lambda}(z, z) = \frac{1}{\sigma}\|z_x\|^2_{L^2} + \left\langle \frac{1}{\lambda - f_u^{*,\sigma}}\{\lambda^2 - (f_u^{*,\sigma} + g_v^{*,\sigma})\lambda + \det{}^{*,\sigma}\}|z|^2, 1 \right\rangle. \tag{2.33}$$

Note that this corresponds to the reduced scalar eigenvalue problem (2.1). Recalling the assumptions (A.3)-(A.5), and $\lambda \in \Lambda_{1,\delta}$, we can see that there exist positive constants μ^* and c_4 such that

$$|\text{the second term of (2.33)}| \geqq c_4\|z\|_{L^2} \tag{2.34}$$

holds for $\lambda > -\mu^*$, where c_4 is independent of λ $(> -\mu^*)$ and σ with $0 \leqq \sigma < \sigma_0$. Combining the estimate (2.34) with (2.33), we easily see that $B^{*,\sigma,\lambda}$ satisfies the positivity property (2.29).

Next we consider the case where λ is a complex noncritical eigenvalue. Our goal is to show that there is a priori bound $-\mu^*$ such that $\operatorname{Re}\lambda < -\mu^*$ for small ε. First we note the following boundedness of complex noncritical eigenvalues.

SUBLEMMA 2.1. *If* $\lambda \in \Lambda_{1,\delta}$ *is a complex eigenvalue of* (LP) (*or* (2.5)), *then, we have*

$$|\lambda| \leqq M_c \quad \textit{for } (\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_0,\sigma_0},$$

where the positive constant M_c *does not depend on* (ε, σ).

Proof. Let z be the eigenfunction associated with λ. Then, the pair (λ, z) satisfies $B^{\varepsilon,\sigma,\lambda}(z, z) = 0$. Taking the Re- and Im- parts of this bilinear form, we have the following.

$$\begin{aligned} &\text{Re part: } \frac{1}{\sigma}\langle z_x, z_x\rangle - \operatorname{Re}\left\{\frac{1}{\zeta^{\varepsilon,\sigma} - \lambda}\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma}\rangle\langle g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}, z\rangle\right\} \\ &\qquad -\operatorname{Re}\langle (L^{\varepsilon,\sigma} - \lambda)^\dagger(-f_v^{\varepsilon,\sigma} z), g_u^{\varepsilon,\sigma} z\rangle \\ &\qquad -\langle g_v^{\varepsilon,\sigma} z, z\rangle + \operatorname{Re}\lambda\langle z, z\rangle = 0, \end{aligned} \tag{2.35}$$

$$\begin{aligned} &\text{Im part: } -\operatorname{Im}\left\{\frac{1}{\zeta_0^{\varepsilon,\sigma} - \lambda}\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma}\rangle\langle g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}, z\rangle \right. \\ &\qquad \left. + \langle (L^{\varepsilon,\sigma} - \lambda)^\dagger(-f_v^{\varepsilon,\sigma} z), g_u^{\varepsilon,\sigma} z\rangle\right\} + \operatorname{Im}\lambda\langle z, z\rangle = 0. \end{aligned} \tag{2.36}$$

Let us normalize z as $\|z\|_{L^2} = 1$. By virtue of (2.7), we can see from (2.36) that $\operatorname{Im}\lambda$ must be uniformly bounded for $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_0,\sigma_0}$. If not, the first term would remain bounded, while the second term is equal to $\operatorname{Im}\lambda$, which is a contradiction. As for $\operatorname{Re}\lambda$, it suffices to show that there exists a positive constant M_1 such that $\operatorname{Re}\lambda < M_1 < +\infty$, since $\lambda \in \Lambda_{1,\delta}$ (see (2.2) and (2.25)). We can prove this by applying a contradiction method to (2.35) again. In fact, suppose that there is a sequence of complex eigenvalues and their eigenfunctions (λ_n, z_n) with $(\varepsilon, \sigma) = (\varepsilon_n, \sigma_n) \in \mathring{\Omega}_{\varepsilon_0,\sigma_0}$ and $\|z_n\|_{L^2} = 1$ such that $\operatorname{Re}\lambda_n \uparrow +\infty$ as $n \uparrow +\infty$ as $n \uparrow \infty$. Then, in view of (2.35), we can see that $1/\sigma\langle (z_n)_x, (z_n)_x\rangle + \operatorname{Re}\lambda_n\langle z_n, z_n\rangle$ diverges to $+\infty$ as $n \uparrow \infty$, while all the other terms remain bounded for $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_0,\sigma_0}$, which violates (2.35). Thus, we obtain the conclusion of Sublemma 2.1.

COROLLARY 2.1. *Let* $\lambda \in \Lambda_{1,\delta}$ *be a complex eigenvalue of* (2.5), *and* z *be the associated eigenfunction with* $\|z\|_{L^2} = 1$. *Then, the following inequality holds*

$$1 \leqq \|z\|_{H^1} \leqq \sqrt{1 + \sigma M} \quad \textit{for } (\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_0,\sigma_0}, \tag{2.37}$$

where $M (> 0)$ *is independent of* ε *and* σ.

Proof. We know the uniform boundedness of $|\lambda|$, $\|(L^{\varepsilon,\sigma}-\lambda)^{\dagger}\|$, and partial derivatives $f_u^{\varepsilon,\sigma}, f_v^{\varepsilon,\sigma}$, etc., for $(\varepsilon,\sigma)\in\Omega_{\varepsilon_0,\sigma_0}$. Therefore, we easily obtain the following estimate from (2.35),

$$\frac{1}{\sigma}\|z_x\|^2_{L_2}\leqq \tilde{M}\|z\|^2_{L^2}\quad \text{for } (\varepsilon,\sigma)\in\mathring{\Omega}_{\varepsilon_0,\sigma_0},$$

where $\tilde{M}(>0)$ is independent of ε and σ, which leads to the conclusion.

Now, we shall show that there is a negative upper bound for the Re-parts of complex eigenvalues by using (2.35) and (2.36). Since z is normalized as $\|z\|_{L^2}=1$, $\|z\|_{H^1}\leqq M^*$, where M^* is independent of $(\varepsilon,\sigma)\in\mathring{\Omega}_{\varepsilon_0,\sigma_0}$ from Corollary 2.1.

First, we consider the limiting behavior of each term of (2.35). It follows from $\lambda\in\Lambda_{1,\delta}$, (1.28) and Corollary 1.3 that there exists a positive constant c_5 such that the inequality

$$\text{(2.38)}\qquad |\text{the second term of (2.35)}|\leqq\frac{c_5\varepsilon}{\delta}\|z\|^2_{H^1}\leqq\frac{c_5\varepsilon}{\delta}(M^*)^2$$

holds for $(\varepsilon,\sigma)\in\mathring{\Omega}_{\varepsilon_\delta,\sigma_0}$ and $\lambda\in\Lambda_{1,\delta}$, where c_5 does not depend on δ, (ε,σ) and λ. For the third term of (2.35), we can see from Lemma 2.2 that

$$\text{(2.39)}\qquad \{\text{the third term of (2.35)}\}\xrightarrow[\varepsilon\downarrow 0]{}\left\langle\frac{-f_u^{*,0}+\operatorname{Re}\lambda}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}(-f_v^{*,\sigma}z),\, g_u^{*,\sigma}z\right\rangle$$

uniformly for $\lambda\in\Lambda_{1,\delta}$, $0\leqq\sigma<\sigma_0$ and $\|z\|_{H^1}\leqq M^*$. Therefore, for small ε, we can rewrite (2.35) in the following form:

$$\begin{aligned}&\frac{1}{\sigma}\langle z_x, z_x\rangle+\left\langle\frac{-f_u^{*,\sigma}}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}(-f_v^{*,\sigma}z),\, g_u^{*,\sigma}z\right\rangle\\ &+\operatorname{Re}\lambda\left\langle\frac{-f_v^{*,\sigma}g_u^{*,\sigma}}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}|z|^2, 1\right\rangle-\langle g_v^{*,\sigma}|z|^2, 1\rangle\\ &+\operatorname{Re}\lambda+o(1)=0,\end{aligned}$$

where $o(1)$ denotes an infinitesimal term which goes to zero uniformly for $\lambda\in\Lambda_{1,\delta}$, $0\leqq\sigma<\sigma_0$ and $\|z\|_{H^1}\leqq M^*$. In the following, we use the same notation $o(1)$ for this type of infinitesimal term. Combining the second and the fourth terms of (2.40), we have

$$\text{(2.41)}\ 2\operatorname{Re}\lambda\left\langle\frac{f_u^{*,\sigma}g_v^{*,\sigma}}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}|z|^2, 1\right\rangle-\left\langle\frac{f_u^{*,\sigma}\det{}^{*,\sigma}+g_v^{*,\sigma}|\lambda|^2}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}|z|^2, 1\right\rangle.$$

After some computation, we obtain the following expression for $\operatorname{Re}\lambda$ from (2.40) together with (2.41),

$$\text{(2.42)}\qquad \operatorname{Re}\lambda=\frac{-\dfrac{1}{\sigma}\|z_x\|^2_{L^2}+\left\langle\dfrac{f_u^{*,\sigma}\det{}^{*,\sigma}+g_v^{*,\sigma}|\lambda|^2}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}|z|^2, 1\right\rangle+o(1)}{\left\langle\left(1+\dfrac{\det{}^{*,\sigma}+f_u^{*,\sigma}g_v^{*,\sigma}}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}\right)|z|^2, 1\right\rangle+o(1)}.$$

Recalling (A.3)-(A.5), $\lambda\in\Lambda_{1,\delta}(\subset\Lambda_1)$ and Sublemma 2.1, we can see that

$$\text{(2.43)}\quad 0<c_6<-\frac{f_u^{*,\sigma}\det{}^{*,\sigma}+g_v^{*,\sigma}|\lambda|^2}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2},\ \left(1+\frac{\det{}^{*,\sigma}+f_u^{*,\sigma}g_v^{*,\sigma}}{(f_u^{*,\sigma}-\operatorname{Re}\lambda)^2+(\operatorname{Im}\lambda)^2}\right)<c_7$$

holds, where the positive constants c_6 and c_7 are independent of δ, $\lambda \in \Lambda_1$ and σ. Since $\|z\|_{L^2}^2 = 1$, it follows from (2.42) and (2.43) that there exist positive constants μ^* and ε_δ such that

$$\operatorname{Re} \lambda < -\mu^* \quad \text{for } 0 < \varepsilon < \varepsilon_\delta,$$

where $-\mu^*$ does not depend on δ and $(\varepsilon, \sigma) \in \mathring{\Omega}_{\varepsilon_\delta, \sigma_0}$. Thus, we have obtained a negative upper bound for complex noncritical eigenvalues.

So far, we only consider the case where $\sigma > 0$. For the limiting case $\sigma = 0$, we have to study the following eigenvalue problem instead of (2.5),

$$\int_I \left\{ \frac{\langle -f_v^{\varepsilon,0}\eta, \phi_0^{\varepsilon,0} \rangle}{\zeta_0^{\varepsilon,0} - \lambda} g_u^{\varepsilon,0} \phi_0^{\varepsilon,0} + g_u^{\varepsilon,0}(L^{\varepsilon,0} - \lambda)^\dagger(-f_v^{\varepsilon,0}\eta) + g_v^{\varepsilon,0}\eta \right\} dx = \lambda \eta, \tag{2.44}$$

where η is a constant function. Note that Lemma 2.1 holds also for $\sigma = 0$. If we compare (2.44) with (2.26), we can see that (2.44) is a special form of (2.26) when we set z^i = constant function ($i = 1, 2$). Therefore, the above proof for $\sigma > 0$ is essentially valid to the case $\sigma = 0$ (even simpler than before), and the conclusion of Proposition 2.1 holds for $\sigma = 0$ without any change. In fact, if we remove $(1/\sigma)\|z_x\|_{L^2}^2$ from (2.33) (or (2.42)) and put $z = \eta$, we obtain the key expressions to show Proposition 2.1 for $\sigma = 0$. We leave the details to the reader (see also [19]).

Thus, we have finished the proof of Proposition 2.1.

Remark 2.4. Reconsidering the above proof, we find that in the framework of nonlinearities where

(AI) $\qquad f_v^{*,\sigma} g_u^{*,\sigma} < 0 \quad$ on the reduced set $R_+ \cup R_-$

as well as (A.0)-(A.4) are satisfied, Proposition 2.1 can be proved under a weaker assumption than (A.5). In fact, for real noncritical eigenvalues, we can see from (2.33) that

$$\sup_{x \in I} \{ f_u^{*,\sigma} + g_v^{*,\sigma} \} < 0 \quad \text{for } 0 < \sigma < \sigma_0 (< \sigma_1^*) \tag{2.45}$$

is enough even without (AI) to guarantee (2.34). For complex eigenvalues, using (2.36), we obtain from (2.40) that

$$\operatorname{Re} \lambda \leqq \frac{1}{2} \left\{ -\frac{1}{\sigma} \|z_x\|_{L^2}^2 + \left\langle \left(\sup_{x \in I} f_u^{*,\sigma} + \sup_{x \in I} g_v^{*,\sigma} \right) |z|^2, 1 \right\rangle \right\} + o(1).$$

Consequently, if we assume that

$$\sup_{x \in I} f_u^{*,\sigma} + \sup_{x \in I} g_v^{*,\sigma} < 0 \quad \text{for } 0 < \sigma < \sigma_0 (< \sigma_1^*). \tag{2.46}$$

$\operatorname{Re} \lambda$ has a negative upper bound for small ε and $0 < \sigma < \sigma_0$. Since (2.46) implies (2.45), under (AI), we can replace (A.5) by the weaker condition (2.46). However, the condition (2.46) depends on σ, in fact, when $\sigma_0 \uparrow \sigma_1^*$, we can see from Remark 0.2 that $\sup_{x \in I} f_u^{*,\sigma} \uparrow 0$. Therefore, if $\sup_{(u,v) \in R_+ \cup R_-} g_v(u, v)$ is strictly positive, the valid region of (2.46) is restricted to the subinterval of $[0, \sigma_1^*)$.

As far as noncritical eigenvalues are concerned, the effect of the second term

$$\frac{\langle -f_v^{\varepsilon,\sigma} z, \phi_0^{\varepsilon,\sigma} \rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda} g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma} \tag{2.47}$$

of (2.5) can be neglected, and therefore, we find that they are not dangerous to the stability of SPS1 as in Proposition 2.1. However, if $\lambda = \lambda(\varepsilon, \sigma)$ is a critical eigenvalue (i.e., $\lim_{\varepsilon \downarrow 0} \lambda(\varepsilon, \sigma) = 0$), then, the behavior of (2.47) becomes delicate in the sense that

(i) the denominator of (2.47) approaches to zero as $\varepsilon \downarrow 0$, and the infinitesimal order of $\zeta_0^{\varepsilon,\sigma}-\lambda$ is not a priori known;

(ii) $\phi_0^{\varepsilon,\sigma}$ does not have a limit function in L^2-sense as $\varepsilon \downarrow 0$.

A nice characterization of (2.47) as $\varepsilon \downarrow 0$ is certainly necessary to overcome these difficulties. We give first a heuristic argument how to know the asymptotic order of $\lambda(\varepsilon,\sigma)$ as $\varepsilon \downarrow 0$. We remind that $\phi_0^{\varepsilon,\sigma}$ decays with an exponential order as $\varepsilon \downarrow 0$ outside of any fixed neighborhood of the layer position $x_1^*(\sigma)$. Therefore, if the decaying order of $\zeta_0^{\varepsilon,\sigma}-\lambda$ is milder than it, we can see from (2.5) that the outer part of the limit eigenfunction is governed by (2.1). Namely, letting $z(\varepsilon,\sigma)$ be the corresponding eigenfunction to $\lambda(\varepsilon,\sigma)$, then the limit function $z^*=\lim_{\varepsilon\downarrow 0} z(\varepsilon,\sigma)$ belongs to $H_N^1(I)$ and satisfies (2.1) with $\lambda=0$ in each of the subintervals $[0,x_1^*(\sigma))$ and $(x_1^*(\sigma),1]$ in classical sense. (Note that all coefficients of (LP) have a jump discontinuity at only one point $x=x_1^*(\sigma)$ as $\varepsilon\downarrow 0$.) Then, z^* must be strictly convex in each subinterval due to the negativity of $\det^{*,\sigma}/f_u^{*,\sigma}$. Therefore, taking account of $z_x^*=0$ on ∂I, z_x^* cannot be continuous at $x=x_1^*(\sigma)$, and has a finite jump discontinuity. On the other hand, if we integrate (2.5) over a small neighbourhood $I_\kappa=(x_1^*(\sigma)-\kappa,\ x_1^*(\sigma)+\kappa)$ of the layer position $x_1^*(\sigma)$, we obtain

$$\frac{1}{\sigma}\{z_x(x_1^*(\sigma)+\kappa)-z_x(x_1^*(\sigma)-\kappa)\}+\frac{\langle -f_v^{\varepsilon,\sigma}z,\ \phi_0^{\varepsilon,\sigma}\rangle}{\zeta_0^{\varepsilon,\sigma}-\lambda}\int_{I_\kappa} g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}\,dx+O(\kappa)=0.$$

Consequently, we have

$$\text{(2.48)}\qquad \frac{1}{\sigma}[z_x^*]=\lim_{\kappa\downarrow 0}\lim_{\varepsilon\downarrow 0}-\{\text{integral of (2.47) over } I_\kappa\},$$

where $[\cdot]$ denotes the jump at $x=x_1^*(\sigma)$. Recalling Corollary 1.3 and (1.28), we can see that $\lambda(\varepsilon,\sigma)-\zeta_0^{\varepsilon,\sigma}$ must be $O(\varepsilon)$ as $\varepsilon\downarrow 0$, since the left-hand side of (2.48) is finite. Thus, we may take the asymptotic form of $\lambda(\varepsilon,\sigma)$ as

$$\text{(2.49)}\qquad \lambda(\varepsilon,\sigma)=\varepsilon\tau(\varepsilon,\sigma)\quad\text{with } \tau^{*,\sigma}\neq\sigma\hat{\zeta}_0^{*,\sigma},$$

where $\tau(\varepsilon,\sigma)$ is a continuous function of ε and σ, $\tau^{*,\sigma}=\tau(0,\sigma)$, and $\hat{\zeta}_0^{*,\sigma}$ is defined by (1.30). Using (1.28), we can rewrite (2.47) in the following form

$$\text{(2.50)}\qquad \frac{\langle -f_v^{\varepsilon,\sigma}z,\ \phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}\rangle}{\sigma\hat{\zeta}_0(\varepsilon,\sigma)+\mathrm{E\hat{x}p}\,(\varepsilon,\sigma)-\tau(\varepsilon,\sigma)}g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon},$$

where $\mathrm{E\hat{x}p}\,(\varepsilon,\sigma)=\mathrm{Exp}\,(\varepsilon,\sigma)/\varepsilon$ which still goes to zero with exponential order like (1.31). The following lemma gives a complete characterization of (2.50).

LEMMA 2.3 (*the second key lemma*).

$$\text{(a)}\qquad \frac{-f_v^{\varepsilon,\sigma}}{\sqrt{\varepsilon}}\phi_0^{\varepsilon,\sigma}\xrightarrow[\varepsilon\downarrow 0]{} c_1^*\delta^*\quad \textit{in } H^{-1}(I)\textit{-sense},$$

$$\text{(b)}\qquad \frac{g_u^{\varepsilon,\sigma}}{\sqrt{\varepsilon}}\phi_0^{\varepsilon,\sigma}\xrightarrow[\varepsilon\downarrow 0]{} c_2^*\delta^*\quad \textit{in } H^{-1}(I)\textit{-sense},$$

uniformly for $0\leqq\sigma<\sigma_0$, *where* $\delta^*=\delta(x-x_1^*(\sigma))$, *a Dirac's* δ*-function at* $x_1^*(\sigma)$, *and*

$$c_1^*=-\kappa^*\frac{d}{dv}J(v)|_{v=v^*}>0\quad (\textit{see}\ (\mathrm{A}.2)),$$

$$c_2^*=\kappa^*\{g(h_+(v^*),v^*)-g(h_-(v^*),v^*)\}>0\quad (\textit{see}\ (\mathrm{A}.4\mathrm{a})),$$

with

$$\kappa^{*-1} = \left\| \frac{d}{dy} \tilde{u}^*(y) \right\|_{L^2(\mathbb{R})}.$$

Remark 2.5. Note that each coefficient c_i^* $(i=1,2)$ of δ^* does not depend on σ.

Proof. In order to show that $\{-f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}\}_{\varepsilon>0}$ is a Dirac-sequence to $c_1^*\delta^*$, it suffices to prove the following. For an arbitrary interval $(a,b)\subset I$, it holds that

$$\lim_{\varepsilon\downarrow 0} \int_a^b -f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}\, dx = \begin{cases} 0 & \text{if } x_1^*(\sigma)\notin(a,b), \\ c_1^* & \text{if } x_1^*(\sigma)\in(a,b). \end{cases} \tag{2.51}$$

First, we consider the case where $x_1^*(\sigma)\notin(a,b)$. Using the stretched variable $y=(x-x_1(\varepsilon,\sigma))/\varepsilon$, we have

$$\int_a^b -f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}\, dx = \int_{(a-x_1(\varepsilon,\sigma))/\varepsilon}^{(b-x_1(\varepsilon,\sigma))/\varepsilon} \sqrt{\varepsilon}(-\tilde{f}_v^{\varepsilon,\sigma})\tilde{\phi}_0^{\varepsilon,\sigma}\, dy. \tag{2.52}$$

We can assume without loss of generality that $x^*(\sigma)<a<b\leqq 1$. The right-hand side is majorized as

$$|\text{the right-hand side of (2.52)}| \leqq \|-\tilde{f}_v^{\varepsilon,\sigma}\|_{L^\infty} \int_{(a-x_1(\varepsilon,\sigma))/\varepsilon}^{(b-x_1(\varepsilon,\sigma))/\varepsilon} \sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}\, dy. \tag{2.53}$$

Since $|x_1(\varepsilon,\sigma)-x_1^*(\sigma)|=o(1)$ as $\varepsilon\downarrow 0$ (see (16) and (17) in Appendix 1), we see that

$$(a-x_1(\varepsilon,\sigma))/\varepsilon \geqq \delta/\varepsilon \quad \text{for small } \varepsilon,$$

where δ is a positive constant independent of ε. Consequently, the right-hand side of (2.53) goes to zero by virtue of the exponentially decaying property of $\sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}(=\hat{\phi}_0^{\varepsilon,\sigma}$, see (1.24)), which completes the proof of the first part of (2.51).

Next, we consider the case where $x_1^*(\sigma)\in(a,b)$. In view of Corollary 1.1, Lemma 1.1 and Lemma 1.3, we have

$$\begin{aligned} -\tilde{f}_v^{\varepsilon,\sigma} &\xrightarrow[\varepsilon\downarrow 0]{} -\tilde{f}_v^{*,\sigma} \quad \text{in } C^2_{\text{c.u.}}\text{-sense} \quad \text{and} \\ \sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma} &\xrightarrow[\varepsilon\downarrow 0]{} \kappa^*\frac{d}{dy}\tilde{u}^* \quad \text{in } C^2_{\text{c.u.}}\text{-sense.} \end{aligned} \tag{2.54}$$

Using the decaying property of $\sqrt{\varepsilon}\tilde{\phi}_0^{\varepsilon,\sigma}$again, we can see that for any $\gamma>0$, there exist positive constants M_γ and ε_γ such that

$$\left| \int_{|y|\geqq M_\gamma} \sqrt{\varepsilon}(-\tilde{f}_v^{\varepsilon,\sigma})\tilde{\phi}_0^{\varepsilon,\sigma}\, dy \right| < \frac{\gamma}{3},$$

$$\left| \int_{|y|\geqq M_\gamma} \kappa^*(-\tilde{f}_v^{*,\sigma})\frac{d}{dy}\tilde{u}^*\, dy \right| < \frac{\gamma}{3},$$

and from (2.54),

$$\left| \int_{|y|\leqq M_\gamma} \left\{ \sqrt{\varepsilon}(-\tilde{f}_v^{\varepsilon,\sigma})\tilde{\phi}_0^{\varepsilon,\sigma} - \kappa^*(-\tilde{f}_v^{*,\sigma})\frac{d}{dy}\tilde{u}^* \right\} dy \right| < \frac{\gamma}{3},$$

for $0<\varepsilon<\varepsilon_\gamma$. Thus, we obtain

$$\left| \int_{\tilde{I}} \sqrt{\varepsilon}(-\tilde{f}_v^{\varepsilon,\sigma})\tilde{\phi}_0^{\varepsilon,\sigma}\, dy - \int_{-\infty}^{+\infty} \kappa^*(-\tilde{f}_v^{*,\sigma})\frac{d}{dy}\tilde{u}^*\, dy \right| < \gamma \quad \text{for } 0<\varepsilon<\varepsilon_\gamma.$$

Since γ is arbitrary, we have

$$\begin{aligned}\lim_{\varepsilon\downarrow 0}\int_{\tilde{I}}\sqrt{\varepsilon}(-\tilde{f}_v^{\varepsilon,\sigma})\tilde{\phi}_0^{\varepsilon,\sigma}\,dy &= \int_{-\infty}^{+\infty}\kappa^*(-\tilde{f}_v^{*,\sigma})\frac{d}{dy}\tilde{u}^*\,dy\\ &= -\kappa^*\int_{h_-(v^*)}^{h_+(v^*)} f_v(s, v^*)\,ds\\ &= -\kappa^*\frac{d}{dv}J(v)\bigg|_{v=v^*}\\ &= c_1^*.\end{aligned}$$

Here we use the fact that $\tilde{u}^*$ is a strictly monotone increasing function. Consequently, we have proved (a) of Lemma 2.3. As for (b), the same argument works without any change except the constant c_2^*. The constant c_2^* is computed as follows:

$$\begin{aligned}c_2^* &= \lim_{\varepsilon\downarrow 0}\int_{\tilde{I}}\sqrt{\varepsilon}\,\tilde{g}_u^{\varepsilon,\sigma}\tilde{\phi}_0^{\varepsilon,\sigma}\,dy\\ &= \int_{-\infty}^{+\infty}\kappa^*\tilde{g}_u^{*,\sigma}\frac{d}{dy}\tilde{u}^*\,dy\\ &= \kappa^*\int_{h_-(v^*)}^{h_+(v^*)}\frac{d}{ds}g(s, v^*)\,ds\\ &= \kappa^*\{g(h_+(v^*), v^*) - g(h_-(v^*), v^*)\}.\end{aligned}$$

Finally, noting that all the above estimates do not depend on σ for $0 \leqq \sigma < \sigma_0$, the convergence results (a) and (b) are also uniform for $0 \leqq \sigma < \sigma_0$, which completes the proof of Lemma 2.3.

Remark 2.6. Lemma 2.3 also holds for (SL) in Remark 1.4.

Using Lemma 2.3, we have

$$(2.55)\qquad (2.50) \xrightarrow[\varepsilon\downarrow 0]{} \frac{c_1^* c_2^*\langle z^*, \delta^*\rangle}{\sigma\hat{\zeta}_0^{*,\sigma} - \tau^{*,\sigma}}\delta^* \quad \text{in } H^{-1}(I)\text{-sense.}$$

This characterization is a key to solve the problem $(\mathrm{P2})_{\mathrm{b}}$ at the beginning of this section. Our task is to determine the sign of $\tau^{*,\sigma}$. In the following, using (2.55), we shall derive a limiting eigenvalue problem of (2.5) as $\varepsilon \downarrow 0$, which we call the *singular limit eigenvalue problem* (SLEP), and show that $\tau^{*,\sigma}$ is uniquely determined to be a negative constant. The assumption (2.49) and SLEP will be justified in the next section as well as the uniquenesss and the simplicity of the critical eigenvalue. Since the convergence in (2.55) has a meaning in H^{-1}-sense, we rewrite (2.5) in a weak form:

$$(2.56)\qquad \begin{aligned}&-\frac{1}{\sigma}\langle z_x, \psi_x\rangle + \frac{\langle z, -f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda}\langle g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}, \psi\rangle\\ &\qquad + \langle g_u^{\varepsilon,\sigma}(L^{\varepsilon,\sigma} - \lambda)^{\dagger}(-f_v^{\varepsilon,\sigma}z), \psi\rangle + \langle g_v^{\varepsilon,\sigma}z, \psi\rangle = \lambda\langle z, \psi\rangle,\\ &\qquad\qquad z \in H_N^1(I), \quad \forall\psi \in H^1(I).\end{aligned}$$

Using (2.55), Lemma 2.2 and $\lim_{\varepsilon\downarrow 0}\lambda(\varepsilon, \sigma) = 0$, we see that $z^* = \lim_{\varepsilon\downarrow 0} z(\varepsilon, \sigma)$ satisfy the following limit equation as $\varepsilon \downarrow 0$.

$$(\mathrm{SLEP})_{\mathrm{a}}\quad -\frac{1}{\sigma}\langle z_x^*, \psi_x\rangle + \frac{c_1^* c_2^*\langle z^*, \delta^*\rangle}{\sigma\hat{\zeta}_0^{*,\sigma} - \tau^{*,\sigma}}\langle\delta^*, \psi\rangle + \left\langle\frac{\det^{*,\sigma}}{f_u^{*,\sigma}}z^*, \psi\right\rangle = 0, \qquad z^* \in H_N^1(I),$$

for any $\psi \in H^1(I)$. Roughly speaking, the second term comes from the transition layer part, and the remaining ones represent the outer part of (2.5) as $\varepsilon \downarrow 0$, namely (2.1) with $\lambda = 0$. We call the above limit equation the *singular limit eigenvalue problem* (SLEP). Hereafter, we normalize the limit eigenfunction z^* as

$$\langle z^*, \delta^* \rangle = 1. \tag{2.57}$$

This normalization is always possible, since there are no nontrivial solutions of SLEP which satisfy $\langle z^*, \delta^* \rangle = 0$. We denote the solution of SLEP under (2.57) by the pair $(z^*, \tau^{*,\sigma}) \in H^1_N(I) \times \mathbb{R}$. It is easily seen that $(\text{SLEP})_a$ is equivalent to the following equations.

$$(\text{SLEP-1})_b \quad \begin{cases} \dfrac{1}{\sigma} z^*_{xx} + \dfrac{\det^{*,\sigma}}{f_u^{*,\sigma}} z^* = 0 & \text{in } [0, x_1^*(\sigma)) \cup (x_1^*(\sigma), 1], \\ z^*_x = 0 & \text{on } \partial I, \text{ and } z^* \text{ is continuous at } x = x_1^*(\sigma), \end{cases}$$

$$(\text{SLEP-2})_b \quad \frac{1}{\sigma}[z^*_x] = -\frac{c_1^* c_2^*}{\sigma \hat{\xi}_0^{*,\sigma} - \tau^{*,\sigma}},$$

where $[z^*_x]$ denotes the jump of z^*_x at $x = x_1^*(\sigma)$, namely,

$$[z^*_x] = \lim_{\delta \downarrow 0} \{ z^*_x(x_1^*(\sigma) + \delta) - z^*_x(x_1^*(\sigma) - \delta) \}.$$

Since $\det^{*,\sigma}/f_u^{*,\sigma}$ is strictly negative from (A.3) and (*A*.4) and smooth in each subinterval, there exists a unique solution of $(\text{SLEP-1})_b$ under (2.57), which is strictly convex and smooth in each subinterval (see Fig. 10). We denote this solution by $z_N^{*,\sigma}$. Using the relation $(\text{SLEP-2})_b$, $\tau^{*,\sigma}$ is also uniquely determined from $z_N^{*,\sigma}$, which we denote by $\tau_N^{*,\sigma}$. Thus, we obtain a unique solution $(z_N^{*,\sigma}, \tau_N^{*,\sigma})$ of $(\text{SLEP})_b$ under (2.57).

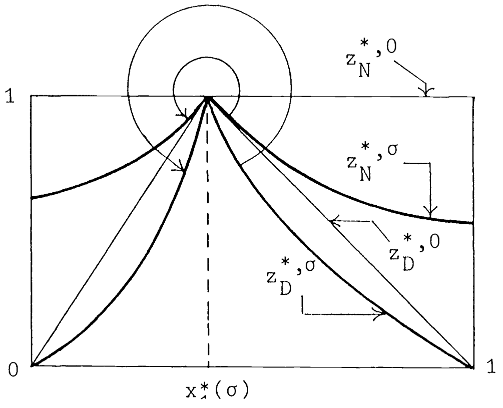


FIG. 10. *Solutions of* SLEP *and a comparison of jumps of* $(z_N^{*,\sigma})_x$ *and* $(z_D^{*,\sigma})_x$ *at* $x = x_1^*(\sigma)$.

In order to judge the sign of $\tau_N^{*,\sigma}$, the following observation is useful. Let $(\text{SLEP})_D$ denote the problem $(\text{SLEP})_b$ with replacing the boundary conditions $z^*_x = 0$ on ∂I by Dirichlet conditions $z^* = 0$ on ∂I. Then, we have the following.

LEMMA 2.4. $(\text{SLEP})_D$ *has a unique solution* $z_D^{*,\sigma}$ *with* $\tau^{*,\sigma} = 0$ *under* (2.57) (*see Fig.* 10).

Proof. By differentiating the SPS1 of the stationary problem (SP) with respect to x, we can see that $\lambda = 0$ is always an eigenvalue of (LP) under Dirichlet boundary conditions. In view of Remarks 1.5, 2.3 and 2.6, we see that all the above procedures to SLEP can be similarly done for Dirichlet boundary conditions, so that we obtain $(\mathrm{SLEP})_{\mathrm{D}}$. Therefore, $V_x^{*,\sigma}=\lim_{\varepsilon\downarrow 0} v(x, \varepsilon, \sigma)$ (see Theorem 1.1) becomes a solution of $(\mathrm{SLEP})_{\mathrm{D}}$ with $\tau^{*,\sigma}=0$, and from a normalization (2.57), $z_D^{*,\sigma}$ is given by

$$z_D^{*,\sigma}=\frac{V_x^{*,\sigma}}{V_x^{*,\sigma}(x_1^*(\sigma))}. \tag{2.58}$$

Now, a comparison of two solutions $(z_N^{*,\sigma}, \tau_N^{*,\sigma})$ and $(z_D^{*,\sigma}, 0)$ leads to the following (see Fig. 10).

LEMMA 2.5.

$$[(z_D^{*,\sigma})_x]<[(z_N^{*,\sigma})_x]<0. \tag{2.59}$$

Proof. By virtue of the negative definiteness of $\det^{*,\sigma}/f_u^{*,\sigma}$ and the positivity of $z_N^{*,\sigma}$ and $z_D^{*,\sigma}$, a simple comparison argument implies that

$$D_x^+ z_N^{*,\sigma}> D_x^+ z_D^{*,\sigma} \quad \text{and} \quad D_x^- z_N^{*,\sigma}< D_x^- z_D^{*,\sigma} \quad \text{at } x=x_1^*(\sigma),$$

where $D_x^+(D_x^-)$ denotes the right(left) x-derivative, which concludes (2.59).

In view of $(\mathrm{SLEP\text{-}2})_{\mathrm{b}}$, we can see from Lemma 2.5 that

$$\tau_N^{*,\sigma}<0 \quad \text{for } 0<\sigma<\sigma_0,$$

which shows the stability of SPS1. It is easy to see that $\tau_N^{*,\sigma}$ is a continuous function of σ for $0<\sigma<\sigma_0$, since $\det^{*,\sigma}/f_u^{*,\sigma}$ depends smoothly on σ in each subinterval and from (1.30).

Let us consider the limiting behavior of $\tau_N^{*,\sigma}$ as $\sigma\downarrow 0$. It is convenient to write $(\mathrm{SLEP})_{\mathrm{b}}$ in the following form.

$$z_{xx}^*+\sigma\frac{\det^{*,\sigma}}{f_u^{*,\sigma}}z^*=0 \quad \text{in each subinterval and } \langle z^*, \delta^*\rangle=1, \tag{2.60}$$

$$z_x^* \text{ (or } z^*)=0 \quad \text{on } \partial I \text{ and } z^* \text{ is continuous at } x=x_1^*(\sigma), \tag{2.61}$$

$$[z_x^*]=-\frac{\sigma c_1^* c_2^*}{\sigma\hat{\zeta}_0^{*,\sigma}-\tau^{*,\sigma}}. \tag{2.62}$$

We can see from the negativity of $\det^{*,\sigma}/f_u^{*,\sigma}$ and the boundary conditions that $z_N^{*,\sigma}$ and $z_D^{*,\sigma}$ satisfy $0\leqq z_N^{*,\sigma}, z_D^{*,\sigma}\leqq 1$. Therefore, noting that $|\det^{*,\sigma}/f_u^{*,\sigma}|<M$ for $0<\sigma<\sigma_0$, where M is independent of σ, we easily see from (2.60) that

$$(z_N^{*,\sigma})_{xx} \text{ and } (z_D^{*,\sigma})_{xx} \xrightarrow[\sigma\downarrow 0]{} 0 \quad \text{uniformly in each subinterval,}$$

$$(z_N^{*,\sigma})_x \xrightarrow[\sigma\downarrow 0]{} 0 \quad \text{uniformly in each subinterval.}$$

Using the boundary conditions, we can derive the following convergence:

$$z_N^{*,\sigma} \xrightarrow[\sigma\downarrow 0]{} 1 \quad \text{uniformly in } C^2\text{-sense in each subinterval,} \tag*{$(2.63)_{\mathrm{N}}$}$$

$$z_D^{*,\sigma} \xrightarrow[\sigma\downarrow 0]{} z_D^{*,0} \quad \text{uniformly in } C^2\text{-sense in each subinterval,} \tag*{$(2.63)_{\mathrm{D}}$}$$

where $z_D^{*,0}$ is a piecewise linear function as is shown in Fig. 10. Now, if we put $\psi=1$

in $(\mathrm{SLEP})_{\mathrm{a}}$, we obtain

$$\frac{c_1^* c_2^*}{\sigma\hat{\zeta}_0^{*,\sigma}-\tau_N^{*,\sigma}}+\left\langle\frac{\det^{*,\sigma}}{f_u^{*,\sigma}}z_N^{*,\sigma}, 1\right\rangle=0. \tag{2.64}$$

Therefore, using $(2.63)_{\mathrm{N}}$, we can see that $\tau_N^{*,\sigma}$ converges to $\tau_N^{*,0}$ as $\sigma\downarrow 0$, which is determined by the relation

$$-\frac{c_1^* c_2^*}{\tau_N^{*,0}}+\left\langle\frac{\det^{*,0}}{f_u^{*,0}}, 1\right\rangle=0. \tag{2.65}$$

On the other hand, starting from (2.44), we can derive the SLEP for the limiting case $\sigma=0$. It turns out after some computation that the SLEP for $\sigma=0$ is exactly the same as (2.65) (see also Nishiura and Fujii [19]). Thus, $\tau_N^{*,0}$ is a continuous function of σ up to $\sigma=0$, and $\tau_N^{*,0}$ is explicitly given by

$$\tau_N^{*,0}=\frac{c_1^* c_2^*}{\langle(\det^{*,0}/f_u^{*,0}), 1\rangle}. \tag{2.66}$$

We summarize the above results in the following theorem.

THEOREM 2.1. *The singular limit eigenvalue problem* (SLEP) *has a unique solution* $(z_N^{*,\sigma}, \tau_N^{*,\sigma})\in H_N^1(I)\times\mathbb{R}$ *under a normalization* (2.57) *such that*

(i) $z_N^{*,\sigma}$ *is smooth, strictly positive and convex in each of the subintervals* $[0, x_1^*(\sigma)]$ *and* $[x_1^*(\sigma), 1]$ (*see Fig.* 10),

(ii) $\tau_N^{*,\sigma}$ *is a continuous function of* σ, *and strictly negative for* $0\leqq\sigma<\sigma_0$. *Moreover,* $\tau_N^{*,0}=\lim_{\sigma\downarrow 0}\tau_N^{*,\sigma}$ *is given by* (2.66).

3. Justification of SLEP and the uniqueness of critical eigenvalue. In this section, we give a justification of SLEP and show the uniqueness and simplicity of the critical eigenvalue.

THEOREM 3.1. *There exists a positive constant* δ *such that* (2.5) *has a unique critical eigenvalue* $\lambda=\lambda_c(\varepsilon, \sigma)$ *in* B_δ *for small* ε. $\lambda_c(\varepsilon, \sigma)$ *is real, simple and takes the form* $\lambda_c(\varepsilon, \sigma)=\varepsilon\tau(\varepsilon, \sigma)$, *where* $\tau(\varepsilon, \sigma)$ *is a strictly negative continuous function for* $(\varepsilon, \sigma)\in\bar{\Omega}_{\varepsilon_0,\sigma_0}$. *Moreover,* $\tau(0, \sigma)$ *is just equal to* $\tau_N^{*,\sigma}$ *determined by* SLEP *in Theorem* 2.1.

This result combined with Proposition 2.1 guarantees that the negativity of $\tau_N^{*,\sigma}$ in Theorem 2.1 implies the stability of SPS1 for small $\varepsilon>0$.

The strategy to show Theorem 3.1 is to reduce the problem of finding critical eigenvalues of (2.5) to solve the algebra-like equation for λ by applying the inverse operator $K^{\varepsilon,\sigma,\lambda}$ (see Lemma 3.1) to (2.5).

Recalling Lemma 2.3, it is convenient to consider (2.5) in a weak form (2.56). First, we will define the operator $K^{\varepsilon,\sigma,\lambda}$, which is, roughly speaking, equal to

$$\left(-\frac{1}{\sigma}\frac{d^2}{dx^2}-g_u^{\varepsilon,\sigma}(L^{\varepsilon,\sigma}-\lambda)^{\dagger}(-f_v^{\varepsilon,\sigma}\cdot)-g_v^{\varepsilon,\sigma}\cdot+\lambda\cdot\right)^{-1}.$$

Let us introduce the bilinear form $\hat{B}^{\varepsilon,\sigma,\lambda}$ by

$$\hat{B}^{\varepsilon,\sigma,\lambda}(z^1, z^2)=\frac{1}{\sigma}\langle z_x^1, z_x^2\rangle-\langle\{g_u^{\varepsilon,\sigma}(L^{\varepsilon,\sigma}-\lambda)^{\dagger}(-f_v^{\varepsilon,\sigma}\cdot)+g_v^{\varepsilon,\sigma}-\lambda\}z^1, z^2\rangle \tag{3.1}$$

for any $z^i\in H_N^1(I)$ $(i=1, 2)$, which is equal to $B^{\varepsilon,\sigma,\lambda}$ (see (2.26)) after removing the second term of $B^{\varepsilon.\sigma,\lambda}$. For a given $h\in H^{-1}(I)$, we consider the equation for $z\in H_N^1(I)$:

$$\hat{B}^{\varepsilon,\sigma,\lambda}(z, \psi)=\langle h, \psi\rangle \quad \text{for any } \psi\in H_N^1(I). \tag{3.2}$$

If z is uniquely determined, we can define the mapping $K^{\varepsilon,\sigma,\lambda}$ as

$$K^{\varepsilon,\sigma,\lambda}h = z; \qquad H^{-1}(I) \to H^1_N(I). \tag{3.3}$$

We have the following result for $K^{\varepsilon,\sigma,\lambda}$.

LEMMA 3.1. *For any $\sigma_0(<\sigma_1^*)$, there exist positive constants ε_0 and δ_0 such that $K^{\varepsilon,\sigma,\lambda}$ is a well-defined uniformly bounded mapping from $H^{-1}(I)$ to $H^1_N(I)$ for $0 \leqq \varepsilon < \varepsilon_0$, $0 \leqq \sigma < \sigma_0$, and $\lambda \in \mathring{B}_{\delta_0}$, and depends continuously on (ε, σ) and analytically on λ in operator norm sense, respectively. Moreover, $K^{\varepsilon,0,\lambda}$ is the operator which maps h $(\in H^{-1}(I))$ to the constant function c_z defined by*

$$c_z = -\langle h, 1\rangle / \langle \{g_u^{\varepsilon,0}(L^{\varepsilon,0}-\lambda)^\dagger(-f_v^{\varepsilon,0}\,\cdot) + g_v^{\varepsilon,0} - \lambda\}1, 1\rangle. \tag{3.4}$$

Proof. First we consider the case where σ is positive. Using the similar arguments to obtain (2.28) and (2.29) in § 2, we can see that, for a given σ_0 $(<\sigma_1^*)$, there exist positive constants ε_0 and δ_0 such that

$$|\hat{B}^{\varepsilon,\sigma,\lambda}(z^1, z^2)| \leqq \left(\frac{1}{\sigma} + c_2 + |\lambda|\right)\|z^1\|_{H^1}\|z^2\|_{H^1} \tag{3.5}$$

and

$$|\hat{B}^{\varepsilon,\sigma,\lambda}(z, z)| \geqq \hat{c}_3\|z\|^2_{H^1} \tag{3.6}$$

hold for $0 \leqq \varepsilon < \varepsilon_0$, $0 < \sigma < \sigma_0$ and $\lambda \in \mathring{B}_{\delta_0}$, where c_2 and $\hat{c}_3$ are positive constants which are independent of ε, σ and λ. Note that ε_0 can be taken uniformly for $\lambda \in \mathring{B}_{\delta_0}$, since $\hat{B}^{\varepsilon,\sigma,\lambda}$ lacks the second term of (2.26). Therefore, it follows from the Lax-Milgram theorem that, for a given $h \in H^{-1}(I)$, there exists a unique $z \in H^1_N(I)$ such that

$$\hat{B}^{\varepsilon,\sigma,\lambda}(z, \psi) = \langle h, \psi\rangle \quad \text{for any } \psi \in H^1_N(I). \tag{3.7}$$

Thus, the generalized inverse operator $K^{\varepsilon,\sigma,\lambda}$

$$z = K^{\varepsilon,\sigma,\lambda}h; \quad H^{-1}(I) \to H^1_N(I)$$

is well defined, and it holds from (3.6) and (3.7) that

$$\|K^{\varepsilon,\sigma,\lambda}\| \leqq \hat{c}_3^{-1} \quad \text{for } 0 \leqq \varepsilon < \varepsilon_0, \quad 0 < \sigma < \sigma_0 \quad \text{and} \quad \lambda \in \mathring{B}_{\delta_0}, \tag{3.8}$$

where $\|K^{\varepsilon,\sigma,\lambda}\|$ denotes the operator norm of $K^{\varepsilon,\sigma,\lambda}$, i.e., $K^{\varepsilon,\sigma,\lambda}$ is uniformly bounded in this parameter region.

Next, we consider the parametric dependence of $K^{\varepsilon,\sigma,\lambda}$ on ε, σ and λ. From the definition of $K^{\varepsilon,\sigma,\lambda}$, we have

$$\hat{B}^{\varepsilon,\sigma,\lambda}(K^{\varepsilon,\sigma,\lambda}h, \psi) = \langle h, \psi\rangle = \hat{B}^{\varepsilon',\sigma',\lambda'}(K^{\varepsilon',\sigma',\lambda'}h, \psi)$$

for any $h \in H^{-1}(I)$ and $\psi \in H^1_N(I)$. Using this, we have the following equality

$$\begin{aligned} &\hat{B}^{\varepsilon,\sigma,\lambda}(K^{\varepsilon,\sigma,\lambda}h, \psi) - \hat{B}^{\varepsilon,\sigma,\lambda}(K^{\varepsilon',\sigma',\lambda'}h, \psi) \\ &\quad = \hat{B}^{\varepsilon',\sigma',\lambda'}(K^{\varepsilon',\sigma',\lambda'}h, \psi) - \hat{B}^{\varepsilon,\sigma,\lambda}(K^{\varepsilon',\sigma',\lambda'}h, \psi). \end{aligned} \tag{3.9}$$

The left-hand side of (3.9) is equal to

$$\hat{B}^{\varepsilon,\sigma,\lambda}(\{K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'}\}h, \psi).$$

Let $\psi = (K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'})h$, then, using the inequality (3.6), we obtain

$$|\hat{B}^{\varepsilon,\sigma,\lambda}(\{K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'}\}h, \{K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'}\}h)| \geqq \hat{c}_3\|(K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'})h\|^2_{H^1}. \tag{3.10}$$

On the other hand, the right-hand side of (3.9) is estimated from above for $\psi = (K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'})h$ as follows:

$$
\begin{aligned}
&|\text{the right-hand side of } (3.9)|\\
&\quad \leqq \left[\left| \frac{1}{\sigma} - \frac{1}{\sigma'} \right| \| K^{\varepsilon',\sigma',\lambda'} h \|_{H^1} \right.\\
&\qquad \left. + \| \{ (g_u^{\varepsilon',\sigma'} (L^{\varepsilon',\sigma'} - \lambda')^{\dagger} (-f_v^{\varepsilon',\sigma'} \cdot) + g_v^{\varepsilon',\sigma'} - \lambda') - (\text{without } ') \} K^{\varepsilon',\sigma',\lambda'} h \|_{L^2} \right]\\
&\qquad \times \| (K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'}) h \|_{H^1}.
\end{aligned}
\tag{3.11}
$$

Combining (3.10) with (3.11), we obtain the following inequality:

$$
\begin{aligned}
\| (K^{\varepsilon,\sigma,\lambda} - K^{\varepsilon',\sigma',\lambda'}) h \|_{H^1} &\leqq \hat{c}_3^{-1} \left| \frac{1}{\sigma} - \frac{1}{\sigma'} \right| \| K^{\varepsilon',\sigma',\lambda'} h \|_{H^1}\\
&\quad + \| \{ (g_u^{\varepsilon',\sigma'} (L^{\varepsilon',\sigma'} - \lambda')^{\dagger} (-f_v^{\varepsilon',\sigma'} \cdot) + g_v^{\varepsilon',\sigma'} - \lambda') - (\text{without } ') \} K^{\varepsilon',\sigma',\lambda'} h \|_{L^2}.
\end{aligned}
\tag{3.12}
$$

Note that when $\|h\|_{H^{-1}} \leqq 1$, we have $\|K^{\varepsilon',\sigma',\lambda'} h\|_{H^1} \leqq \hat{c}_3^{-1}$ from (3.8). Recalling that $(L^{\varepsilon,\sigma} - \lambda)^{\dagger}(-f_v^{\varepsilon,\sigma} \cdot)$ converges to $(-f_v^{*,\sigma} \cdot)/(f_u^{*,\sigma} - \lambda)$ as $\varepsilon \downarrow 0$ uniformly on a bounded set in $H^1(I)$ (see Lemma 2.2), L^2-continuity of the coefficients with respect to parameters up to $\varepsilon = 0$, and the resolvent formula in Remark 2.2, we can see from (3.12) that $K^{\varepsilon,\sigma,\lambda}$ depends continuously on (ε, σ), and analytically on λ in operator norm sense for $0 \leqq \varepsilon < \varepsilon_0$, $0 < \sigma < \sigma_0$ and $\lambda \in B_{\delta_0}$, respectively.

Finally, we consider the limiting behavior of $K^{\varepsilon,\sigma,\lambda}$ as $\sigma \downarrow 0$. Although the boundedness of $\hat{B}^{\varepsilon,\sigma,\lambda}$ breaks down in this limit (see (3.5)), we can recover it by dividing the solution space into the average free space and the constant function space as follows,

$$
z = c_z + \hat{z} \quad \text{and} \quad h = c_h + \hat{h},
\tag{3.13}
$$

where $c_z = \langle z, 1 \rangle$, $c_h = \langle h, 1 \rangle$ and $\hat{z}$ and $\hat{h}$ are average free parts of z and h, respectively. Let $\hat{H}_N^1(I)$ denote the average free space, namely,

$$
\hat{H}_N^1(I) = \{ \hat{\psi} \in H_N^1(I) \,|\, \langle \hat{\psi}, 1 \rangle = 0 \}.
$$

Substituting (3.13) into (3.2), and multiplying σ on both sides, we obtain for $\psi = \hat{\psi} \in \hat{H}_N^1(I)$,

$$
\begin{aligned}
&\langle \hat{z}_x, \hat{\psi}_x \rangle - \sigma \langle \{ g_u^{\varepsilon,\sigma} (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} \cdot) + g_v^{\varepsilon,\sigma} + \lambda \} \hat{z}, \hat{\psi} \rangle\\
&\qquad = \sigma \langle \{ g_u^{\varepsilon,\sigma} (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} \cdot) + g_v^{\varepsilon,\sigma} + \lambda \} c_z + \hat{h}, \hat{\psi} \rangle.
\end{aligned}
\tag{3.14}
$$

Noting that $\langle \hat{\psi}_x, \hat{\psi}_x \rangle$ is equivalent to H^1-norm in $\hat{H}_N^1$-space, we see that there exist a $\hat{\sigma}(>0)$ such that the left side of (3.14) is a positive bounded bilinear form for $0 \leqq \sigma < \hat{\sigma}$. Thus, we can solve (3.14) with respect to $\hat{z}$ as a function of c_z and $\hat{h}$ as $\hat{z} = \hat{z}(c_z, \hat{h}) \in \hat{H}_N^1(I)$ with the aid of the Lax-Milgram theorem. Furthermore, $\hat{z}$ satisfies

$$
\| \hat{z} \|_{\hat{H}_N^1} \leqq C \sigma (c_z + \| \hat{h} \|_{H^{-1}}),
\tag{3.15}
$$

where C is a positive constant independent of ε and σ. On the other hand, if we set $\psi = 1$ in (3.2), we have

$$
-\langle \{ g_u^{\varepsilon,\sigma} (L^{\varepsilon,\sigma} - \lambda)^{\dagger} (-f_v^{\varepsilon,\sigma} \cdot) + g_v^{\varepsilon,\sigma} - \lambda \} (c_z + \hat{z}), 1 \rangle = \langle c_h, 1 \rangle.
\tag{3.16}
$$

Substituting $\hat{z} = \hat{z}(c_z, \hat{h})$ into (3.16), we have an equation for c_z. Note that the coefficient of c_z in the left-hand side of (3.16) is strictly negative by virtue of Lemma 2.2, (A.3) and (A.4b) for $0 \leqq \varepsilon < \varepsilon_0$, $0 \leqq \sigma < \sigma_0$, and $\lambda \in B_{\delta_0}$. Here, we take ε_0 and δ_0 smaller, if necessary. In view of (3.15), we can see that (3.16) is uniquely solvable with respect to c_z for small $\hat{\sigma}$ as

$$c_z = c_z^{\varepsilon,\sigma,\lambda}(h). \tag{3.17}$$

Thus, $z = K^{\varepsilon,\sigma,\lambda} h$ takes the following form

$$z = c_z^{\varepsilon,\sigma,\lambda}(h) + \hat{z}(c_z^{\varepsilon,\sigma,\lambda}(h), \hat{h}) \tag{3.18}$$

for $0 \leqq \varepsilon < \varepsilon_0$, $0 \leqq \sigma < \hat{\sigma}$ and $\lambda \in B_{\delta_0}$.

Integrating the results from (3.15) to (3.18), we can see that

$$K^{\varepsilon,\sigma,\lambda} \xrightarrow[\sigma \downarrow 0]{} K^{\varepsilon,0,\lambda} \quad \text{in operator norm}$$

uniformly for $0 \leqq \varepsilon < \varepsilon_0$ and $\lambda \in B_{\delta_0}$, where $K^{\varepsilon,0,\lambda}$ is the operator which maps $h \in H^{-1}(I)$ to the constant function c_z defined by

$$c_z = -\langle h, 1\rangle / \langle \{ g_u^{\varepsilon,0}(L^{\varepsilon,0} - \lambda)^{\dagger}(-f_v^{\varepsilon,0} \cdot) + g_v^{\varepsilon,0} - \lambda \} \cdot 1, 1 \rangle, \tag{3.19}$$

which completes the proof of Lemma 3.1.

Remark 3.1. When $h \in L^2(I)$, $K^{\varepsilon,\sigma,\lambda} h$ belongs to $H^2(I) \cap H_N^1(I)$. Moreover, for positive ε, when h is smooth, so is $K^{\varepsilon,\sigma,\lambda}\ h$.

Now, we will derive a scalar equation for λ, which is equivalent to (2.56). Applying the operator $K^{\varepsilon,\sigma,\lambda}$ in Lemma 3.1 to (2.56), and dividing both denominator and numerator by ε, we have

$$z = \frac{\langle z, -f_v^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma} / \sqrt{\varepsilon} \rangle}{(\zeta_0^{\varepsilon,\sigma} - \lambda)/\varepsilon} K^{\varepsilon,\sigma,\lambda}(g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma} / \sqrt{\varepsilon}), \tag{3.20}$$

which implies that z is a scalar multiple of $K^{\varepsilon,\sigma,\lambda}(g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon})$, namely, $z = \alpha K^{\varepsilon,\sigma,\lambda}(g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon})$ with α being a scalar constant. Substituting this into (3.20), we can see that (3.20) has a nontrivial solution z if and only if λ satisfies the following equation,

$$\left\langle K^{\varepsilon,\sigma,\lambda}\left(\frac{g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}}{\sqrt{\varepsilon}} \right), \frac{-f_v^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}}{\sqrt{\varepsilon}} \right\rangle = \frac{\zeta_0^{\varepsilon,\sigma} - \lambda}{\varepsilon}. \tag{3.21}$$

It follows from Lemma 2.3 and Lemma 3.1 that the left-hand side of (3.21) is continuous with respect to (ε, σ), and analytic with respect to λ for $0 \leqq \varepsilon < \varepsilon_0$, $0 \leqq \sigma < \sigma_0$ and $\lambda \in B_{\delta_0}$. Therefore, recalling the asymptotic form of $\zeta_0^{\varepsilon,\sigma}$ (see (1.28) in Lemma 1.4), λ must be $O(\varepsilon)$ in order that (3.21) has a solution $\lambda = \lambda_c(\varepsilon, \sigma)$ with $\lambda_c(0, \sigma) = 0$. Hence, without loss of generality, we can set

$$\lambda = \varepsilon \tau(\varepsilon, \sigma), \tag{3.22}$$

where τ is a bounded continuous function for $(\varepsilon, \sigma) \in \bar{\Omega}_{\varepsilon_0,\sigma_0}$. Substituting (3.22) into (3.21), it is easily seen that (3.21) is equivalent to the following scalar equation for τ:

$$\begin{aligned} \mathscr{F}(\varepsilon, \sigma, \tau) = \tau - \sigma \tilde{\zeta}_0(\varepsilon, \sigma) - \hat{\text{Exp}}\,(\varepsilon, \sigma) \\ + \left\langle K^{\varepsilon,\sigma,\varepsilon\tau}\left(\frac{g_u^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}}{\sqrt{\varepsilon}} \right), \frac{-f_v^{\varepsilon,\sigma} \phi_0^{\varepsilon,\sigma}}{\sqrt{\varepsilon}} \right\rangle = 0. \end{aligned} \tag{3.23}$$

When $\varepsilon = 0$, there exists a unique solution $\tau_N^{*,\sigma}$ of $\mathscr{F}(0, \sigma, \tau) = 0$ for any σ with $0 \leqq \sigma < \sigma_0$, namely

$$\tau_N^{*,\sigma} = \sigma \hat{\zeta}_0^{*,\sigma} - c_1^* c_2^* \langle K^{0,\sigma,0} \delta^*, \delta^* \rangle. \tag{3.24}$$

Moreover, in view of (3.23), we can see from the uniform continuity with respect to σ that, for any $\kappa_0 > 0$, there exists an $\varepsilon_{\kappa_0} > 0$ such that, if $(\varepsilon, \sigma, \tau)$ is a solution of (3.23) with $0 \leqq \varepsilon < \varepsilon_{\kappa_0}$ and $0 \leqq \sigma < \sigma_0$ then τ must belong to the κ_0-neighbourhood of $\tau_N^{*,\sigma}$ in the complex plane. Recalling (SLEP)$_a$ in § 2, (2.57) and the definition of $K^{\varepsilon,\sigma,\lambda}$, $\tau_N^{*,\sigma}$ defined by (3.24) is just equal to $\tau_N^{*,\sigma}$ given in Theorem 2.1. On the other hand, since $\mathscr{F}(0, \sigma, \tau_N^{*,\sigma}) = 0$ and $(\partial\mathscr{F}/\partial\tau)(0, \sigma, \tau_N^{*,\sigma}) = 1$, we can apply the implicit function theorem to (3.23) and obtain a unique solution $\tau = \tau_c(\varepsilon, \sigma)$ with $\tau_c(0, \sigma) = \tau_N^{*,\sigma}$ in the appropriate region $0 \leqq \varepsilon < \varepsilon^*$, $0 \leqq \sigma < \sigma_0$ and $|\tau - \tau_N^{*,\sigma}| < \kappa^*$, where ε^* and κ^* do not depend on σ. If we take κ_0 to be smaller than κ^* and ε_0 to be $\min\{\varepsilon_{\kappa_0}, \varepsilon^*\}$, we can see from the above arguments that $\tau = \tau_c(\varepsilon, \sigma)$ is a unique solution of (3.23) for $0 \leqq \varepsilon < \varepsilon_0$ and $0 \leqq \sigma < \sigma_0$, which implies that

$$\lambda = \lambda_c(\varepsilon, \sigma) \overset{\text{def}}{\equiv} \varepsilon\tau_c(\varepsilon, \sigma) \tag{3.25}$$

is a unique solution of (3.21) for $0 \leqq \varepsilon < \varepsilon_0$, $0 \leqq \sigma < \sigma_0$ and $\lambda \in B_\delta$, where $0 < \delta \leqq \varepsilon_0\kappa_0$. Noting that $\mathscr{F}$ is a real operator, it is clear from the uniqueness that $\tau_c(\varepsilon, \sigma)$ is real.

Finally, we show the simplicity of the critical eigenvalue (3.25). Apparently, the geometric multiplicity of $\lambda_c(\varepsilon, \sigma)$ is equal to one, since the z-component of the eigenvector ${}^t(w_c, z_c)$ is uniquely determined by $z_c = K^{\varepsilon,\sigma,\varepsilon\tau_c(\varepsilon,\sigma)}(g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma})$ up to a constant multiple, and so is the w-component through $w_c = (L^{\varepsilon,\sigma} - \lambda_c(\varepsilon, \sigma))^{-1}(-f_v^{\varepsilon,\sigma} z_c)$.

In order to show that the algebraic multiplicity is also equal to one, we introduce the adjoint operator of $\mathscr{L}^{\varepsilon,\sigma}$ (see (LP)) defined by

$$(\mathscr{L}^{\varepsilon,\sigma})^*\begin{pmatrix} w^* \\ z^* \end{pmatrix} = \begin{bmatrix} \varepsilon^2\dfrac{d^2}{dx^2} + f_u^{\varepsilon,\sigma} & g_u^{\varepsilon,\sigma} \\ f_v^{\varepsilon,\sigma} & \dfrac{1}{\sigma}\dfrac{d^2}{dx^2} + g_v^{\varepsilon,\sigma} \end{bmatrix}\begin{pmatrix} w^* \\ z^* \end{pmatrix} = \lambda^*\begin{pmatrix} w^* \\ z^* \end{pmatrix}, \tag{3.26}$$

with $\mathscr{D}((\mathscr{L}^{\varepsilon,\sigma})^*) = (H^2(I) \cap H_N^1(I))^2 (\subset (L^2(I))^2)$ into $(L^2(I))^2$. It is not difficult to verify that $(\mathscr{L}^{\varepsilon,\sigma})^*$ has the same isolated eigenvalue $\lambda^* = \lambda_c(\varepsilon, \sigma)$ with geometric multiplicity being equal to one. The associated eigenfunction is given by ${}^t(w_c^*, z_c^*) = {}^t((L^{\varepsilon,\sigma} - \lambda_c(\varepsilon, \sigma))^{-1}\{-g_u^{\varepsilon,\sigma}(K^{\varepsilon,\sigma,\lambda_c})^*\ (f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma})\}, (K^{\varepsilon,\sigma,\lambda_c})^*(f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}))$, where $(K^{\varepsilon,\sigma,\lambda_c})^*$ is the generalized inverse from $H^{-1}(I)$ to $H_N^1(I)$ analogously defined as $K^{\varepsilon,\sigma,\lambda}$. From the closed range theorem, we have $(\mathscr{N}(\mathscr{L}^{\varepsilon,\sigma} - \lambda_c)^*)^\perp = \mathscr{R}(\mathscr{L}^{\varepsilon,\sigma} - \lambda_c)$, where $\mathscr{N}$ and $\mathscr{R}$ denotes the nullspace and range, respectively. Therefore, in order to show the simplicity of $\lambda_c(\varepsilon, \sigma)$, it suffices to prove

$$\left\langle\!\!\left\langle \begin{pmatrix} w_c \\ z_c \end{pmatrix}, \begin{pmatrix} w_c^* \\ z_c^* \end{pmatrix} \right\rangle\!\!\right\rangle \neq 0, \tag{3.27}$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ denotes the inner product in $(L^2(I))^2$-space. Let us compute the left-hand side of (3.27). Using the eigenfunction expansion of $L^{\varepsilon,\sigma}$, we rewrite w_c as

$$\begin{aligned} w_c &= (L^{\varepsilon,\sigma} - \lambda_c)^{-1}\{-f_v^{\varepsilon,\sigma}K^{\varepsilon,\sigma,\lambda_c}(g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma})\} \\ &= \frac{\langle -f_v^{\varepsilon,\sigma}K^{\varepsilon,\sigma,\lambda_c}(g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}), \phi_0^{\varepsilon,\sigma}\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda_c}\phi_0^{\varepsilon,\sigma} + (L^{\varepsilon,\sigma} - \lambda_c)^\dagger\{\cdot\} \\ &= \frac{\langle K^{\varepsilon,\sigma,\lambda_c}(g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}), -f_v^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}/\sqrt{\varepsilon}\rangle}{\sigma\hat{\zeta}_0(\varepsilon, \sigma) - \tau_c}\phi_0^{\varepsilon,\sigma} + (L^{\varepsilon,\sigma} - \lambda_c)\dagger\{\cdot\}. \end{aligned} \tag{3.28}$$

Since $\|g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma}\|_{H^{-1}} = O(\sqrt{\varepsilon})$ from Corollary 1.3, it follows from Lemma 3.1 that $\|K^{\varepsilon,\sigma,\lambda_c}(g_u^{\varepsilon,\sigma}\phi_0^{\varepsilon,\sigma})\|_{H^1} = O(\sqrt{\varepsilon})$, which shows that the second term of (3.28) is of $O(\sqrt{\varepsilon})$ due to the L^2-uniform boundedness of $(L^{\varepsilon,\sigma} - \lambda_c)^\dagger$ (see (2.7)). On the other hand, the first term of (3.28) is just equal to $\phi_0^{\varepsilon,\sigma}$, since $\lambda_c = \varepsilon\tau_c$ is a solution of (3.21). Therefore, the principal part of w_c for small ε is given by $\phi_0^{\varepsilon,\sigma}$, namely,

$$w_c = \phi_0^{\varepsilon,\sigma} + O(\sqrt{\varepsilon}) \quad \text{as } \varepsilon \downarrow 0 \text{ in } L^2\text{-sense.} \tag{3.29}$$

Note that the above computation also shows that

$$z_c = O(\sqrt{\varepsilon}) \quad \text{as } \varepsilon \downarrow 0 \text{ in } L^2\text{-sense.} \tag{3.30}$$

Similarly, we have the analogous result for the adjoint eigenfunction;

$$\left.\begin{array}{l} w_c^* = \phi_0^{\varepsilon,\sigma} + O(\sqrt{\varepsilon}) \\ z_c^* = O(\sqrt{\varepsilon}) \end{array}\right. \quad \text{as } \varepsilon \downarrow 0 \text{ in } L^2\text{-sense.} \tag{3.31}$$

Thus, using (3.29)-(3.31), we can see that the inner product (3.27) becomes

$$\left\langle\!\!\left\langle \begin{pmatrix} w_c \\ z_c \end{pmatrix}, \begin{pmatrix} w_c^* \\ z_c^* \end{pmatrix} \right\rangle\!\!\right\rangle = 1 + O(\varepsilon) \quad \text{for small } \varepsilon,$$

which concludes the simplicity of $\lambda_c(\varepsilon, \sigma)$ for small ε and finishes the proof of Theorem 3.1.

4. Concluding remarks—instability theorems. As mentioned in the Introduction, concerning the stability of SPS, and in particular that of mode 1 patterns, it has been widely believed that they are stable so long as they can be constructed. Also, "numerical evidences leave little doubt that the pattern is quite robust" (according to Conway [3]).

So far, our study has been concentrated on the stability of SPS of mode 1, from the viewpoint that "under what conditions SPS1 are stable solutions?" Our Main Theorem says that the above *belief* is *partially* correct, that we needed (A.5) as the stability condition of SPS1, together with the existence conditions (A.1)-(A.4).

We have an important remark here. If the direction of the inequality in (A.5) is reversed, then an instability theorem follows for SPS of mode 1.

Another possibility of instability is that the sign of dJ/dv at v^* is reversed in (A.2). Of course, this means a violation of our ⟨Existence Conditions⟩. However, it is not difficult to see that (i) SPS can still be constructed under such a reversed condition on dJ/dv at least for small σ, and that (ii) there can exist a nonlinearity satisfying the reversed condition on dJ/dv, without destroying the other existence conditions. In fact, we can construct a function $f(u, v)$ by an appropriate deformation around the central branch $h_0(v)$ so that $J(v)$ has three zeros at $v = \underline{v}^*$, v^*, and $\bar{v}^*$ as in Fig. 11, and $dJ/dv > (<)0$ at $v = v^*$ ($\underline{v}^*$ and $\bar{v}^*$), respectively. We remark that the mathematical structures of these two instabilities are not the same, namely, *the noncritical instability* for the first case and *the critical case* for the second case. More precisely, the instability occurs in two different ways as follows. Hereafter, let us fix σ to be an appropriate positive constant.

(1) Instability of noncritical type. We shall show that there exists a *real positive eigenvalue* of (LP) for small ε, if g_v satisfies the following:

$$\begin{gathered} \min_{(u,v)\in R_+\cup R_-} \{f_u + g_v\} > 0, \\ \min_{(u,v)\in R_+\cup R_-} \{(f_u + g_v)^2 - 4(f_u g_v - f_v g_u)\} > 0, \\ \min_{x\in I} \bar{\lambda}(x) > \max_{x\in I} \underline{\lambda}(x), \end{gathered} \tag{4.1}$$

where $\bar{\lambda}(x)$ and $\underline{\lambda}(x)$ ($\underline{\lambda}(x) < \bar{\lambda}(x)$) are two real positive solutions of

$$\lambda^2 - \{f_u^{*,\sigma}(x) + g_v^{*,\sigma}(x)\}\lambda + \det{}^{*,\sigma}(x) = 0.$$

Recall that $f_u^{*,\sigma}(x) = f_u(U^{*,\sigma}(x), V^{*,\sigma}(x))$ and so on.

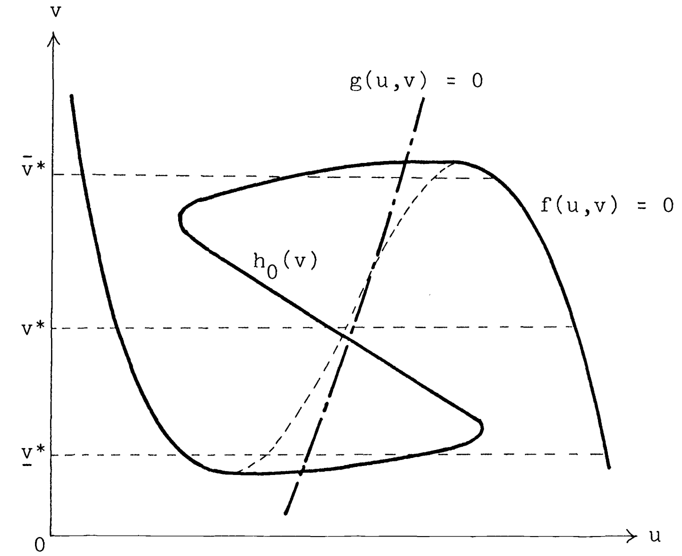


FIG. 11. *An example of the nonlinearity satisfying* $(dJ/dv)>0$ *at* $v=v^*$.

First, let us consider the following minimization problem for real λ associated with (2.33):

$$m(\sigma,\lambda)\overset{\text{def}}{=}\min_{z\in H^2(I)\cap H^1_N(I),\|z\|_{L^2}=1} B^{*,\sigma,\lambda}. \tag{4.2}$$

Suppose $m(\sigma,\lambda)$ is equal to zero and attained by $(z,\lambda)=(z^*,\lambda^*)$, then (z^*,λ^*) is a solution of the eigenvalue problem

$$\frac{1}{\sigma}z_{xx}-\frac{\det^{*,\sigma}-g_v^{*,\sigma}}{\lambda-f_u^{*,\sigma}}z=\lambda z. \tag{4.3}$$

Under the conditions (4.1), we can find a $\lambda_1>0$ such that

$$\frac{\lambda_1^2-(f_u^{*,\sigma}+g_v^{*,\sigma})\lambda_1+\det^{*,\sigma}}{\lambda_1-f_u^{*,\sigma}}<0 \quad \text{for any } x\in I,$$

while, for large $\lambda_2>0$, we have

$$\frac{\lambda_2^2-(f_u^{*,\sigma}+g_v^{*,\sigma})\lambda_2+\det^{*\sigma}}{\lambda_2-f_u^{*,\sigma}}>0 \quad \text{for any } x\in I.$$

Let $z=1$ in (4.2), we see that $m(\sigma,\lambda_1)<0$. On the other hand, it is clear that $m(\sigma,\lambda_2)>0$. Therefore, by using a continuity argument, it is easily seen that there exists a λ_0^* such that $m(\sigma,\lambda_0^*)=0$ with $0<\lambda_1<\lambda_0^*<\lambda_2$. This implies the existence of the real positive eigenvalue of (4.3). Applying a regular perturbation technique to (2.5) with the aid of $K^{\varepsilon,\sigma,\lambda}$ in Lemma 3.1, it is not difficult to obtain a real positive eigenvalue $\lambda=\lambda^*(\varepsilon,\sigma)$ with $\lambda^*(0,\sigma)=\lambda_0^*$ of (LP) for $\varepsilon>0$. We leave the details to the reader.

(2) Instability of critical type. If $dJ/dv(v^*)$ changes its sign from negative to positive, two important quantities also change signs. The first one is $\hat{\zeta}_0^{*,\sigma}$ (see (1.30)), which changes sign from positive to negative. The second one is the coefficient c_1^* of δ^* in Lemma 2.3, which becomes positive when $dJ/dv(v^*)>0$. In view of SLEP in § 2, we see that SLEP is invariant under the transformation from $(c_1^*, \hat{\zeta}_0^{*,\sigma}, \tau^{*,\sigma})$ to

$(-c_1^*, -\hat{\zeta}_0^{*,\sigma}, -\tau^{*,\sigma})$. Therefore, $\tau^{*,\sigma}$ takes a positive value when $dJ/dv(v^*) > 0$, which implies the existence of *the unstable real critical eigenvalue* which tends to zero from the positive side with $O(\varepsilon)$.

Thus, we have shown the following proposition.

PROPOSITION 4.1. *If the sign of* $g_v|_{R_+ \cup R_-}$ *(Stability Assumption* (A.5)) *or* $dJ/dv(v^*)$ ((A.2)) *is reversed,* SPS1 $U^{\varepsilon,\sigma}$ *becomes unstable in the following sense.*

(i) *If* $g_v > 0$ *on* $R_+ \cup R_-$ *and satisfies* (4.1), *then there exists a real eigenvalue* $\lambda^*(\varepsilon, \sigma)$, *which is strictly positive for small* ε.

(ii) *If* $dJ/dv(v^*) > 0$, *the unique critical eigenvalue* $\lambda_c(\varepsilon, \sigma)$, *which is real and simple, approaches to zero from the positive side with* $O(\varepsilon)$ *as* $\varepsilon \downarrow 0$, *namely,* $\lambda_c(\varepsilon, \sigma) \simeq \tau^{*,\sigma}\varepsilon(\tau^{*,\sigma} > 0)$.

Remark 4.1. For the limiting case $\sigma = 0$, the proof for the instability of critical type is also valid. While, for noncritical type, the same instability as in Proposition 4.1 occurs under weaker conditions than (4.1), namely,

$$\int_I (f_u^{*,0} + g_v^{*,0})\, dx > 0,$$

$$\left\{\int_I (f_u^{*,0} + g_v^{*,0})\, dx\right\}^2 - 4\int_I \det{}^{*,0}\, dx > 0.$$

This can be proved by making use of (2.44) in an analogous way.

Appendix 1. We show the outline of the construction of SPS1 and its parametric dependency in Theorem 1.1. The method here is essentially a modified version of Mimura, Tabata and Hosono [15]. However, this improves the original version at the three points:

(i) The σ-family of SPS1 is uniformly constructed with respect to σ up to $\sigma = 0$. (Note that the method in [15] breaks down as σ tends to 0.)

(ii) Although the smallness of σ is assumed in [15], our method enables us to construct SPS1 for any $\sigma \in [0, \sigma_1^*)$, thanks to the device by Ito [11].

(iii) Construction of SPS1 solutions proceeds in the function space $X_\varepsilon := C_\varepsilon^2 \times C^2$, as compared with $X_\varepsilon := C_\varepsilon^2 \times H^2$ in [15]. This is due to Hosono and Mimura [10].

A.1. Reduced problem. We begin with the C^1-matching solution of (RSP). For a given $x_1 \in (0, 1)$, we consider the following problem in each of the subintervals $I_{-x_1} = (0, x_1)$ and $I_{+x_1} = (x_1, 1)$:

$$\text{(1)} \qquad \begin{aligned} &\frac{1}{\sigma}(V_\pm)_{xx} + G_\pm(V_\pm) = 0, \qquad x \in I_{\pm x_1}, \\ &(V_-)_x(0) = 0 = (V_+)_x(1), \\ &V_-(x_1) = v^* = V_+(x_1), \end{aligned}$$

where $G_\pm(v) = g(h_\pm(v), v)$ for $v \in I_\pm$. It is a useful trick for the uniform construction up to $\sigma = 0$ to put

$$\text{(2)} \qquad V_\pm = v^* + \sigma W_\pm .$$

Substituting (2) into (1), we have

$$\text{(3)} \qquad \begin{aligned} &(W_\pm)_{xx} + G_\pm(v^* + \sigma W_\pm) = 0, \qquad x \in I_{\pm x_1}, \\ &(W_-)_x(0) = 0 = (W_+)_x(1), \\ &W_-(x_1) = 0 = W_+(x_1). \end{aligned}$$

The problem is to find an x_1 so that W_+ and W_- are matched in C^1-sense at $x=x_1$. Here we only consider the monotone increasing solution of (3). Using similar arguments as in [15] and [6], we can prove the following two lemmas.

LEMMA A.1. *Under* (A.0), (A.1) *and* (A.4), *there exists a uniquely determined positive constant* σ_1^* *and a unique* C^1*-function* $x_1^*(\sigma)$ *of* $\sigma \in [0, \sigma_1^*]$ *such that the function* $W^*(x;\sigma)$ *defined by*

$$W^*(x;\sigma)=\begin{cases} W_-^*(x;\sigma), & x\in \bar{I}_-^{*,\sigma}=[0, x_1^*(\sigma)], \\ W_+^*(x;\sigma), & x\in \bar{I}_+^{*,\sigma}=[x_1^*(\sigma), 1], \end{cases}$$

belongs to $C^1(\bar{I})$ *and depends continuously on* σ *in* $C^1(\bar{I})$*-topology for* $\sigma\in[0,\sigma_1^*]$ (*see* Fig. A.1), *where* $W_\pm^*(x;\sigma)$ *is the solution of* (3) *for* $x\in \bar{I}_\pm^{*,\sigma}$, *respectively. Moreover, as* $\sigma\downarrow 0$, $x_1^*(0)=\lim_{\sigma\downarrow 0} x_1^*(\sigma)$ *is given by*

$$x_1^*(0)=\frac{G_+(v^*)}{G_+(v^*)-G_-(v^*)}. \tag{4}$$

The constant σ_1^* *and the function* $x_1^*(\sigma)$ *are the same as appeared in Propositions* 1.1 *and* 1.2.

Remark A.1. When $\sigma=0$, $W_\pm^*(x;0)$ are quadratic functions of x in $I_\pm^{*,0}$.

For later use, we also construct reduced solutions which have perturbed matching points and values in C^0-sense.

LEMMA A.2. *For any* $\sigma_0\in[0,\sigma_1^*)$, *and small positive constants* δ_0 *and* ω_0, *let* $W_-=W_-(x;x_1,v,\sigma)=W_-(x;\delta,\omega,\sigma)$ *be the solution of*

$$\begin{aligned} &(W_-)_{xx}+G_-(v+\sigma W_-)=0, \qquad 0<x<x_1, \\ &(W_-)_x(0)=0, \\ &W_-(x_1^*(\sigma)+\delta)=0, \end{aligned}$$

for $(\delta,\omega,\sigma)\in\Gamma_0$, *where* $x_1=x_1^*(\sigma)+\delta$ *and* $v=v^*+\omega$, *and* $\Gamma_0=\{(\delta,\omega,\sigma)\in\mathbb{R}^3 \mid |\delta|<\delta_0, |\omega|<\omega_0, \sigma\in[0,\sigma_0)\}$. *Similarly, we can define* $W_+(x;\delta,\omega,\sigma)$. *See Fig.* A.1. *Then, we have*

(i) $W_\pm(x;\delta,\sigma)$ *depends continuously on* (δ,ω,σ) *in* C^k*-topology* $(k\geqq 2)$.

(ii) $$\begin{aligned} &\|W_-(x;\delta,\omega,\sigma)-W^*(x;\sigma)\|_{C^1([0,x_1^*(\sigma)+\delta])}\to 0 \\ &\|W_+(x;\delta,\omega,\sigma)-W^*(x;\sigma)\|_{C^1([x_1^*(\sigma)+\delta,1])}\to 0 \end{aligned} \quad \textit{as } \delta,\omega\to 0,$$

uniformly for $\sigma\in[0,\sigma_0)$.

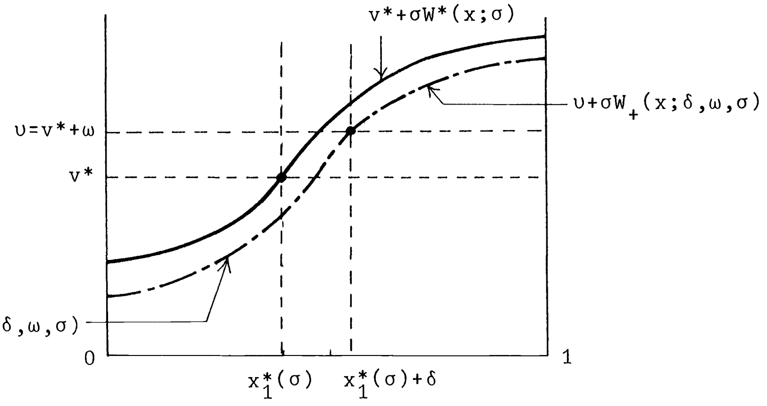


FIG. A.1. *Reduced solution which depends on parameters* δ *and* ω.

(iii) $\pm\dfrac{d}{d\delta}\left(\dfrac{\partial}{\partial x}W_{\pm}(\cdot\,;\delta,\omega,\sigma)\Big|_{x=x_1^*(\sigma)+\delta}\right)<0$ *for* $(\delta,\omega,\sigma)\in\Gamma_0$.

A.2. Approximate solutions in subintervals. We shall construct the solutions $(u_{\pm}, v_{\pm})$ of the following two problems:

$$\text{(5a)}\qquad \begin{aligned} &\varepsilon^2(u_-)_{xx}+f(u_-,v_-)=0,\\ &\frac{1}{\sigma}(v_-)_{xx}+g(u_-,v_-)=0, \end{aligned}\qquad x\in(0,x_1),\ x_1=x_1^*(\sigma)+\delta,$$

with the boundary conditions

$$\text{(5b)}\qquad \begin{aligned} &(u_-)_x(0)=0, && u_-(x_1)=h_0(v^*),\\ &(v_-)_x(0)=0, && v_-(x_1)=v^*+\omega, \end{aligned}$$

and

$$\text{(6a)}\qquad \begin{aligned} &\varepsilon^2(u_+)_{xx}+f(u_+,v_+)=0,\\ &\frac{1}{\sigma}(v_+)_{xx}+g(u_+,v_+)=0, \end{aligned}\qquad x\in(x_1,1),$$

with the boundary conditions

$$\text{(6b)}\qquad \begin{aligned} &(u_+)_x(1)=0, && u_+(x_1)=h_0(v^*),\\ &(v_+)_x(1)=0, && v_+(x_1)=v^*+\omega. \end{aligned}$$

Using the solution $W_{\pm}$ in Lemma A.2, we define $\hat{V}_{\pm}$ by

$$\text{(7)}\qquad \hat{V}_{\pm}(x;\delta,\omega,\sigma)=v^*+\omega+\sigma W_{\pm}(x;\delta,\omega,\sigma).$$

The 0th approximate solution (U_+, V_+) for (6) is defined as follows ((U_-, V_-) is analogously defined):

$$\text{(8a)}\qquad U_+(x;\delta,\omega,\varepsilon,\sigma)=h_+(\hat{V}_+(x;\delta,\omega,\sigma))+z_+(x;\delta,\omega,\varepsilon,\sigma),$$

$$\text{(8b)}\qquad V_+(x;\delta,\omega,\varepsilon,\sigma)=\hat{V}_+(x;\delta,\omega,\sigma)+\varepsilon^2\sigma Y_+(y),$$

where y is a stretched variable, namely $y=(x-x_1)/\varepsilon$. Here, Y_+ is defined by

$$Y_+(y)=\tilde{Y}(y)-\tilde{Y}(0)\zeta(x),$$

$$\tilde{Y}(y)=-\int_y^{+\infty}\int_\eta^{+\infty}\{g(h_+(\hat{V}_+(\eta',\delta,\omega,\sigma))+\tilde{z}_+(\eta',\delta,\omega,\varepsilon),v)$$
$$-g(h_+(\tilde{\hat{V}}_+(\eta',\delta,\omega,\sigma)),v)\}\,d\eta'\,d\eta,$$

where $\zeta(x)$ is a C^∞-cutoff function defined by

$$\zeta(x)=\begin{cases}1 & \text{for } x\in[0,\frac{1}{4}],\\ 0 & \text{for } x\in[\frac{1}{2},1].\end{cases}$$

Remark A.2. $Y_+(y)$ is uniformly bounded in C^2-sense for $(\varepsilon, \sigma) \in \Omega_{\varepsilon_0, \sigma_0}$, and z_+ is defined by

$$z_+(x;\, \delta, \omega, \varepsilon, \sigma) = z_+(x;\, x_1, \upsilon, \varepsilon) = \left\{ \tilde{z}_+^*\left(\frac{x-x_1}{\varepsilon};\, \upsilon\right) - h_+(\upsilon) \right\} \cdot \zeta\left(\frac{x-x_1}{1-x_1}\right),$$

where $\tilde{z}_+^*(y;\, \upsilon)$ is the layer correction term defined by the solution of

$$(9) \qquad \frac{d^2}{dy^2}\tilde{z} + f(\tilde{z}, \upsilon) = 0, \quad \tilde{z}(0) = h_0(v^*), \quad \tilde{z}(+\infty) = h_+(\upsilon).$$

The layer correction term $\tilde{z}_+^*$ satisfies the following.

LEMMA A.3 (*Fife* [4]). *The solution* $\tilde{z} = \tilde{z}_+^*(y;\, \upsilon)$ *of* (9) *exists uniquely and satisfies the following properties*:

(i) $\tilde{z}_+^*(y;\, \upsilon)$ *is monotone increasing and*

$$\left| \left\{ \left(\frac{d}{d\xi}\right)^j (\tilde{z}_+^*(y;\, \upsilon) - h_+(\upsilon)) \right\} \right| \leqq C \exp(-\kappa y) \qquad (j = 0, 1, 2),$$

where the positive constants C *and* κ *are independent of* ω *for* $|\omega| < \omega_0$ (*recall that* $\upsilon = v^* + \omega$).

$$\text{(ii)} \qquad \frac{1}{2}\left\{\frac{dz_+^*}{dy}(0;\, \upsilon)\right\}^2 = \int_{h_0(v^*)}^{h_+(\upsilon)} f(s, \upsilon)\, ds.$$

We seek the desired solution of (6) in the following form:

$$(10\text{a}) \qquad u_+ = U_+ + r_+ + \sigma \frac{d}{d\upsilon} h_+(\hat{V}_+(x;\, \delta, \omega, \sigma)) \cdot s_+,$$

$$(10\text{b}) \qquad v_+ = V_+ + \sigma s_+ = \upsilon + \sigma(W_+ + \varepsilon^2 Y_+ + s_+),$$

where (U_+, V_+) is defined by (8), and (r_+, s_+) is the unknown vector to be determined so that (u_+, v_+) becomes a true solution of (SP) in $(x_1, 1)$. Let us introduce several notations; $t = (r_+, s_+)$, $\Delta = (\delta, \omega, \sigma)$, $X_\varepsilon = C_{\varepsilon 0}^2 \times C_0^2$, and $Y = C^0 \times C^0$. We define the operator T from X_ε to Y by

$$(11) \qquad T(t, \varepsilon;\, \Delta) = \begin{pmatrix} \varepsilon^2 \dfrac{d^2}{dx^2} u_+ + f(u_+, v_+) \\ \dfrac{1}{\sigma} \dfrac{d^2}{dx^2} v_+ + g(u_+, v_+) \end{pmatrix}.$$

T is a continuously differentiable mapping of t for $(\varepsilon, \Delta) \in (0, \varepsilon_0) \times \Gamma_0$. Analogously as in Lemma 4.3 of [15], we can obtain the following lemma.

LEMMA A.4. *There exist positive constants* ε_0, δ_0 *and* ω_0 *such that the following estimates hold for* $(\varepsilon, \Delta) \in (0, \varepsilon_0) \times \Gamma_0$;

(i) $\|T(0, \varepsilon;\, \Delta)\|_Y \leqq K_0 \varepsilon$,

(ii) $\|T_t^{-1}(0, \varepsilon;\, \Delta)\|_{(Y, X_\varepsilon)} \leqq K_1 < +\infty$,

(iii) $\|T_t(t_1, \varepsilon;\, \Delta) - T_t(t_2, \varepsilon;\, \Delta)\|_{(X_\varepsilon, Y)} \leqq K_2 \|t_1 - t_2\|_{X_\varepsilon}$,

where K_i $(i = 1, 2)$ *are positive constants independent of* (ε, Δ).

Applying the generalized implicit function arguments (see [15] and [4]), we can find solutions of $T = 0$, namely, true solutions of (6) in the right-half interval $(x_1, 1)$.

THEOREM A.1. *There exist solutions* $t=t(\varepsilon;\Delta)=(r_+(\varepsilon;\Delta), s_+(\varepsilon;\Delta))$ *of* $T=0$ *for* $(\varepsilon;\Delta)\in(0,\varepsilon_0)\times\Gamma_0$ *such that* $t(\varepsilon;\Delta)$ *depends continuously on* $(\varepsilon;\Delta)$ *in* X_ε*-topology, and when* $\varepsilon\downarrow 0$, *it goes to zero with the asymptotic order as*

$$\|r_+\|_{C^2_{\varepsilon 0}}+\|s_+\|_{C^2_0}\leqq C\varepsilon,$$

where $C>0$ *is independent of* $(\varepsilon,\Delta)\in(0,\varepsilon_0)\times\Gamma_0$. *Moreover, if we denote the solution of* (6) *by* $(u_+(x;\delta,\omega,\varepsilon,\sigma), v_+(x;\delta,\omega,\varepsilon,\sigma))$, *then it satisfies*

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0}\left(\varepsilon\frac{d}{dx}u_+(x;\delta,\omega,\varepsilon,\sigma)\big|_{x=x_1^*(\sigma)+\delta}\right)^2=\int_{h_0(v^*)}^{h_+(v)}f(s,v)\,ds\quad(v=v^*+\omega),\\(12)\qquad&\lim_{\varepsilon\downarrow 0}\left(\frac{1}{\sigma}\frac{d}{dx}v_+(x;\delta,\omega,\varepsilon,\sigma)\big|_{x=x_1^*(\sigma)+\delta}\right)=\frac{d}{dx}W_+(x;\delta,\omega,\sigma)\big|_{x=x_1^*(\sigma)+\delta}.\end{aligned}$$

In a similar fashion, we can construct the solution $(u_-(x;\delta,\omega,\varepsilon,\sigma), v_-(x;\delta,\omega,\varepsilon,\sigma))$ of (5).

A.3. Matching problem and the asymptotic behavior of δ(ε, σ) and ω(ε, σ). In order to obtain a solution of (SP) on the whole interval, we have to solve the C^1-matching problem at $x=x_1^*(\sigma)+\delta$. Let us introduce the matching functions (Φ,Ψ):

$$(13a)\quad \Phi(\delta,\omega,\varepsilon,\sigma)=\left(\varepsilon\frac{d}{dx}u_+(x_1^*(\sigma)+\delta,\delta,\omega,\varepsilon,\sigma)\right)^2-\left(\varepsilon\frac{d}{dx}u_-\ \text{(the same value)}\right)^2,$$

$$(13b)\quad \Psi(\delta,\omega,\varepsilon,\sigma)=\frac{1}{\sigma}\left\{\frac{d}{dx}v_+(x_1^*(\sigma)+\delta,\delta,\omega,\varepsilon,\sigma)-\frac{d}{dx}v_-\ \text{(the same value)}\right\}.$$

Note that the factor $1/\sigma$ is important to construct the solution up to $\sigma=0$. Analogously as in [15], Φ and Ψ can be extended to be a uniform continuous function for $(\varepsilon,\Delta)\in[0,\varepsilon_0)\times\Gamma_0$, and satisfy

$$(14)\qquad \begin{aligned}&\lim_{\varepsilon\downarrow 0}\Phi(0,0,\varepsilon,\sigma)=0\\&\lim_{\varepsilon\downarrow 0}\Psi(0,0,\varepsilon,\sigma)=0\end{aligned}\qquad\text{uniformly with respect to }\sigma.$$

Moreover, it follows from assumptions (A.2) and (A.4) that

$$(15)\qquad \lim_{\varepsilon\downarrow 0}\begin{pmatrix}\Phi_\delta & \Phi_\omega\\ \Psi_\delta & \Psi_\omega\end{pmatrix}_{\substack{\delta=0\\ \omega=0}}$$

becomes an invertible matrix as in [15]. Therefore, applying again the generalized implicit function theorem of [4], we obtain a solution $\delta=\delta(\varepsilon,\sigma)$ and $\omega=\omega(\varepsilon,\sigma)$ of $\Phi=0=\Psi$ such that they are continuous functions with respect to (ε,σ) up to $\varepsilon=0$, and

$$(16)\qquad \begin{aligned}&\delta=\delta(\varepsilon,\sigma)\xrightarrow[\varepsilon\downarrow 0]{}0\\&\omega=\omega(\varepsilon,\sigma)\xrightarrow[\varepsilon\downarrow 0]{}0\end{aligned}\qquad\text{uniformly with respect to }\sigma.$$

Substituting this solution $(\delta(\varepsilon,\sigma),\omega(\varepsilon,\sigma))$ into $(u_\pm, v_\pm)$, we obtain the SPS1 $U^{\varepsilon,\sigma}$ in Theorem 1.1 which satisfies (1.5)-(1.8).

Note that the matching point $x=x_1(\varepsilon,\sigma)$ is defined by

$$(17)\qquad x_1(\varepsilon,\sigma)=x_1^*(\sigma)+\delta(\varepsilon,\sigma).$$

As for the parametric dependency of $U^{\varepsilon,\sigma}$, the only problem is the continuity at $\varepsilon=0$. However, in view of Theorem A.1, (16) and the continuity of the reduced solution $(U^{*,\sigma}, V^{*,\sigma})$ with respect to σ in $L^2\times C^1$-topology, we can see that $U^{\varepsilon,\sigma}$ is a continuous function of (ε, σ) up to $\varepsilon=0$.

Appendix 2. We prove Lemma 2.1 by contradiction. Suppose that this lemma does not hold, we can choose a sequence $\{\varepsilon_n\}_{n\geqq 1}$ with $\lim_{n\downarrow\infty}\varepsilon_n=0$ such that $\zeta_0^{\varepsilon_n,\sigma}$ is the eigenvalue of (LP) for $n\geqq 1$. Let (w_n, z_n) be an associated eigenfunction with $\zeta_0^{\varepsilon_n,\sigma}$. Then, from the solvability condition of the first equation of (LP), we have

$$(18)\qquad \langle f_v^{\varepsilon_n,\sigma} z_n, \phi_0^{\varepsilon_n,\sigma}\rangle=0 \quad \text{for } n\geqq 1.$$

When (18) is satisfied, the general form of w_n is given by

$$(19)\qquad w_n=k_n\phi_0^{\varepsilon_n,\sigma}+(L^{\varepsilon_n,\sigma}-\zeta_0^{\varepsilon_n,\sigma})^{\dagger}(-f_v^{\varepsilon_n,\sigma}z_n), \qquad k_n;\ \text{real constant.}$$

Substituting (19) into the second equation of (LP), we obtain

$$(20)\qquad \frac{1}{\sigma}(z_n)_{xx}+g_u^{\varepsilon_n,\sigma}k_n\phi_0^{\varepsilon_n,\sigma}+g_u^{\varepsilon_n,\sigma}(L^{\varepsilon_n,\sigma}-\zeta_0^{\varepsilon_n,\sigma})^{\dagger}(-f_v^{\varepsilon_n,\sigma}z_n)+g_v^{\varepsilon_n,\sigma}z_n=\zeta_0^{\varepsilon_n,\sigma}z_n.$$

Recalling $K^{\varepsilon_n,\sigma,\zeta_0^{\varepsilon_n,\sigma}}$ exists and uniformly bounded for all $n\geqq 1$ (see Lemma 3.1), we can see that $k_n\neq 0$, otherwise $(w_n, z_n)=(0,0)$, which is a contradiction. Dividing (w_n, z_n) by $k_n\sqrt{\varepsilon}_n$, we have a new eigenvector $(\hat{w}_n, \hat{z}_n)$;

$$\hat{w}_n=w_n/k_n\sqrt{\varepsilon}_n=\phi_0^{\varepsilon_n,\sigma}/\sqrt{\varepsilon}_n+(L^{\varepsilon_n,\sigma}-\zeta_0^{\varepsilon_n,\sigma})^{\dagger}(-f_v^{\varepsilon_n,\sigma}\hat{z}_n)\qquad \hat{z}_n=z_n/k_n\sqrt{\varepsilon}_n.$$

Then, $\hat{z}_n$ satisfies (20) with replacing $k_n\phi_0^{\varepsilon_n,\sigma}$ by $\phi_0^{\varepsilon_n,\sigma}/\sqrt{\varepsilon}_n$ in the second term. Therefore, recalling that $\phi_0^{\varepsilon_n,\sigma}/\sqrt{\varepsilon_n}$ converges to the constant multiple of the Dirac-function δ^* (see Lemma 2.3), we can see that $\hat{z}^*=\lim_{n\uparrow\infty}\hat{z}_n$ exists in $H_N^1(I)$, and $\langle\hat{z}^*, \delta^*\rangle\neq 0$. On the other hand, it holds from the (18) and Lemma 2.3 that $\langle\hat{z}^*, \delta^*\rangle=0$, which is a contradiction, and completes the proof.

REFERENCES

[1] D. G. Aronson, A. Tesei and H. Weinberger, *On a simple density-dependent diffusion system*, preprint.

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

[3] E. D. Conway, *Diffusion and predator-prey interaction: pattern in closed systems*, Res. Notes in Math., 101 (1984), pp. 85-133.

[4] P. C. Fife, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497-521.

[5] H. Fujii and Y. Hosono, *Neumann layer phenomena in nonlinear diffusion-systems*, in Recent Topics in Nonlinear PDE, M. Mimura and T. Nishida, eds., Math. Studies 98, North-Holland, Amsterdam, 1983, 21-38.

[6] H. Fujii, M. Mimura and Y. Nishiura, *A picture of the global bifurcation diagram in ecological interacting and diffusing systems*, Physics 5D, (1982), pp. 1-42.

[7] H. Fujii and Y. Nishiura, *Global bifurcation diagram in nonlinear diffusion systems*, in Nonlinear PDE in Applied Sciences, U.S.-Japan Seminar, Tokyo, Math. Studies, 81, North-Holland, Amsterdam, 1982, pp. 17-35.

[8] H. Fujii, Y. Nishiura and Y. Hosono, *On the structure of multiple existence of stable stationary solutions in systems of reaction-diffusion equations*, in Patterns and Waves—Qualitative Analysis of Nonlinear Differential Equations, T. Nishida, M. Mimura and H. Fujii, eds., Stud. Math. Appl., 18, North-Holland, 1986, pp. 157-219.

[9] A. Gierer and H. Meinhardt: *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30-39.

[10] Y. HOSONO AND M. MIMURA, *Singular perturbation approach to traveling waves in competing and diffusing species models*, J. Math. Kyoto Univ., 22 (1982), pp. 435-461.

[11] M. ITO, *A remark on singular perturbation methods*, Hiroshima Math. J., 14 (1985), pp. 619-629.

[12] M. MIMURA AND J. D. MURRAY, *Spatial structures in a model substrate-inhibition reaction-diffusion system*, Z. Naturforsch, 33C (1978), pp. 580-586.

[13] M. MIMURA, Y. NISHIURA, A. TESEI AND T. TSUJIKAWA, *Coexistence problem for two competing species models with density-dependent diffusion*, Hiroshima Math. J., 14 (1984), pp. 425-449.

[14] M. MIMURA, Y. NISHIURA AND M. YAMAGUTI, *Some diffusive prey and predator systems and their bifurcation problems*, Ann. New York Acad. Sci., 316 (1979), pp. 490-521.

[15] M. MIMURA, M. TABATA AND Y. HOSONO, *Multiple solutions of two-point boundary value problems of Neumann type with a small parameter*, this Journal, 11 (1980), pp. 613-631.

[16] Y. NISHIURA, *Global structure of bifurcating solutions of some reaction-diffusion systems*, this Journal, 13 (1982), pp. 555-593.

[17] ——, *Every multi-mode singularly perturbed solution recovers its stability—from a global bifurcation view point*, Lecture Notes in Biomath. 55, Springer, Berlin-New York, 1984, pp. 292-301.

[18] Y. NISHIURA AND H. FUJII, *Stability theorem for singularly perturbed solutions to systems of reaction-diffusion equations*, Proc. Japan Acad. Ser. A Math. Sci., 61 (1985), pp. 329-332.

[19] Y. NISHIURA AND H. FUJII, *An approach to the stability of singularly perturbed solutions in reaction-diffusion systems*, preprint.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

018

MULTIPLE STEADY STATES IN A BIOCHEMICAL SYSTEM*

CHUNQING LU†

Abstract. The differential equation

$$-s''+\sigma s/(1+s+ks^2)=0, \qquad 0<x<1,$$
$$s(0)=s(1)=s_0$$

for $k>0$ and for $\sigma>0$ governs the steady state in a reaction-diffusion equation which describes a substrate in a mono-enzymatic artificial membrane. We present a proof that for given $k>0$ there exists an $s^*>0$ (depending on k only) such that for $s_0>s^*$ and for any $\sigma>0$ there exist at most three positive solutions for this steady state problem.

Key words. enzyme, steady state, response function

AMS (MOS) subject classification. 92

1. Introduction. M. C. Duban [3] and J. P. Kernevez and D. Thomas [5], [8] introduced a nonlinear evolution equation in 1974 and 1975 which describes the diffusion and reaction of a substrate in a mono-enzymatic artificial membrane, when the enzyme is inhibited by an excess of substrate:

$$S_t-D_SS_{xx}+R(S)=0, \tag{1.0}$$

together with boundary conditions

$$S=S_0 \quad \text{at } x=0 \quad \text{and} \quad x=L \quad (L \text{ membrane thickness}),$$

and the given condition, where $S=S(x,t)$ represents the concentration of the substrate of the membrane, D_S is a constant, the coefficient of diffusion, and

$$R(S)=V_MS/[k_S+S(1+S/k_{SS})]$$

is the velocity due to the reaction where k_S is the Michaelis constant, k_{SS} the inhibition constant of S for the enzyme and V_M the maximal value of the reaction rate. Nondimensionalizing the quantities, one obtains the following equation:

$$\begin{aligned} &s_t-s_{xx}+\sigma F(s)=0, \quad 0<x<1, \quad t>0,\\ &s(0,t)=s(1,t)=s_0,\\ &s(x,0) \text{ given}\end{aligned} \tag{1.1}$$

where $s=S/k_S$, $F(s)=s/(1+s+ks^2)$ and $\sigma=(V_M/k_S)L^2$, $k=k_S/k_{SS}$ and $s_0=S_0/k_S$ are positive constants. Then the steady state equation associated with (1.1) is the two-point boundary value problem

$$-s''+\sigma F(s)=0, \qquad s(0)=s(1)=s_0. \tag{1.2}$$

J. P. Kernevez proved that for σ large enough and s_0 conveniently chosen, (1.2) admits at least three solutions, and presented numerical evidence that there are at most three solutions for any positive s_0, k and σ [6, pp. 64–73]. In 1976, C. M. Brauner and B. Nicolaenko [2] rewrote (1.2) in the following form:

$$\begin{aligned} &-u''+u/(\alpha^2+\beta\alpha u+k\beta^2u^2)=0, \qquad 0<x<1,\\ &u(0)=u(1)=1\end{aligned} \tag{1.3}$$

* Received by the editors June 3, 1985; accepted for publication (in revised form) October 22, 1986.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260. This author is on leave from Computing Center, Academia Sinica, Beijing, P. R. China.

by defining $u=s/s_0$, $\alpha^2=1/\sigma$, $\beta=s_0\alpha$. They studied the stability of multiple solutions of (1.3) for sufficiently small $\alpha>0$, i.e., for σ large enough and s_0 conveniently chosen in problem (1.2) ([1], [2]). Their discussion is based on their claim that there exist at most three solutions for problem (1.3). However, they did not provide a proof for this.

In this paper we employ another change of variable and parameters from [1] in order to present a rigorous proof that for given $k>0$ there exists an $s^*=s^*(k)$ such that for $s_0>s^*$ and for any positive σ the steady state problem (1.2) admits no more than three solutions. The result we obtain here agrees with that of experiment and numerical analysis because it is known that in order to get multiple steady states we must have σ and s_0 sufficiently large [6, p. 97]. The change of parameters we introduce in (2.2) will let us study sufficiently small ε, which is related to $1/s_0$. The latter is proportional to the concentration on the edge of the membrane. Problems of a similar kind are treated in [4] and [9]. However, the nonlinearities are completely different and very little of the analysis is relevant here.

2. Formulation of the perturbation problem and the response function. We first make a change of variable $u(x)=s(x)/s_0$ where $s=s(x)$, s_0 satisfy (1.2). Then (1.2) becomes

$$\begin{aligned} &-u''+\sigma u/(1+s_0u+ks_0^2u^2)=0, \qquad 0<x<1,\\ &u(0)=u(1)=1. \end{aligned} \tag{2.1}$$

Next we introduce some new parameters β and ε by defining

$$\beta=s_0/\sqrt{\sigma}, \qquad \varepsilon=1/2ks_0, \tag{2.2}$$

where ε will be the perturbation parameter. Thus (2.1) becomes the following equation:

$$-u''+u/[k\beta^2(u^2+2\varepsilon u+4k\varepsilon^2)]=0, \qquad 0<x<1 \tag{2.3}$$

with $u(0)=u(1)=1$.

Our purpose is to prove that for given $k>0$ there exists an $\varepsilon_0>0$ such that 0 (2.3) has no more than three solutions for $0<\varepsilon<\varepsilon_0$ and for any $\beta>0$, where $\varepsilon_0=\varepsilon_0(k)$ depends on k only. Obviously, this means that for given $k>0$ there exists a value $s^*=s^*(k)$ such that (1.2) admits at most three solutions for all $s_0>s^*$ and any $\sigma>0$.

Multiplying (2.3) by u', we see that

$$[u'^2/2]'=uu'/[k\beta^2(u^2+2\varepsilon u+4k\varepsilon^2)]. \tag{2.4}$$

If we integrate both sides of (2.4) with respect to x, it follows that

$$[u'(x)]^2/2=\left(\frac{1}{k\beta^2}\right)\int_{u(1/2)}^{u(x)}\{u/(u^2+2\varepsilon u+4k\varepsilon^2)\}\,du+u'^2(\tfrac12)/2. \tag{2.5}$$

We claim that $u'(\frac12)=0$. Assume that $u=u(x)$ is any solution of (2.3) with $u(0)=u(1)=1$ (for the proof of the existence of such a solution for (2.3) see [7]). From (2.5), $u'(0)^2=u'(1)^2$. Since $u''(x)>0$, u is a convex function of x on $[0,1]$, so it must be that $u'(1)=-u'(0)$. Suppose $u'(1)=m$. Now we see that (2.3) with boundary conditions is equivalent to the same equation with initial conditions either $u(1)=1$, $u'(1)=m$ or $u(0)=1$, $u'(0)=-m$. Let $x=\frac12+z$ and $v(z)=u(\frac12+z)$, where x is in $[0,1]$ and z in $[-\frac12,\frac12]$. Then $v(z)$ satisfies

$$-v''+v/[k\beta^2(v+2\varepsilon v+4k\varepsilon^2)=0, \qquad -\tfrac12<z<\tfrac12 \tag{2.6}$$

with $v(-\frac12)=v(\frac12)=1$; moreover, $v'(\frac12)=-v'(-\frac12)=m$. It is easy to check that $v(-z)=w(z)$ is also a solution for (2.6) with the initial conditions. By the uniqueness of $v(z)$, one obtains $v(z)=v(-z)$, which implies that the $u(x)$ profile is symmetric with respect to $x=\frac{1}{2}$, and hence $u'(\frac{1}{2})=0$. It also follows that $u'(x)>0$ for $x>\frac{1}{2}$, $u'(x)<0$ for $x<\frac{1}{2}$ by the convexity of $u(x)$. One can see that the number of solutions of (2.3) will be determined by the number of values of $u(\frac{1}{2})$.

To determine which condition $u(\frac{1}{2})$ has to satisfy we begin with (2.5). Let $y=1/u(\frac{1}{2})$. Then for x in $(\frac{1}{2}, 1]$, $1/y<u(x)$ and

$$u'(x)=\left\{\left[\frac{2}{k\beta^2}\right]\int_{y^{-1}}^{u(x)}[s/(s^2+2\varepsilon s+4k\varepsilon^2)]\,ds\right\}^{1/2}. \tag{2.7}$$

Computing the integral in (2.7), we obtain

$$\begin{aligned}\int_{y^{-1}}^{u(x)}[s/(s^2+2\varepsilon s+4k\varepsilon^2)]\,ds&=\ln\{[\varepsilon+u(x)]^2+(4k-1)\varepsilon^2\}^{1/2}\\&\quad-\ln[(\varepsilon+y^{-1})^2+(4k-1)\varepsilon^2]^{1/2}+\varepsilon h_0(u(x),y)\end{aligned} \tag{2.8}$$

where

$$h_0(u(x),y)=-\int_{y^{-1}}^{u(x)}\{1/[(s+\varepsilon)^2+(4k-1)\varepsilon^2]\}\,ds. \tag{2.9}$$

Denote the right side of (2.8) by $K(\varepsilon, y, u(x))/2$. From (2.7) we have

$$u'/\sqrt{K}(\varepsilon, y, u)=1/(\beta\sqrt{k}). \tag{2.10}$$

Moreover,

$$\int_{u(1/2)}^{1}K^{-1/2}(\varepsilon, y, u)\,du=\frac{1}{(2\beta\sqrt{K})}. \tag{2.11}$$

For convenience, we make a change of variable as follows:

$$\varepsilon+u=(\varepsilon+1/y)\,e^{t^2},\qquad du=2(\varepsilon+y^{-1})t\,e^{t^2}\,dt.$$

Then (2.11) becomes

$$\begin{aligned}2(\varepsilon+y^{-1})\int_0^{v(\varepsilon,y)}&\{\ln[(1+\varepsilon y)^2e^{2t^2}+(4k-1)(\varepsilon y)^2]\\&-\ln[(1+\varepsilon y)^2+(4k-1)(\varepsilon y)^2]+2h(\varepsilon,y,t)\}^{-1/2}t\,e^{t^2}\,dt\\=\frac{1}{(2\beta\sqrt{k})}&\end{aligned} \tag{2.12}$$

where

$$h(\varepsilon,y,t)=-\varepsilon y\int_1^{(1+\varepsilon y)e^{t^2}-\varepsilon y}\{1/[(s+\varepsilon y)^2+(4k-1)\varepsilon^2y^2]\}\,ds, \tag{2.13}$$

$y\geqq 1$, and $v(\varepsilon, y)=\sqrt{\{\ln(\varepsilon+1)-\ln(\varepsilon+y^{-1})\}}$. Evidently, it is necessary to prove that, for given $k>0$, if ε is small enough, then for any $\beta>0$ there are no more than three y's satisfying (2.12). We denote the function on the left in (2.12) by $f_\varepsilon(y)$, which is called the response function, and then (2.12) becomes

$$f_\varepsilon(y)=1/(2\beta\sqrt{k}). \tag{2.14}$$

3. Notation and conventions. Since we only deal with nonnegative solutions of (2.3), we shall always assume that $u(x)>0$. From § 2, $y\geqq 1$. If we set $\varepsilon=0$ in (2.3), then we obtain the so-called reduced problem, and the corresponding response function will be denoted by $f_0(y)$. In the reduced problem, of course, we have

$$f_0(y) = (2\beta\sqrt{k})^{-1}. \tag{3.1}$$

Since we always fix $k > 0$, we shall omit the words "for given $k > 0$" in all our statements. Let G denote

$$\begin{aligned} G(\varepsilon, y, t) = {} & \ln[(1+\varepsilon y)^2 e^{2t^2} + (4k-1)(\varepsilon y)^2] \\ & - \ln[(1+\varepsilon y)^2 + (4k-1)(\varepsilon y)^2] + 2h(\varepsilon, y, t) \end{aligned} \tag{3.2}$$

where $h(\varepsilon, y, t)$ is given in (2.13).

We shall always denote the derivative of any function with respect to y by "prime," and with respect to t by "dot."

We now list some functions which are repeatedly used in the following sections:

$$v = v(\varepsilon, y) = \sqrt{\{\ln(1+\varepsilon) - \ln(y^{-1}+\varepsilon)\}},$$

$$f_\varepsilon(y) = 2(\varepsilon + y^{-1}) \int_0^v t\, e^{t^2} G^{-1/2}\, dt,$$

$$G_1 = G_1(\varepsilon, y, t) = \frac{\partial G}{\partial y},$$

$$G^{\bullet} = G^{\bullet}(\varepsilon, y, t) = \frac{\partial G}{\partial t},$$

$$G_1^{\bullet} = G_1^{\bullet}(\varepsilon, y, t) = \frac{\partial^2 G}{\partial t \partial y},$$

$$G_2 = G_2(\varepsilon, y, t) = \frac{\partial^2 G}{\partial y^2},$$

$$G^{\bullet\bullet} = G^{\bullet\bullet}(\varepsilon, y, t) = \frac{\partial^2 G}{\partial t^2},$$

$$F_\varepsilon(y) = f_\varepsilon''(y) + \{2 - [\varepsilon y(1 + 8k\varepsilon y + 4k(\varepsilon y)^2)][2(1+\varepsilon y)(1 + 2\varepsilon y + 4k\varepsilon^2 y^2)]^{-1}\} y f_\varepsilon'(y),$$

$$T_1 = T_1(\varepsilon, y, t) = (1+\varepsilon y)^2 e^{2t^2} + (4k-1)(\varepsilon y)^2,$$

$$T_2 = T_2(\varepsilon, y, t) = (1+\varepsilon y)^2 e^{2t^2} + 2(4k-1)\varepsilon y(1+\varepsilon y)\, e^{t^2} - (4k-1)(\varepsilon y)^2.$$

Elementary calculations show that

$$G_1 = 2\varepsilon[(1+\varepsilon y)\, e^{2t^2} + e^{t^2} + (4k-1)\varepsilon y]/T_1 - 4\varepsilon(1 + 2k\varepsilon y)/T_1(\varepsilon, y, 0),$$

$$G_1^{\bullet} = -4\varepsilon t\, e^{t^2} T_2 / T_1^2,$$

$$G^{\bullet} = 4t\, e^{t^2}(1+\varepsilon y)[(1+\varepsilon y)\, e^{t^2} - \varepsilon y] T_1^{-1},$$

$$\begin{aligned} G_2 = {} & 2\varepsilon^2 T_1^{-2}[-(1+\varepsilon y)^2 e^{4t^2} - 2(1+\varepsilon y)\, e^{3t^2} \\ & - (4k-1)(2\varepsilon y + 2\varepsilon^2 y^2 - 1)\, e^{2t^2} - 2(4k-1)\varepsilon y\, e^{t^2} - (4k-1)(\varepsilon y)^2] \\ & - 8\varepsilon^2 T_1^{-2}(\varepsilon, y, 0)(k - 1 - 4k\varepsilon y - 4k^2\varepsilon^2 y^2), \end{aligned}$$

$$\begin{aligned} G^{\bullet\bullet} = {} & 4(1+\varepsilon y)\, e^{t^2}[4t^2 e^{t^2}(1+\varepsilon y) + (1+\varepsilon y)\, e^{t^2} - 2t^2\varepsilon y - \varepsilon y] \cdot T_1^{-1} \\ & - 16(1+\varepsilon y)^3 t^2 e^{3t^2}[(1+\varepsilon y)\, e^{t^2} - \varepsilon y] T_1^{-2}. \end{aligned}$$

For simplicity, we shall only give the details in the case $k=\frac{1}{4}$ in this paper. We do indicate those quantities introduced below which depend on k in the general case. For $k\neq\frac{1}{4}$ the interested reader can refer to [7]. We see that if $k=\frac{1}{4}$, then

$$G=2t^2+2\varepsilon y(e^{-t^2}-1)/(1+\varepsilon y), \tag{3.3}$$

$$T_1=T_2=(1+\varepsilon y)^2\,e^{2t^2}, \tag{3.4}$$

$$G_1=-2\varepsilon(1-e^{-t^2})/(1+\varepsilon y)^2, \tag{3.5}$$

$$G^{\bullet}=4t(1+\varepsilon y-\varepsilon y\,e^{-t^2})/(1+\varepsilon y), \tag{3.6}$$

$$G_1^{\bullet}=-4\varepsilon t/\{(1+\varepsilon y)^2\,e^{t^2}\}, \tag{3.7}$$

$$G_2=4\varepsilon^2(1-e^{-t^2})/(1+\varepsilon y)^3, \tag{3.8}$$

$$G^{\bullet\bullet}=4\{(1+\varepsilon y)\,e^{t^2}+2\varepsilon yt^2-\varepsilon y\}/\{(1+\varepsilon y)\,e^{t^2}\}, \tag{3.9}$$

$$f_\varepsilon(y)=2(\varepsilon+y^{-1})\int_0^v t\,e^{t^2}G^{-1/2}\,dt, \tag{3.10}$$

$$\begin{aligned} f_\varepsilon'(y)&=(1+\varepsilon)/[(1+\varepsilon y)G^{1/2}(\varepsilon,y,v)]\\ &\quad-\left(\frac{2}{y^2}\right)\int_0^v t\,e^{t^2}G^{-1/2}\,dt-\left(\varepsilon+\frac{1}{y}\right)\int_0^v t\,e^{t^2}G_1G^{-3/2}\,dt, \end{aligned} \tag{3.11}$$

$$F_\varepsilon(y)=y^2f_\varepsilon''(y)+\{2-\varepsilon y/[2(1+\varepsilon y)]\}yf_\varepsilon'(y). \tag{3.12}$$

4. The reduced problem. At $\varepsilon=0$, $G=G(0,y,t)=2t^2$ and

$$f_0(y)=(\sqrt{2}/y)\int_0^{\sqrt{\ln y}} e^{t^2}\,dt, \tag{4.1}$$

$$f_0'(y)=1/[y\delta/\sqrt{(2\ln y)}]-(\sqrt{2}/y^2)\int_0^{\sqrt{\ln y}} e^{t^2}\,dt, \tag{4.2}$$

$$f_0''(y)=-2/[y^2\sqrt{(2\ln y)}]-1/[y^2\{\sqrt{(2\ln y)}\}^3]+\left(\frac{2\sqrt{2}}{y^3}\right)\int_0^{\sqrt{\ln y}} e^{t^2}\,dt, \tag{4.3}$$

$$F_0(y)=-1/[\sqrt{(2\ln y)}]^3<0 \tag{4.4}$$

for $y>1$. It follows that $f_0(y)$ has at most one critical point, the maximum. Thus the reduced problem admits at most two solutions for any $\beta>0$.

5. Outline of the argument. For our purpose it suffices to prove that there exists an $\varepsilon_0=\varepsilon_0(k)$ such that for ε in $(0,\varepsilon_0)$ and for $y>1$, $f_\varepsilon(y)$ has only two critical points; one is a local maximum and the other is a local minimum. In order to do this we shall prove the following.

THEOREM 1. *There exists an* $\varepsilon_1=\varepsilon_1(k)$ *such that for* ε *in* $(0,\varepsilon_1)$ *and* $y>1/\varepsilon^2$, $f_\varepsilon'(y)>0$.

THEOREM 2. *There exist constants $d = d(k)$, $Y_0 = Y_0(k)$ and $\varepsilon_2 = \varepsilon_2(k)$ such that for y in $[Y_0, d/\varepsilon]$ and ε in $(0, \varepsilon_2)$, $f_\varepsilon'(y) < 0$.*

THEOREM 3. *There exists an $\varepsilon_3 = \varepsilon_3(k)$ such that for $0 < \varepsilon < \varepsilon_3$ and y in $[d/\varepsilon, 1/\varepsilon^2]$, $F_\varepsilon(y) > 0$, where d is given in Theorem 2 and F_ε is listed in* § 2.

It is then clear that $f_\varepsilon(y)$ has only one critical point, the local minimum, for $y > Y_0$ if ε is small enough, because if $f_\varepsilon' = 0$, then by Theorem 3 and the expression of F_ε', $f_\varepsilon'' > 0$.

THEOREM 4. *There exists an $\varepsilon_4 = \varepsilon_4(k)$ such that for $0 < \varepsilon < \varepsilon_4$, $f_\varepsilon(y)$ has exactly one critical point, the local maximum, on $[1, Y_0]$, where Y_0 is given in Theorem 2.*

The technique used in proving Theorems 1-3 is to estimate the integrals which appear in the expressions of $f_\varepsilon'(y)$ and $F_\varepsilon(y)$ carefully. For Theorem 4, however, we shall compare $f_\varepsilon(y)$ with $f_0(y)$ for small ε.

Now if we set $\varepsilon_0 = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$, then all the conclusions of Theorems 1-4 hold, as desired.

6. The proofs of Theorems 1 and 2. Let δ be any constant such that $0 < \delta < v = v(\varepsilon, y)$. Then

$$\int_0^v t e^{t^2} G^{-1/2}\, dt = \int_0^\delta + \int_\delta^v, \qquad \int_0^v t e^{t^2} G^{-3/2} G_1\, dt = \int_0^\delta + \int_\delta^v. \tag{6.1}$$

By integration by parts,

$$\begin{aligned} \int_\delta^v t e^{t^2} G^{-1/2}\, dt &= (1+\varepsilon)y/[2(1+\varepsilon y)\sqrt{G}(\varepsilon, y, \delta)] \\ &\quad - e^{\delta^2}/[2\sqrt{G}(\varepsilon, y, \delta)] + \frac{1}{4}\int_\delta^v e^{t^2} G^{-3/2} G^{\cdot}\, dt. \end{aligned} \tag{6.2}$$

Substituting (6.1) and (6.2) into (3.11), we obtain

$$\begin{aligned} f_\varepsilon'(y) &= e^{\delta^2}/[y^2\sqrt{G}(\varepsilon, y, \delta)] - \left(\frac{1}{y^2}\right)\int_0^v t e^{t^2} G^{-3/2}[2G + (1+\varepsilon y) y G_1]\, dt \\ &\quad - \left(\frac{1}{y^2}\right)\int_0^v e^{t^2} G^{-3/2}[G^{\cdot}/2 + (1+\varepsilon y) y t G_1]\, dt. \end{aligned} \tag{6.3}$$

From (3.5) and (3.6),

$$\frac{G^{\cdot}}{2} + (1+\varepsilon y) y t G_1 = 2t/(1+\varepsilon y), \tag{6.4}$$

and then (6.3) becomes

$$\begin{aligned} f_\varepsilon'(y) &= e^{\delta^2}/[y^2\sqrt{G}(\varepsilon, y, \delta)] - \left(\frac{1}{y^2}\right)\int_0^\delta t e^{t^2} G^{-3/2}[2G + (1+\varepsilon y) y G_1]\, dt \\ &\quad - \frac{2}{y^2(1+\varepsilon y)}\int_\delta^v t e^{t^2} G^{-3/2}\, dt. \end{aligned} \tag{6.5}$$

LEMMA 6.1. *For given $\varepsilon y > 0$, $tG^{-3/2}(\varepsilon, y, t)$ is a decreasing function of t, where $t > 0$.*

Proof. $G^{\cdot} > 0$ for $t > 0$ and for given $\varepsilon y > 0$; hence from $G > 0$

$$(1/G)^{\cdot} < 0. \tag{6.6}$$

All we need is to have $(tG^{-1/2})^{\cdot} < 0$, which is equivalent to

$$(6.7) \qquad (t^2/G^{-1})^{\cdot} < 0 \quad \text{for } t > 0.$$

Notice that $(t^2/G^{-1})^{\cdot} = t(2G - tG^{\cdot})/G^2$ and $G = G^{\cdot} = 0$ at $t = 0$. Since

$$(2G - tG^{\cdot})^{\cdot} = G^{\cdot} - tG^{\cdot\cdot} = -8\varepsilon y t^3/(1+\varepsilon y) < 0$$

for $t > 0$, (6.7) holds. This proves the lemma.

LEMMA 6.2. *For each $c > 0$ there is a $Y(c)$ such that for $0 < a < 1$ and $Y > Y(c)$,*

$$(6.8) \qquad Y/[2\sqrt{\ln Y}] < \int_a^{\sqrt{\ln Y}} e^{t^2}\, dt < (\tfrac{1}{2} + c)Y/\sqrt{\ln Y}.$$

Also for each $b > 0$ there is a $W(b)$ such that for $W > W(b)$

$$(6.9) \qquad W/[2\sqrt{\ln W}] < \int_b^{\sqrt{\ln W}} e^{t^2}\, dt.$$

Proof. By integration by parts,

$$(6.10) \qquad \begin{aligned} \int_1^{\sqrt{\ln y}} e^{t^2}\, dt &= y/[2\sqrt{\ln y}] - e/2 + \frac{1}{2}\int_1^{\sqrt{\ln y}} e^{t^2}/t^2\, dt, \\ &\geqq y/[2\sqrt{\ln y}] - e/2 + \frac{1}{2}\int_1^2 e\, dt = y/[2\sqrt{\ln y}] \end{aligned}$$

for $Y \geqq e^4$, since $e^{t^2}/t^2 \geqq e$ for $t \geqq 1$. We then see that (6.10) implies the first inequality in (6.8) for $Y > e^4$. In the same way one can prove (6.9). The right half of the inequality (6.8) follows from

$$(6.11) \qquad \lim_{y\to\infty} \left\{ \int_0^{\sqrt{\ln y}} e^{t^2}\, dt / [y/\sqrt{\ln y}] \right\} = \frac{1}{2}. \qquad \text{Q.E.D.}$$

LEMMA 6.3. *There exist an $\varepsilon_{*1} = \varepsilon_{*1}(k)$ and a $\delta_1 = \delta_1(k)$ such that for t in $(0, \delta)$, where $0 < \delta < \delta_1$, and for $y > 1/\varepsilon^2$, where ε is in $(0, \varepsilon_{*1})$,*

$$2G + (1+\varepsilon y)yG_1 < 0.$$

Proof. Let $g = g(\varepsilon, y, t) = 2G + (1+\varepsilon y)yG_1$. Then $g(\varepsilon, y, 0) = 0$ for any $y > 0$ and $g^{\cdot} = 4t(2 + 2\varepsilon y - 3\varepsilon y/e^{t^2})/(1+\varepsilon y)$. Let $g_1(\varepsilon, y, t) = 2 + 2\varepsilon y - 3\varepsilon y\, e^{-t^2}$. Then for given $\varepsilon y > 0$, $g_1(\varepsilon, y, t) = 2 - \varepsilon y + g_1^{\cdot}(\varepsilon, y, \alpha)t$ for some $\alpha = \alpha(\varepsilon, y, t)$, with $0 < \alpha < t$. Take $\delta_1 = \sqrt{2}/2$ if $k = \frac{1}{4}$; then $g_1^{\cdot}(\varepsilon, y, \alpha) < 6\varepsilon yt\, e^{-t^2}$ for $0 < \alpha < t < \delta < \delta_1$. Thus

$$\begin{aligned} g_1(\varepsilon, y, t) &< 2 - 2\varepsilon y(1 - 3t^2 e^{-t^2}) < 2 - 2\varepsilon y(1 - 3\delta^2 e^{-\delta^2}) \\ &< 2 - 2(1 - 3\delta_1^2 e^{-\delta_1^2})/\varepsilon \end{aligned}$$

for $y > 1/\varepsilon^2$. Fix δ_1 and let ε tend to 0 to complete the proof. Q.E.D.

LEMMA 6.4. *For any $\delta > 0$ there is a $r = r(k, \delta)$ such that $G(\varepsilon, y, \delta) > r > 0$ for all $\varepsilon y > 0$.*

Proof. $\partial G(\varepsilon, y, t)/\partial(\varepsilon y) = G_1/\varepsilon < 0$ and $\lim_{\varepsilon y\to\infty} G(\varepsilon, y, \delta) = 2(\delta^2 - 1 + e^{-\delta^2}) > 0$. Q.E.D.

We now proceed with the proof of Theorem 1. First of all we set $\delta < \min\{\delta_1, 1\}$ and $\varepsilon < \varepsilon_{*1}$ in (6.5), where δ_1 and ε_{*1} are given in Lemma 6.3. Then by Lemmas 6.1

and 6.3 it follows that

$$(6.12)\qquad f'_\varepsilon(y) > e^{\delta^2}/[y^2\sqrt{G}(\varepsilon, y, \delta)] - 2/[y^2(1+\varepsilon y)] \int_\delta^v [\max(tG^{-3/2})]\, e^{t^2}\, dt.$$

Notice that $\max tG^{-3/2} = \delta G^{-3/2}(\varepsilon, y, \delta)$ for t in $[\delta, v]$ and $v < \sqrt{[\ln(1+\varepsilon) - \ln \varepsilon]}$ for $y > 1$. We then use $1/(1+\varepsilon y) < 3/(2\varepsilon y)$ in (6.12) and obtain that, for $\varepsilon < \varepsilon_{*1}$ and $y \geqq 1/\varepsilon^2$,

$$(6.13)\qquad y^2 G^{3/2}(\varepsilon, y, \delta) f'_\varepsilon(y) > e^{\delta^2} G(\varepsilon, y, \delta) - (3\delta/\varepsilon y) \int_0^{v_0(\varepsilon)} e^{t^2}\, dt$$

where $v_0(\varepsilon) = \sqrt{\{\ln(1+\varepsilon) - \ln \varepsilon\}}$. Let ε_{*2} be a number such that $v_0(\varepsilon) > Y(\frac{1}{2})$ for $\varepsilon < \varepsilon_{*2}$, where $Y(\frac{1}{2})$ is the value of $Y(c)$ at $c = \frac{1}{2}$ in Lemma 6.2. Denote the function on the left in (6.13) by $n(\varepsilon, y, \delta)$ and apply Lemmas 6.2 and 6.4 to show that

$$(6.14)\qquad n(\varepsilon, y, \delta) > r(k, \delta) - 3\delta(1+\varepsilon)/v_0(\varepsilon).$$

Since the choice for δ does not depend upon ε and $(1+\varepsilon)/v_0(\varepsilon)$ tends to 0 as $\varepsilon \to 0$, there exists $\varepsilon_{*3} = \varepsilon_{*3}(k)$ such that

$$(1+\varepsilon)/v_0(\varepsilon) < r(k, \delta)/6\delta.$$

Thus, setting $\varepsilon_1 = \min\{\varepsilon_{*1}, \varepsilon_{*2}, \varepsilon_{*3}\}$, we obtain Theorem 1.

To prove Theorem 2 we require the following lemma which will prove that the second term in the expression for $f'_\varepsilon(y)$ is negative.

LEMMA 6.5. *There exists a number $d = d(k)$ such that for εy in $(0, d]$ and $t > 0$, $2G + (1+\varepsilon y) y G_1 > 0$.*

Proof. We shall use the same notation as in the proof of Lemma 6.3. Since $g_1(0,0,0) = 2$, there is a number $d = d(k)$ such that $g_1(\varepsilon, y, 0) \geqq 0$ for εy in $[0, d]$ and $g_1(\varepsilon, y, 0) = 0$ at $\varepsilon y = d$. Also $\dot{g}_1(\varepsilon, y, t) > 0$ for $t > 0$ and $\varepsilon y > 0$. This implies the lemma.

Remark. In the case of $k = \frac{1}{4}$, $d = 2$.

Proof of Theorem 2. We first take d as in Lemma 6.5, and then set $\delta = 1$ in (6.5). Using Lemmas 6.1 and 6.6, we see that for $1 < y \leqq d/\varepsilon$,

$$(6.15)\qquad f'_\varepsilon(y) < e/[y^2\sqrt{G}(\varepsilon, y, 1)] - \{2v/[y^2(1+\varepsilon y)G^{3/2}(\varepsilon, y, v)]\} \int_1^v e^{t^2}\, dt.$$

Further, assume that $\varepsilon < 1/(2e^4)$. Then $v > 2$ for $y > 2e^4$ and hence from (6.10)

$$(6.16)\qquad \int_1^v e^{t^2}\, dt > (1+\varepsilon)y/[2v(1+\varepsilon y)].$$

This shows that for $\varepsilon < 1/(2e^4)$ and for $2e^4 \leqq y \leqq d/\varepsilon$

$$(6.17)\qquad f'_\varepsilon(y) < e/[y^2\sqrt{G}(\varepsilon, y, 1)] - (1+\varepsilon)/[y(1+\varepsilon y)^2 G^{3/2}(\varepsilon, y, v)].$$

Since $G(\varepsilon, y, 1) = 2 + 2\varepsilon y(1/e - 1)/(1+\varepsilon y) \geqq m(d)$, where $m(d) = 2 + 2d(1/e - 1)/(1+d)$, for $y < d/\varepsilon$, and $G(\varepsilon, y, v) < 2v^2 < 2\ln y$ for $y > 1$, it follows that for y in $[2e^4, d/\varepsilon]$,

$$(6.18)\qquad yf'_\varepsilon(y) < e/\sqrt{m}(d) - y/[(1+d)^2(2\ln y)^{3/2}].$$

Thus there exists an $\varepsilon_2 = \varepsilon_2(k)$ and $Y_0 = Y_0(k)$ such that $f'_\varepsilon(y) < 0$ for $0 < \varepsilon < \varepsilon_2$ and y in $[Y_0, d/\varepsilon]$, which is desired.

7. The proof of Theorem 3. To begin with, let δ be chosen as in the beginning of § 6. We start at (6.5). Then

$$\begin{aligned}
[y^2 f_\varepsilon'(y)]' &= 2y f_\varepsilon'(y) + y^2 f_\varepsilon''(y) \\
&= -e^{\delta^2} G_1(\varepsilon, y, \delta)/[2G^{3/2}(\varepsilon, y, \delta)] \\
&\quad - \int_0^\delta t\, e^{t^2} \{ G^{-3/2}[(3+2\varepsilon y)G_1 + (1+\varepsilon y) y G_2] \\
&\qquad - (\tfrac{3}{2}) G_1 G^{-5/2}[2G + (1+\varepsilon y) y G_1]\}\, dt \\
&\quad - (1+\varepsilon)/[(1+\varepsilon y)^3 G^{3/2}(\varepsilon, y, v)] \\
&\quad + [1/(1+\varepsilon y)^2] \int_\delta^v t\, e^{t^2} G^{-3/2}[2\varepsilon + 3(1+\varepsilon y) G_1 G^{-1}]\, dt,
\end{aligned} \tag{7.1}$$

$$\begin{aligned}
F_\varepsilon(y) &= [y^2 f_\varepsilon'(y)]' - \varepsilon y^2 f_\varepsilon'(y)/[2(1+\varepsilon y)] \\
&= -e^{\delta^2} G_1(\varepsilon, y, \delta)/[2G^{3/2}(\varepsilon, y, \delta)] \\
&\quad - \varepsilon\, e^{\delta^2}/[2(1+\varepsilon y)\sqrt{G}(\varepsilon, y, \delta) - (1+\varepsilon)/[(1+\varepsilon y)^3 G^{3/2}(\varepsilon, y, v)] \\
&\quad - \int_0^\delta t\, e^{t^2} \{ G^{-3/2}[(3+2\varepsilon y)G_1 + (1+\varepsilon y) y G_2 - \varepsilon[2G + yG_1(1+\varepsilon y)] \\
&\qquad \cdot (1+\varepsilon y)^{-1}/2] - (\tfrac{3}{2}) G_1 G^{-5/2}[2G + (1+\varepsilon y) y G_1]\}\, dt \\
&\quad + [3/(1+\varepsilon y)^2] \int_\delta^v t\, e^{t^2} G^{-3/2}[\varepsilon + (1+\varepsilon y) G_1 G^{-1}]\, dt.
\end{aligned} \tag{7.2}$$

We now define several functions which will be used in this section as follows:

$$\begin{aligned}
N = N(\varepsilon, y, t) &= (3+2\varepsilon y) G_1/G + y(1+\varepsilon y) G_2/G \\
&\quad - \varepsilon[2 + (1+\varepsilon y) y G_1/G]/(1+\varepsilon y)/2 \\
&\quad - 3G_1[2 + (1+\varepsilon y) G_1/G]/(2G), \\
M = M(\varepsilon, y, \delta) &= -e^{\delta^2} G_1(\varepsilon, y, \delta)/[2G^{3/2}(\varepsilon, y, \delta)] \\
&\quad - \varepsilon\, e^{\delta^2}/[2(1+\varepsilon y)\sqrt{G}(\varepsilon, y, \delta)], \\
Q = Q(\varepsilon, y, t) &= \varepsilon + (1-\varepsilon y) G_1/G.
\end{aligned} \tag{7.3}$$

Then rewrite (7.2) as

$$\begin{aligned}
F_\varepsilon(y) &= M - (1+\varepsilon)/[(1+\varepsilon y)\sqrt{G}(\varepsilon, y, v)]^3 \\
&\quad - \int_0^\delta t\, e^{t^2} N/\sqrt{G}\, \mathrm{dt} + [3/(1+\varepsilon y)^2] \int_\delta^v t\, e^{t^2} Q/G^{3/2}\, dt.
\end{aligned} \tag{7.4}$$

LEMMA 7.1. *G_1/G is an increasing function of $t > 0$ for any given $\varepsilon y > 0$.*

Proof. $(G_1/G)^{\cdot} = (GG_1^{\cdot} - G_1 G^{\cdot})/G^2 = G_1^{\cdot}(G - G_1 G^{\cdot}/G_1^{\cdot})/G^2$. Since $G_1^{\cdot} < 0$ and $G > 0$ for $\varepsilon y > 0$ and $t > 0$, it suffices to show that $G - G_1 G^{\cdot}/G_1^{\cdot} < 0$. It is clear that if $k = \frac{1}{4}$, then $G - G_1 G^{\cdot}/G_1^{\cdot} = 2(1 + t^2 - e^{t^2}) < 0$ for $t > 0$. The proof of Lemma 7.1 for $k = \frac{1}{4}$ is complete.

Considering that $Q(\varepsilon, y, 0) = 0$ and $Q^{\cdot} \geqq 0$, from Lemma 7.1 we have the following corollary.

COROLLARY 7.1. *$Q(\varepsilon, y, t) > 0$ for $t > 0$ and for any $\varepsilon y > 0$.*

Remark. Lemma 7.1 is the hardest lemma to prove if $k \neq \frac{1}{4}$, but it is true for any $k > 0$.

LEMMA 7.2. *There exists a number $b=b(k)>0$ such that for $\varepsilon y \geqq d$, where d is chosen as in Theorem 2, and for $t>b$, $Q(\varepsilon, y, t)>2/3y$.*

Proof. From the remark after the proof of Lemma 6.5, $d=2$ for $k=\frac{1}{4}$; hence $3\varepsilon y-2>0$ for $\varepsilon y \geqq d$. Define

$$p(\varepsilon y, t)=3yQ(\varepsilon, y, t)/[2(1+\varepsilon y)].$$

We see that $\lim_{\varepsilon y\to\infty} p(\varepsilon y, t)=3/2$ for fixed $t>0$. From Lemma 7.1, the function $\{p(\varepsilon y, t)-1/(1+\varepsilon y)\}$ has at most a single zero if y is given. Notice that $p(\varepsilon y, 1)=3\varepsilon y/[2e(1+\varepsilon y/e)]>3/(e+2)>1/(1+\varepsilon y)$ for $\varepsilon y \geqq d$ and $p(\varepsilon y, 0)=0$, and that $p(\varepsilon y, t)>0$. Then there is a $b=b(k)$ such that $0<b<1$ and $p(\varepsilon y, t)>1/(1+\varepsilon y)$ for $t>b$ and $\varepsilon y \geqq d$, which implies Lemma 7.2.

COROLLARY 7.2. *There is an $\tilde{\varepsilon}=\tilde{\varepsilon}(k)$ such that for $\varepsilon y>d$, where ε is in $(0, \tilde{\varepsilon})$ and d is chosen as in Theorem 2,*

$$-[(1+\varepsilon)/(1+\varepsilon y)]G^{-3/2}(\varepsilon, y, v)+3\int_b^v t\,e^{t^2}QG^{-3/2}\,dt>0, \tag{7.5}$$

where b is given as in Lemma 7.2.

Proof. Let $W=(\varepsilon+1)/(\varepsilon+1/y)$ in Lemma 6.2. Since $W(b)$ depends on $b=b(k)$, and $W>(1+\varepsilon)/[\varepsilon(1/d+1)]$ for $\varepsilon y \geqq d$, there is an $\tilde{\varepsilon}=\tilde{\varepsilon}(k)$ such that $W>(1+\varepsilon)/[\varepsilon(1/d+1)]>W(b)$ for all $\varepsilon y \geqq d$, where ε is in $(0, \tilde{\varepsilon})$. Hence by Lemma 6.2,

$$\int_b^v e^{t^2}\,dt>(1+\varepsilon)/[2v(1+\varepsilon y)]. \tag{7.6}$$

Using Lemmas 7.2 and 6.1, we obtain

$$3\int_b^v tG^{-3/2}Q\,e^{t^2}\,dt>2v/[yG^{3/2}(\varepsilon, y, v)]\int_b^v e^{t^2}\,dt. \tag{7.7}$$

From (7.6), the corollary is proved.

LEMMA 7.3. *For given $\varepsilon y>0$, there exists a $\delta_1=\delta_1(\varepsilon, y, k)$ such that for $0<t<\delta_1$, where δ is in $(0, \delta_1)$, $N(\varepsilon, y, t)<0$.*

Proof. Let $c_1(\varepsilon, y)=\lim_{t\to 0} G_1/G$ and $c_2(\varepsilon, y)=\lim_{t\to 0} G_2/G$. From (7.3),

$$N(\varepsilon, y, 0)=(1+2\varepsilon y)c_1(\varepsilon, y)+y(1+\varepsilon y)[c_2(\varepsilon, y)-c_1^2(\varepsilon, y)]<0.$$

In fact $c_1(\varepsilon, y)=-\varepsilon/(1+\varepsilon y)$, $c_2(\varepsilon, y)=2\varepsilon^2/(1+\varepsilon y)^2$ and hence $N(\varepsilon, y, 0)=-\varepsilon$ for $k=\frac{1}{4}$. For $k\neq\frac{1}{4}$ a tedious calculation shows that the lemma also holds.

LEMMA 7.4. *For given $\varepsilon y>0$, there exists a $\delta_2=\delta_2(\varepsilon, y, k)$ such that for δ in $(0, \delta_2)$*

$$M(\varepsilon, y, \delta)+[3/(1+\varepsilon y)^2]\int_{b/2}^{b} t\,e^{t^2}G^{-3/2}Q\,dt>0, \tag{7.8}$$

where $b>0$ is any constant.

Proof. Let $q(\varepsilon, y, k)$ be the second term on the left side of (7.8). By Corollary 7.1, $q(\varepsilon, y, k)>0$. From (7.3),

$$2M(\varepsilon, y, \delta)\,e^{-\delta^2}=[-G_1(\varepsilon, y, \delta)/G(\varepsilon, y, \delta)+c_1(\varepsilon, y)]/\sqrt{G}(\varepsilon, y, \delta) \tag{7.9}$$

where $c_1(\varepsilon, y)$ is in the proof of Lemma 7.3. By L'Hôpital's rule, $\lim_{\delta\to 0} M(\varepsilon, y, \delta)=0$. This completes the proof of Lemma 7.4.

We can now proceed to prove Theorem 3. First fix an ε in $(0, \tilde{\varepsilon})$, where $\tilde{\varepsilon}$ is chosen in Corollary 7.2. Second, for any given $y>d/\varepsilon$, where d is in Theorem 2, take b as in Lemma 7.2 (note that b is independent of ε). Then let $\delta=\min\{\delta_1, \delta_2\}/2$, where

δ_1 and δ_2 are given in Lemmas 7.3 and 7.4 respectively. By Corollaries 7.1 and 7.2 and Lemmas 7.3 and 7.4, we prove the theorem immediately, if we rewrite (7.4) as follows:

$$
\begin{aligned}
F_\varepsilon(y) = M(\varepsilon, y, \delta) &+ [3/(1+\varepsilon y)^2] \int_{b/2}^{b} t\, e^{t^2} G^{-3/2}\, Q\, dt \\
&- (1+\varepsilon)/[(1+\varepsilon y)^3 G^{3/2}(\varepsilon, y, v)] \\
&+ [3/(1+\varepsilon y)^2] \int_{b}^{v} t\, e^{t^2} G^{-3/2} Q\, dt \\
&+ \int_0^{\delta} t\, e^{t^2}(-N)/\sqrt{G}\, dt + [3/(1+\varepsilon y)^2] \int_0^{b/2} t\, e^{t^2} G^{-3/2} Q\, dt.
\end{aligned}
\tag{7.10}
$$

8. The proof of Theorem 4. From (3.11) and $v' = 1/[2vy(1+\varepsilon y)]$,

$$
\begin{aligned}
f''_\varepsilon(y) &= -2(1+\varepsilon)/[y^2(1+\varepsilon y)\sqrt{G}(\varepsilon, y, v)] \\
&\quad - (1+\varepsilon) G_1(\varepsilon, y, v)/[y(1+\varepsilon y) G^{3/2}(\varepsilon, y, v)] \\
&\quad - 1/[y^2(1+\varepsilon y)^2 G^{3/2}(\varepsilon, y, v)] \\
&\quad + \left(\frac{4}{y^3}\right) \int_0^v t\, e^{t^2}/\sqrt{G}\, dt + \left(\frac{2}{y^2}\right) \int_0^v t\, e^{t^2} G_1 G^{-3/2}\, dt \\
&\quad - \left(\varepsilon + \frac{1}{y}\right) \int_0^v [3G_1^2/(2G^{5/2}) - G_2/G^{3/2}]\, t\, e^{t^2}\, dt.
\end{aligned}
\tag{8.1}
$$

LEMMA 8.1. *As ε tends to* 0,

$$
\int_0^v t\, e^{t^2}/\sqrt{G}\, dt \to \frac{\sqrt{2}}{2} \int_0^{\sqrt{\ln y}} e^{t^2}\, dt,
\tag{8.2}
$$

$$
\int_0^v t\, e^{t^2} G_1 G^{-3/2}\, dt \to 0,
\tag{8.3}
$$

$$
\int_0^v t\, e^{t^2}(G_1^2 G^{-5/2} - G_2 G^{-3/2})\, dt \to 0
\tag{8.4}
$$

uniformly on $[1, E]$, *where* $E > 1$ *is any constant.*

Proof. From (3.1) or (3.3), we see that

$$
\frac{\partial G}{\partial \varepsilon} = \left(\frac{\partial G}{\partial y}\right) y/\varepsilon = \frac{G_1 y}{\varepsilon} < 0
\tag{8.5}
$$

for $t, \varepsilon, y > 0$. Thus $t\, e^{t^2}/\sqrt{G}$ monotonely increases as ε decreases for given $t > 0$ and for given $t \geqq 1$. By Dini's theorem,

$$
t\, e^{t^2}/\sqrt{G} \to \sqrt{2}\, e^{t^2}/2
\tag{8.6}
$$

uniformly on any compact set of (y, t) in R^2. In addition, $v \to \ln y$ uniformly on $[1, E]$, so that (8.2) holds. From (3.5) and (3.8), both G_1/t^2 and G_2/t^2 converge to 0 as ε tends to 0 uniformly on any compact set of (y, t) in R^2. Then considering that $t\, e^{t^2} G_1/G^{3/2} = (t\, e^{t^2}/\sqrt{G})(G_1/t^2)(t^2/G)$, we get (8.3). Similarly, one can prove (8.4). Q.E.D.

We then see that in order to determine the sign of $f'_\varepsilon(y)$ when y is near to $y = 1$ the integral terms in (3.11) are not important for small ε. For the first term in (3.11) we need the following lemma.

LEMMA 8.2. *$1/G(\varepsilon, y, v)$ decreases as ε decreases, and hence for $\varepsilon > 0$,*

$$1/\sqrt{G}(\varepsilon, y, v) > 1/\sqrt{(2 \ln y)}. \tag{8.7}$$

Proof. We see that

$$\begin{aligned} \frac{\partial G(\varepsilon, y, v(\varepsilon))}{\partial \varepsilon} &= G_1(\varepsilon, y, v)y/\varepsilon + \frac{\partial G(\varepsilon, y, v)}{\partial v} \cdot \frac{\partial v}{\partial \varepsilon} \\ &= G_1(\varepsilon, y, v)y/\varepsilon + G'(\varepsilon, y, v)/(2v)[1/(1+\varepsilon) - y/(1+\varepsilon y)] \qquad (8.8) \\ &= 2(1-y)(2+\varepsilon+\varepsilon y)/[(1+\varepsilon y)^2(1+\varepsilon)^2] < 0. \end{aligned}$$

At $\varepsilon = 0$, $G(\varepsilon, y, v) = 2 \ln y$, which implies the lemma.

LEMMA 8.3. *There exist two numbers E_1 and α_1 such that for y in $(1, E_1]$ and for α in $(0, \alpha_1]$*

$$1/[y(1+\alpha)\sqrt{(2 \ln y)}] - (2/y^2) \int_0^{\sqrt{\ln y}} e^{t^2}\, dt - 2\alpha > 0. \tag{8.9}$$

Proof. Denote the left side of (8.9) by $j(y, \alpha)$. Obviously, $j(y, \alpha)$ is continuous in $(1, E) \times (0, D)$ where $E > 1$ and $D > 0$ are constants. Because $\lim_{y \to 1} j(y, 0) = +\infty$ as y approaches 0, such an α_1 and E_1 exist. The lemma is proved.

LEMMA 8.4. *There exists an $\varepsilon_{-1} = \varepsilon_{-1}(k)$ such that for ε in $(0, \varepsilon_{-1})$ and y in $(1, E_1]$, where E_1 is chosen as in Lemma 8.3, $f'_\varepsilon(y) > 0$.*

Proof. By Lemma 8.1, there exists an $\varepsilon_{-1} = \varepsilon_{-1}(E_1, k)$ such that

$$\left(\frac{2}{y^2}\right) \int_0^v t e^{t^2}/\sqrt{G}\, dt < \alpha_1 + \left(\frac{\sqrt{2}}{y^2}\right) \int_0^{\sqrt{\ln y}} e^{t^2}\, dt, \tag{8.10}$$

$$\left(\varepsilon + \frac{1}{y}\right) \int_0^v t e^{t^2} G_1 G^{-3/2}\, dt < \alpha_1 \tag{8.11}$$

and $(1+\varepsilon)/[y(1+\varepsilon y)] > 1/[y(1+\alpha_1)]$ for ε in $(0, \varepsilon_{-1})$ and y in $(1, E_1]$, where α_1 is chosen as in Lemma 8.3. Then from (3.11), (8.10) and (8.9), we obtain the lemma.

LEMMA 8.5. *$f'_\varepsilon(y)$ and $f''_\varepsilon(y)$ converge to $f'_0(y)$ and $f''_0(y)$ respectively as ε tends to 0 uniformly on $[E_1, Y_0]$, where E_1 is given as in Lemma 8.4 and Y_0 in Theorem 2.*

Proof. We see that $(1+\varepsilon)/[y(1+\varepsilon y)\sqrt{G}(\varepsilon, y, v)]$ converges to $y/\sqrt{(2 \ln y)}$ uniformly on $[E_1, Y_0]$ as ε tends to 0; the rest follows by Lemma 8.1. Q.E.D.

COROLLARY 8.1. *As ε approaches 0, $F_\varepsilon(y)$ converges to $F_0(y)$ uniformly on $[E_1, Y_0]$.*

From Corollary 8.1, there is an $\varepsilon_{-2} = \varepsilon_{-2}(k)$ such that for (ε, y) in $(0, \varepsilon_{-2}) \times [E_1, Y_0]$, $F_\varepsilon(y) > 0$. To see this we take $\delta = (2 \ln Y_0)^{-3/2}/2$ in the definition of uniform convergence. Then such an ε_{-2} exists and for (ε, y) in $(0, \varepsilon_{-2}) \times [E_1, Y_0]$,

$$\begin{aligned} F_\varepsilon(y) < F_0(y) + \delta &= -(2 \ln y)^{-3/2} + (2 \ln Y_0)^{-3/2}/2 \\ &< -(2 \ln Y_0)^{-3/2}/2. \end{aligned}$$

Now, combining this fact and Lemma 8.4 and taking $\varepsilon_4 = \min\{\varepsilon_{-1}, \varepsilon_{-2}\}$, we have proved Theorem 4.

9. Comments. (i) If k is not equal to $\frac{1}{4}$, most calculations are very tedious. For example, the proof of Lemma 7.1 takes about 15 pages of computation! Nevertheless, the functions which appear in the above proofs for general $k > 0$ still have all the same properties as for $k = \frac{1}{4}$. Much of this has been checked using the symbolic manipulator Macsyma [7].

(ii) If we make a change of parameters as in [1], i.e., define $\beta = s_0/\sqrt{\sigma}$, $\varepsilon = 1/\sqrt{\sigma}$, then the corresponding problem becomes (1.3) with $u(0) = u(1) = 1$. The response function $f_\varepsilon(y, \beta)$ is a function of both y and β. In this case we need to determine the number of solutions of the equation

$$f_\varepsilon(y, \beta) = \frac{\sqrt{2}}{(4\beta\sqrt{k})}. \tag{9.1}$$

We may apply the same ideas to prove that, for given $k > 0$ and for given $\beta > 0$, there is an $\varepsilon_0 = \varepsilon_0(k, \beta)$ such that $f_\varepsilon(y, \beta)$ has exactly two critical points for ε in $(0, \varepsilon_0)$. Hence (1.3) admits no more than three solutions. This is to say that for σ large enough and s_0 conveniently chosen, problem (1.2) admits at most three solutions.

(iii) We can also study the stability of the steady states associated with problem (1.1) in the case where s_0 is large by using the Conley Index Theory or Morse Index Theorem. The conclusion is that if there exist three steady state solutions, then two of them are stable and the other is unstable (cf. [7]).

Acknowledgment. I wish to thank my advisor, Professor Stuart P. Hastings, for the suggestions he contributed to the thesis upon which this paper is based.

REFERENCES

[1] C. M. Brauner and B. Nicolaenko, *Singular perturbation, multiple solutions, and hysteresis in a nonlinear problem*, in Singular Perturbation Bound. Layer Theory, Proc. Cong. Lyin 1976, Lecture Notes in Math. 594, Springer, Berlin-New York, 1977.

[2] ———, *Perturbation singulière, solutions multiples et hystérésis dans un problème de biochimie*, C.R. Acad. Sci. Paris Sér. I Math., 283 (1976), pp. 775-778.

[3] M. C. Duban, *Hystérésis, oscillation et struturation en espace dans des systèmes biochemiques distribués*, Thèse de 3ième cycle, Compeigne 1975.

[4] S. P. Hastings and J. B. McLeod, *The number of solutions of an equation from catalysis*, MRC Tech. Summary Report #2579, Math. Research Center, Univ. of Wisconsin, 1983.

[5] J. P. Kernevez and D. Thomas, *Numerical analysis and control of some biochemical systems*, Appl. Math. Optim., 1 (1975), pp. 222-285.

[6] J. P. Kernevez, *Enzyme Mathematics*, North-Holland, Amsterdam-New York, 1980.

[7] C. Lu, *Multiple steady states and their stability in a biochemical system*, Ph.D. dissertation, State Univ. of New York, Buffalo, NY, February 1986.

[8] A. Naparstek, J. L. Romette, J. P. Kernevez and D. Thomas, *Memory in enzyme membranes*, Nature, 249 (1974), pp. 490-491.

[9] J. Smoller and A. Wasserman, *Global bifurcation of steady-state solutions*, J. Differential Equations, 39 (1981), pp. 269-290.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

019

WEAK SOLUTIONS TO STEFAN PROBLEMS WITH PRESCRIBED CONVECTION*

JIM RULLA†

Abstract. This paper deals with the equation

$$w_t + \operatorname{div}(\mathbf{v}w - \nabla\alpha(w)) = g,$$

which is to hold in a smooth, bounded domain $G \subset \mathbb{R}^n$. The function $\alpha: \mathbb{R} \to \mathbb{R}$ is uniformly Lipschitz and nondecreasing, but may be identically zero. For certain smooth functions $\mathbf{v}: \bar{G} \to \mathbb{R}^n$ satisfying $\operatorname{div}(\mathbf{v}) \geqq 0$, there are integral solutions to the Cauchy problem associated with this equation, provided the boundary conditions on w are chosen appropriately. We prove that these integral solutions are weak solutions in the usual sense. Moreover, these weak solutions are unique; hence the notion of a weak solution is adequate for problems of this type.

Key words. Stefan problem, free boundary, convection, uniqueness

AMS(MOS) subject classifications. 35D05, 35K65, 35R35

1. Introduction and notation. Consider a mixture of crude oil and melted paraffin flowing through a pipe. As the temperature of the mixture falls, the paraffin solidifies, thus changing the thermal properties of the mixture. We shall develop a heat equation modeling this problem, and we shall prove existence and uniqueness of solutions (in an appropriate sense) to the evolution equation. In [5], Fasano, Primicerio and Rubenstein address this problem classically and obtain solutions if the free boundary is assumed to satisfy a simplifying assumption. Visintin, in [12], considers a similar class of equations and obtains existence results for weak solutions. His solutions are weaker than ours, however. In particular, we prove that solutions are bounded and Lipschitz continuous: $[0, T] \to L^1(G)$ (so the notion of an initial value makes sense in $L^1(G)$) while the similar class of problems yields solutions which are only measurable into L^∞ (hence the notion of initial value must be weakened). By interpolation, we prove that solutions are weak solutions in an L^2 sense. Moreover, these weak solutions are unique provided the data in the problem are appropriately regular; consequently, the notion of a weak solution is appropriate for this problem.

In § 2 of this paper, we present a brief derivation of the Stefan problem. In § 3, we find solutions to an appropriate stationary problem. The operator involved is a degenerate elliptic operator, and existence of solutions depends strongly on our making a suitable choice of boundary conditions. The properties of the solutions which we shall need follow from the uniqueness of solutions, which must be proven independently from the existence argument. In § 4, we prove the existence of integral solutions to the evolution equation and we show that these integral solutions are actually weak solutions. Finally, we prove in § 5 the uniqueness of these weak solutions. This proof mimics the uniqueness proof from § 3.

Let $G \subseteq \mathbb{R}^n$ be a bounded domain with smooth (C^2) boundary, ∂G, which we denote by S. We let $L^p(G)$ and $H^m(G)$ denote the usual Lebesgue and Hilbert spaces of measurable functions on G, and we omit the reference to the set G when such reference is unnecessary. We use the notation $C([0, T]; L^p)$ to denote the space of continuous functions from $[0, T]$ into L^p.

* Received by the editors February 3, 1986; accepted for publication (in revised form) December 15, 1986.

† Department of Mathematics, University of Texas, Austin, Texas 78712. Present address, Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

We denote the spatial divergence and gradient operators, respectively, by div $(\cdot)$ and ∇. The spatial Laplace operator is denoted by Δ. A natural domain for the divergence operator is the space $\mathscr{E}$ defined by

$$\mathscr{E}=\{\mathbf{v}\in(L^2)^n \mid \operatorname{div}(\mathbf{v})\in L^2\}.$$

The space $(L^2)^n$ here is the n-fold product of $L^2(G)$. The trace operators γ_0 and γ_ν, which represent restricting scalar and normal components of vector valued functions, on G to S, respectively, are then related by

$$\int_G [\operatorname{div}(\mathbf{v})u+\mathbf{v}\cdot\nabla u]=\langle\gamma_\nu(\mathbf{v}),\ \gamma_0(u)\rangle$$

for all $\mathbf{v}\in\mathscr{E}$ and $u\in H^1$. The duality on the right takes place between $H^{1/2}(S)$ and its dual (cf. [11]). If $\boldsymbol{\nu}:S\to\mathbb{R}^n$ is the unit outward normal on S, then

$$\langle\gamma_\nu(\mathbf{v}),\ \gamma_0(u)\rangle=\int_S(\boldsymbol{\nu}\cdot\mathbf{v})u$$

provided $\mathbf{v}$ and u are smooth enough. In addition, $C^\infty(\bar{G})$ is dense in $\mathscr{E}$ (when $\mathscr{E}$ is supplied with its natural topology).

We shall split S into disjoint measurable subsets S_1 and S_2. If S_1 has positive (surface) measure, then the space

$$V\equiv\{v\in H^1(G) \mid \gamma_0(v)=0 \text{ a.e. (w.r.t. surface measure) on } S_1\}$$

is a Hilbert space with inner product

$$(u,v)_V\equiv\int_G\nabla u\cdot\nabla v.$$

We shall make use of the notation

$$\langle u,v\rangle\equiv\int_G uv$$

whenever the product $uv\in L^1$. We use the same notation if $\mathbf{u}\cdot\mathbf{v}\in L^1$, i.e.,

$$\langle\mathbf{u},\mathbf{v}\rangle=\int_G\mathbf{u}\cdot\mathbf{v}.$$

We conclude this section by recalling that a subset $\mathscr{A}$ of $L^1\times L^1$ is *accretive* provided

$$\|x_1-x_2\|_{L^1}\leqq\|(x_1+\lambda y_1)-(x_2+\lambda y_2)\|_{L^1}$$

for all pairs $[x_i,y_i]\in\mathscr{A}$ $(i=1,2)$ and all (sufficiently small) $\lambda>0$. This is equivalent to having

$$\int_G(y_1-y_2)\sigma\geqq0,$$

where σ is some measurable function on G chosen so that

$$\sigma\in\operatorname{sgn}(x_1-x_2)\quad\text{a.e. in } G.$$

The signum relation is given by

$$\operatorname{sgn}(s)=\begin{cases}\{1\}, & s>0,\\ [-1,1], & s=0,\\ \{-1\}, & s<0.\end{cases}$$

We view the set $\mathscr{A}$ as the graph of an operator, which we denote by $\mathscr{A}$. The operator $\mathscr{A}$ is *m-accretive* if, for every $\lambda > 0$ and every $f \in L^1$, there is a solution u to

$$u + \lambda \mathscr{A}(u) \ni f.$$

For a thorough treatment of m-accretive operators and the semigroups they generate, see [1, Chaps. 2 and 3]. We denote the *zero section* of a graph by a subscripted zero. If β is a graph, then $\beta_0(s)$ is the element of the set $\beta(s)$ which is closest to zero. For example,

$$\operatorname{sgn}_0(s) = \begin{cases} 1, & s > 0, \\ 0, & s = 0, \\ -1, & s < 0. \end{cases}$$

2. The formulation of the Stefan problem. We consider a fluid flowing through a bounded domain, $G \subseteq \mathbb{R}^n$, with a prescribed velocity $\mathbf{v} \in C^1(\bar{G}; \mathbb{R}^n)$. We shall make further restrictions on the velocity in § 3. Let $u(x)$ denote the temperature and $w(x)$ the enthalpy or heat energy of the fluid at the point $x \in G$. When the fluid changes phase, the enthalpy increases by an amount L per unit mass, while the temperature remains constant. The quantity L is the latent heat of the material. This fact suggests that the enthalpy, not the temperature, is the natural function to consider in problems of heat transfer (cf. [10]). If $c = c(u)$ is the heat capacity of the material (which depends on the phase and may be discontinuous) and $\xi \in [0, 1]$ is the fraction of the material in the more energetic phase, then a unit mass of the material has enthalpy

$$w = c(u)u + \xi L.$$

If the phase change occurs when $u = u_0$, then $\xi \in H(u - u_0)$, where $H(x) = \operatorname{sgn}^+(x)$ is the Heaviside graph. We may then write

$$w \in c(u)u + LH(u - u_0),$$

so w is given as a nondecreasing relation of u. In fact, we may invert this relation to obtain the temperature as a nondecreasing function of the enthalpy:

$$u = \theta(w).$$

If $k = k(u)$ denotes the conductivity of the material (which again depends on the phase and may be discontinuous), then the flux or flow of energy at any point in G is given by

$$\begin{aligned} \text{flux} &= \mathbf{v}w - k(u)\nabla u \\ &= \mathbf{v}w - \nabla \alpha(w), \end{aligned}$$

where $\alpha : \mathbb{R} \to \mathbb{R}$ is the antiderivative of $k \circ \theta$ with $\alpha(0) = 0$. We remark that α is nondecreasing and satisfies $\alpha'(s) = 0$ for $s \in [\theta_0^{-1}(u_0), \theta_0^{-1}(u_0) + L]$. The temperature u and $\alpha(w)$ are related by a one-to-one correspondence and we abuse the distinction by referring to $\alpha(w)$ as the temperature. We shall, naturally, require that $\alpha(w) \in H^1$. Notice that this requirement may force $\alpha(w) \notin H^2$ and $w \notin H^1$. In fact, since energy is conserved, the free boundary (where the two phases of material meet) advances at a speed proportional to the jump in the normal component of the flux across the boundary. In [8], Rubenstein comments on the adequacy of this model.

Let $\Omega = G \times (0, T)$ and divide Ω into three regions:

$$\Omega_0 = \{\alpha(w(x, t)) = u_0\},$$

$$\Omega_1 = \{\alpha(w(x, t)) < u_0\},$$

$$\Omega_2 = \{\alpha(w(x, t)) > u_0\}.$$

In each of these regions, energy is conserved, so

$$\frac{\partial w}{\partial t}+\operatorname{div}(\mathbf{v}w-\nabla\alpha(w))=g \quad \text{in } \Omega_j, \qquad j=0,1,2,$$

where g represents any external heat sources or sinks. If we multiply by a smooth function $\varphi\in C^\infty(\bar\Omega)$ and formally integrate by parts over each of the three regions, we get

$$\begin{aligned}\int_{\Omega_0}+\int_{\Omega_1}+\int_{\Omega_2}&\left[\frac{\partial w}{\partial t}+\operatorname{div}(\mathbf{v}w-\nabla\alpha(w))\right]\varphi\\ &=-\int_{\Omega}\left[w\frac{\partial\varphi}{\partial t}+\mathbf{v}w\cdot\nabla\varphi-\nabla\alpha(w)\cdot\nabla\varphi\right]\\ &\quad+\int_{\partial\Omega_1}+\int_{\partial\Omega_2}+\int_{\partial\Omega_3}\mathbf{N}\cdot(\mathbf{v}w-\nabla\alpha(w),w)\varphi.\end{aligned}$$

The vector N is the unit normal out of each of the regions $\Omega_j, j=0,1,2$, and $(\cdot,\cdot)$ denotes vectors in $\mathbb{R}^n\times\mathbb{R}$. The jump condition on the flux is equivalent to having the boundary integrals make a nonzero contribution only on $\partial\Omega$, not on any of the interior interfaces. We conclude that

$$\begin{aligned}&\int_{\Omega} g\varphi+w\frac{\partial\varphi}{\partial t}+\mathbf{v}w\cdot\nabla\varphi-\nabla\alpha(w)\cdot\nabla\varphi\\ &\quad=\int_0^T\int_{\partial G}(\mathbf{v}w-\nabla\alpha(w))\cdot\boldsymbol{\nu}\varphi+\int_G[\varphi(\cdot,T)w(\cdot,T)-\varphi(\cdot,0)w(\cdot,0)].\end{aligned}\tag{2.1}$$

If $\varphi\in C_0^\infty(\Omega)$, then we are led to the differential equation

$$\frac{\partial w}{\partial t}+\operatorname{div}(\mathbf{v}w-\nabla\alpha(w))=g.\tag{2.2a}$$

It is important to choose the boundary conditions on w carefully in order to be able to solve this problem. We let $S=S_1\cup S_2$, where S_1 and S_2 are disjoint, measurable subsets of $S=\partial G$. We require that

$$S_1\subseteq\{s\in S\,|\,\boldsymbol{\nu}(s)\cdot\mathbf{v}(s)\geqq 0\}$$

and

$$S_2\subseteq\{s\in S\,|\,\boldsymbol{\nu}(s)\cdot\mathbf{v}(s)\leqq 0\}$$

and that the measure of S_1 be nonzero. For a specified $g_1\in H^1$ and $\mathbf{g}_2\in\mathscr{C}$, we shall require that

$$\begin{aligned}&\gamma_0(\alpha(w))=\gamma_0(g_1)\quad\text{a.e. on } S_1,\\ &\langle\gamma_\nu(\mathbf{g}_2-(\mathbf{v}w-\nabla\alpha(w))),\gamma_0(v)\rangle=0\quad\text{for all } v\in V.\end{aligned}\tag{2.2b}$$

(Recall that $V=\{v\in H^1\,|\,\gamma_0(v)=0 \text{ a.e. on } S_1\}$.) Notice that we specify the *temperature* on S_1 and the normal component of the *flux* on S_2.

The last piece of data which we need to make the Cauchy problem well posed is the initial enthalpy

$$w(x,0)=w_0(x).\tag{2.2c}$$

3. The stationary problem. A standard approach to solving the Cauchy problem (2.2) is to define an operator $\mathcal{A}$ corresponding to the spatial derivatives $\operatorname{div}(\mathbf{v}(\cdot)-\nabla\alpha(\cdot))$ on an appropriate domain $D(\mathcal{A})$ of functions in some Banach space X which satisfy the boundary conditions (2.2b). One then proves that $\mathcal{A}$ is m-accretive on X and then applies the generation theorems of Crandall and Liggett [4] or Benilan [2], [1] to obtain "solutions" $w:[0, T]\to X$ to

$$
\begin{aligned}
&\frac{dw(t)}{dt}+\mathcal{A}(w(t))\ni g(t),\\
&w(0)=w_0.
\end{aligned}
\tag{3.1}
$$

If X is reflexive and the data $g(t)$ and w_0 are sufficiently well behaved, then $w(t)\in D(\mathcal{A})$ a.e. on $[0, T]$ and w satisfies (3.1) a.e. in $[0, T]$ (cf. [1]). It is easy to see by a formal computation, however, that $\mathcal{A}$ cannot be accretive on L^p for $p>1$. Since L^1 is not reflexive, the above regularity result fails, and "solutions" to (3.1) may satisfy $w(t)\notin D(\mathcal{A})$ for all $t\in[0, T]$. In addition, it is difficult to identify $D(\mathcal{A})$ if $G\subseteq\mathbb{R}^n$ with $n>1$. These observations motivate our search for weak solutions to (2.1), i.e., functions $w\in C([0, T]; L^2)$ satisfying $\alpha(w)-g_1\in L^2([0, T]; V)$ and

$$
\begin{aligned}
&\int_0^T \langle g,\varphi\rangle+\left\langle w,\frac{d\varphi}{dt}+\mathbf{v}\cdot\nabla\varphi\right\rangle-\langle\nabla\alpha(w),\nabla\varphi\rangle\,dt\\
&\qquad=\langle w(T),\varphi(T)\rangle-\langle w(0),\varphi(0)\rangle+\int_0^T\langle\gamma_\nu(\mathbf{g}_2),\gamma_0(\varphi)\rangle\,dt
\end{aligned}
\tag{3.2}
$$

for all appropriately smooth testing functions $\varphi:[0, T]\to V$.

In order to prove the existence of solutions to (3.2) we employ the following plan:

(1) Define an operator in $L^2\times L^2$ and prove that this operator is accretive when considered as a subset of $L^1\times L^1$.

(2) Prove that the closure of the operator in (1) is m-accretive in $L^1\times L^1$.

(3) Prove that the integral solutions generated by the operator in (2) are weak solutions (solutions to (3.2)).

Step (1) is similar in nature to the technique used in [3]. The uniqueness proof used in Step (1) contains the idea of the proof of uniqueness of weak solutions. Steps (2) and (3) consist primarily of applying the estimates found in Theorem 1 in an obvious way to obtain the necessary estimates.

Throughout the remainder of this discussion, we shall assume that the following "regularity" conditions are met:

The function $\alpha:\mathbb{R}\to\mathbb{R}$ is uniformly Lipschitz continuous with a derivative at all but finitely many points.

The functions $g_1, \mathbf{g}_2$ and g satisfy $g_1\in H^1(G)$, $\mathbf{g}_2\in\mathscr{E}$ and $g\in L^1([0, T]; L^\infty(G))$. We shall assume that $\gamma_0(g_1)\in L^\infty(S_1)$, that $\gamma_0(g_1)\in$ Range of (α) a.e. on S_1 and that $\gamma_0(g_1)$ satisfies $\|\alpha_0^{-1}(\gamma_0(g_1))\|_{L^\infty(S_1)}<\infty$, where α_0^{-1} is the zero section of the graph of α^{-1}. We also suppose that there is a constant M chosen so that

$$
\langle\gamma_\nu(\mathbf{g}_2),\gamma_0(v)\rangle\leqq M|\langle\gamma_\nu(\mathbf{v}),\gamma_0(v)\rangle|\quad\forall v\in V.
\tag{3.3}
$$

In particular, if $\mathbf{g}_2$ is smooth, this means that

$$
|\boldsymbol{\nu}\cdot\mathbf{g}_2|\leqq M|\boldsymbol{\nu}\cdot\mathbf{v}|\quad\text{a.e. on } S_2.
$$

The function $\mathbf{v}\in C^1(\bar{G};\mathbb{R}^n)$ satisfies $\operatorname{div}(\mathbf{v})\geqq 0$ in G. This means that the fluid may be incompressible ($\operatorname{div}(\mathbf{v})=0$) or that there may be sources of flow ($\operatorname{div}(\mathbf{v})\geqq 0$). Since the fluid will adhere to the sides of the pipe, we expect that $\mathbf{v}=\mathbf{0}$ on that portion of

S corresponding to the pipe. In fact, we shall only require that $\mathbf{v}=\mathbf{0}$ on the set $\partial S_1=\partial S_2$. We define the operator

$$\mathcal{A}_2(w)=\operatorname{div}(\mathbf{v}w-\nabla\alpha(w))$$

on the set

$$D(\mathcal{A}_2)=\{w\in L^2\,|\,\alpha(w)-g_1\in V,\ \mathcal{A}_2(w)\in L^2,\text{ and}$$
$$\langle\gamma_\nu(\mathbf{v}w-\nabla\alpha(w)-\mathbf{g}_2),\ \gamma_0(v)\rangle=0\ \forall v\in V\}.$$

THEOREM 1. *There is a constant* $\delta=\delta(\mathbf{v})$ *so that if* $\lambda\leqq\delta$ *and* $f\in L^\infty$, *then there is a unique* $w\in D(\mathcal{A}_2)$ *satisfying*

$$w+\lambda\mathcal{A}_2(w)=f \tag{3.4}$$

(*notice that* $\mathcal{A}_2$ *is single valued*; *hence equality holds rather than containment*). *If* w_1 *and* w_2 *satisfy* (3.4) *with* f *replaced by* f_1 *and* f_2, *respectively, then*

$$\|(w_1-w_2)^+\|_{L^1}\leqq\|(f_1-f_2)^+\|_{L^1}. \tag{3.5}$$

In particular, $\mathcal{A}_2$ *is accretive in* $L^1\times L^1$ *and the resolvent* $(I+\lambda\mathcal{A}_2)^{-1}$ *is order preserving. Finally, if* $\beta=\alpha^{-1}$,

$$K_1\equiv\min\left\{\beta_0\left(\operatorname*{ess\,inf}_{S_1}(\gamma_0(g_1))\right),-M\right\}\leqq 0$$

and

$$K_2\equiv\max\left\{\beta_0\left(\operatorname*{ess\,sup}_{S_1}(\gamma_0(g_1))\right),M\right\}\geqq 0,$$

then

$$\min\left\{\operatorname*{ess\,inf}_G(f),K_1\right\}\leqq w\leqq\max\left\{\operatorname*{ess\,sup}_G(f),K_2\right\} \tag{3.6}$$

a.e. in G.

Proof. We prove uniqueness of solutions first. Uniqueness of solutions to (3.4) follows from (3.5), but estimate (3.5) is established assuming that solutions to (3.4) are unique; therefore, we require an independent uniqueness proof.

Let w_1 and $w_2\in D(\mathcal{A}_2)$ satisfy (3.4). Then

$$(w_1-w_2)+\lambda\operatorname{div}(\mathbf{v}(w_1-w_2)-\nabla(\alpha(w_1)-\alpha(w_2)))=0$$

and $\alpha(w_1)-\alpha(w_2)\in V$. For any testing function $\varphi\in V$ we have

$$0=\int_G(w_1-w_2)\varphi-\lambda[\mathbf{v}(w_1-w_2)-\nabla(\alpha(w_1)-\alpha(w_2))]\cdot\nabla\varphi \tag{3.7}$$

since the boundary term in the integration by parts formula is zero. The proof of uniqueness follows by choosing an appropriate testing function φ. Readers familiar with the duality between H_0^1 and H^{-1} will recognize the similarity between the techniques used there and here.

Let $\varphi\in V$ be the solution to the mixed boundary value problem

$$-\Delta\varphi=w_1-w_2\in L^2,\qquad\frac{\partial\varphi}{\partial\nu}=0\quad\text{a.e. on }S_2.$$

Then $\varphi \in V$ and for all $v \in V$,

$$\int_G \nabla\varphi\cdot\nabla v=\int_G (w_1-w_2)v.$$

This problem is well posed (cf. [9]) and so $w_1=w_2$ if and only if $\varphi\equiv 0$.

With this choice of φ, (3.7) becomes

$$0=\int_G |\nabla\varphi|^2+\lambda(\Delta\varphi\mathbf{v})\cdot\nabla\varphi+\lambda(\alpha(w_1)-\alpha(w_2))(w_1-w_2) \tag{3.8}$$

since $\alpha(w_1)-\alpha(w_2)\in V$. The last term is nonnegative a.e. because α is nondecreasing. By the regularity theory for elliptic operators (cf. [9]), $\varphi\in H^2(G')$ for any $G'\subset G$ whose intersection with a neighborhood of ∂S_1 is void. If ψ is a smooth function in G with support $\subset \bar{G}'$, then $\psi\mathbf{v}\cdot\nabla\varphi\in H^1(G)$ and

$$\int_G \Delta\varphi(\psi\mathbf{v}\cdot\nabla\varphi)=\langle\gamma_\nu(\nabla\varphi),\gamma_0(\psi\mathbf{v}\cdot\nabla\varphi)\rangle-\int_G \nabla\varphi\cdot\nabla(\psi\mathbf{v}\cdot\nabla\varphi).$$

The boundary term is

$$\int_S (\boldsymbol{\nu}\cdot\nabla\varphi)\psi\mathbf{v}\cdot\nabla\varphi=\int_{S_1}(\boldsymbol{\nu}\cdot\nabla\varphi)\psi\mathbf{v}\cdot\nabla\varphi$$

because $\boldsymbol{\nu}\cdot\nabla\varphi=0$ on $S_2\cap\bar{G}'$. Suppose for the moment that φ is smooth in a neighborhood of $S_1\cap\bar{G}'$. Then $\nabla\varphi$ lies in the direction of $\pm\boldsymbol{\nu}$ on $\{s\in S_1|\nabla\varphi(s)\neq\mathbf{0}\}$ because $\varphi=0$ on S_1. In this case,

$$\int_S (\boldsymbol{\nu}\cdot\nabla\varphi)\psi\mathbf{v}\cdot\nabla\varphi=\int_{S_1}\boldsymbol{\nu}\cdot\mathbf{v}|\nabla\varphi|^2\psi.$$

This equality remains valid for $\varphi\in V\cap H^2(G')$ with $\partial\varphi/\partial\nu=0$ on S_2 by a simple approximation argument.

From the above computation, we have

$$\int_G \Delta\varphi(\psi\mathbf{v}\cdot\nabla\varphi)=\int_{S_1}\boldsymbol{\nu}\cdot\mathbf{v}|\nabla\varphi|^2\psi-\int_G \varphi_j[(\psi\mathbf{v}^i)_j\varphi_i+\psi\mathbf{v}^i\varphi_{ij}],$$

where $\mathbf{v}=(v^1,\cdots,v^n)$, $\varphi_j=\partial\varphi/\partial x_j$, and the summation over $1\leqq i,j\leqq n$ is implicit. The last integrand is

$$\varphi_j\varphi_{ij}\psi\mathbf{v}^i=\varphi_j\varphi_{ji}\psi\mathbf{v}^i=\tfrac{1}{2}\psi\mathbf{v}^i\,\partial_i(\varphi_j)^2=\tfrac{1}{2}\psi\mathbf{v}\cdot\nabla|\nabla\varphi|^2.$$

We integrate by parts to get

$$\begin{aligned}\int_G \varphi_j\psi\mathbf{v}^i\varphi_{ij}&=\int_S \frac{1}{2}\psi\mathbf{v}\cdot\boldsymbol{\nu}|\nabla\varphi|^2-\int_G \frac{1}{2}\operatorname{div}(\psi\mathbf{v})|\nabla\varphi|^2\\ &\leqq\int_{S_1}\frac{1}{2}\psi\mathbf{v}\cdot\boldsymbol{\nu}|\nabla\varphi|^2-\int_G \frac{1}{2}\operatorname{div}(\psi\mathbf{v})|\nabla\varphi|^2,\end{aligned}$$

provided $\psi\geqq 0$, since $\mathbf{v}\cdot\boldsymbol{\nu}\leqq 0$ on S_2. With this estimate, we have

$$\begin{aligned}\int_G \Delta\varphi(\psi\mathbf{v}\cdot\nabla\varphi)&\geqq\int_G \frac{1}{2}\operatorname{div}(\psi\mathbf{v})|\nabla\varphi|^2-(\psi\mathbf{v}^i)_j\varphi_j\varphi_i+\int_{S_1}\frac{1}{2}\boldsymbol{\nu}\cdot\mathbf{v}|\nabla\varphi|^2\psi\\ &\geqq\int_G \frac{1}{2}\operatorname{div}(\psi\mathbf{v})|\nabla\varphi|^2-(\psi\mathbf{v}^i)_j\varphi_j\varphi_i,\end{aligned}$$

provided $\psi\geqq 0$, because $\boldsymbol{\nu}\cdot\mathbf{v}\geqq 0$ on S_1.

Let $\varepsilon > 0$ and define

$$\sigma_\varepsilon^+(s) = \begin{cases} 1, & s > \varepsilon, \\ s/\varepsilon, & 0 \leqq s \leqq \varepsilon, \\ 0, & s < 0. \end{cases}$$

If we let $\psi \in \sigma_\varepsilon^+(|\mathbf{v}|^2 - \varepsilon)$, then the above calculations are valid, and $\psi\mathbf{v} \to \mathbf{v}$ boundedly in $\bar{G}$ as $\varepsilon \downarrow 0$. We calculate

$$\partial_j(\psi \mathbf{v}^i) = \frac{2\mathbf{v} \cdot \mathbf{v}_j \mathbf{v}^i}{\varepsilon} \chi_{\{0 \leqq |\mathbf{v}|^2 - \varepsilon \leqq \varepsilon\}} + \psi \mathbf{v}_j^i.$$

Both terms remain bounded and the sum tends to $\mathbf{v}_j^i$ a.e. as $\varepsilon \downarrow 0$. We apply the dominated convergence theorem to verify

$$\begin{aligned} \int_G \Delta\varphi(\mathbf{v} \cdot \nabla\varphi) &\geqq \int_G \frac{1}{2} \operatorname{div}(\mathbf{v})|\nabla\varphi|^2 - \mathbf{v}_j^i \varphi_i \varphi_j \\ &\geqq -K(\mathbf{v}) \int_G |\nabla\varphi|^2. \end{aligned} \tag{3.9}$$

We return now to estimate (3.8). Using (3.9), we have

$$0 \geqq \int_G |\nabla\varphi|^2 - \lambda K(\mathbf{v})|\nabla\varphi|^2 \geqq \frac{1}{2}\|\varphi\|_V^2$$

provided $\lambda K(\mathbf{v}) \leqq \frac{1}{2}$. We conclude that $\varphi \equiv 0$; hence $w_1 = w_2$.

Proof of existence. It is convenient to work with temperature rather than with enthalpy when proving existence of solutions. Let $\beta = \alpha^{-1}$ and assume for the moment that β is Lipschitz with Lipschitz constant μ. We begin by establishing estimates which are independent of μ.

Let $u = \alpha(w)$, so $w = \beta(u)$. Then $w \in D(\mathcal{A}_2)$ is a solution to (3.4) if and only if

$$u - g_1 \in V, \tag{3.10a}$$

$$\mathbf{v}\beta(u) - \nabla u \in \mathscr{E}, \tag{3.10b}$$

and

$$b(u, v) = \langle f, v \rangle \quad \text{for all } v \in V, \tag{3.10c}$$

where $b(u, v) \equiv \int_G [\beta(u)(v - \lambda\mathbf{v} \cdot \nabla v) + \lambda \nabla u \cdot \nabla v] + \lambda\langle \gamma_\nu(\mathbf{g}_2), \gamma_0(v)\rangle$.

LEMMA. *If* $f \in L^\infty$ *and* u *satisfies* (3.10), *then*

$$\min\left\{\operatorname*{ess\,inf}_G (f), K_1\right\} \leqq \beta(u) \leqq \max\left\{\operatorname*{ess\,sup}_G (f), K_2\right\} \quad \textit{a.e. in } G.$$

Proof. We shall establish the right inequality; the left inequality is proved in a similar manner.

Let K be chosen so that $\beta(K) \geqq K_2$, and set $v = \sigma_\varepsilon^+(u - K)$. Then $v \in V$ (cf. [6, p. 54]) and

$$\begin{aligned} \int_G (f - \beta(K))v = \int_G &\Big[(\beta(u) - \beta(K))v - \lambda\mathbf{v}[\beta(u) - \beta(K)] \cdot \nabla v + \lambda\nabla(u - K) \cdot \nabla v\Big] \\ &- \lambda \int_G [\mathbf{v}\beta(K) \cdot \nabla v] + \lambda\langle \gamma_\nu(\mathbf{g}_2), \gamma_0(v)\rangle. \end{aligned}$$

We calculate

$$\nabla v = (\nabla(u-K)/\varepsilon)\chi_{E_\varepsilon} \quad \text{a.e. in } G,$$

where χ_{E_ε} is the characteristic function of the set $E_\varepsilon = \{x \in G \mid 0 < u(x) - K < \varepsilon\}$. If we integrate the term in the second integral on the right by parts, we get

$$\begin{aligned}\|(f-\beta(K))^+\|_{L^1} &\geqq \int_G (f-\beta(K))\sigma_\varepsilon^+(u-K)\\ &\geqq \int_G \Big[(\beta(u)-\beta(K))v - \lambda\mathbf{v}\Big[\frac{\beta(u)-\beta(K)}{\varepsilon}\Big]\nabla(u-K)\chi_{E_\varepsilon}\\ &\qquad + \lambda\beta(K)\operatorname{div}(\mathbf{v})v\Big]\\ &\qquad + \lambda\langle\gamma_\nu(\mathbf{g}_2 - \mathbf{v}\beta(K)), \gamma_0(v)\rangle.\end{aligned}$$

The boundary term is nonnegative since $v \in V$, $v \geqq 0$, $\beta(K) \geqq M$, $\mathbf{v} \cdot \boldsymbol{\nu}\beta(K) \leqq 0$ on S_2, and condition (3.3) holds. We let $\varepsilon \downarrow 0$. Since β is Lipschitz, the second term on the right is dominated by a constant times a function which goes to zero boundedly on G. Hence this term tends to zero. The third term on the right is nonnegative since $v \geqq 0$, $\beta(K) \geqq 0$ and $\operatorname{div}(\mathbf{v}) \geqq 0$ a.e. in G. Finally, as $\varepsilon \downarrow 0$, v tends to $\operatorname{sgn}_0^+(u-K) = \operatorname{sgn}_0^+(\beta(u)-\beta(K))$ boundedly a.e. in G, so

$$\|(f-\beta(K))^+\|_{L^1} \geqq \|(\beta(u)-\beta(K))^+\|_{L^1}.$$

The lemma now follows by choosing $\beta(K) = \max\{\operatorname{ess\,sup}_G(f), K_2\}$. □

Let $\mathscr{B}: V \to V^*$, the dual of V, be the operator defined by

$$(\mathscr{B}y, v)_{V^*,V} = b(y+g_1, v).$$

If we prove that $\mathscr{B}$ is bounded, coercive, and Type-M, then there is a solution $y \in V$ to

$$(\mathscr{B}y, v)_{V^*,V} = \langle f, v\rangle \quad \text{for all } v \in V$$

(cf. [7]). The function $y + g_1 \equiv u$ is then the solution to (3.10).

We begin by estimating

$$\begin{aligned}|(\mathscr{B}y, v)_{V^*,V}| = |b(y+g_1, v)| &\leqq \|\beta(y+g_1)\|_{L^2} \cdot K\|v\|_V\\ &\quad + \lambda\|y+g_1\|_V \cdot \|v\|_V + \lambda\|\mathbf{g}_2\|_{\mathscr{G}} \cdot \|v\|_V,\end{aligned}$$

where $1 + \lambda\|\mathbf{v}\|_{L^\infty} = K$. Since β is Lipschitz and $\beta(0) = 0$,

$$\|\beta(y+g_1)\|_{L^2} \leqq \mu\|y+g_1\|_{L^2} \leqq \mu C\|y+g_1\|_V,$$

where C depends only on G (by Poincaré's lemma). We conclude that $\mathscr{B}$ is bounded.

Next, we observe that if $y_n \rightharpoonup y$ (converges weakly) in V, then $y_n \to y$ in L^2 and $\beta(y_n + g_1) \to \beta(y+g_1)$ in L^2 since β is Lipschitz. We conclude from the definition of $\mathscr{B}$ that $\mathscr{B}$ is weakly continuous, hence Type-M.

Finally, we want to prove that $\mathscr{B}$ is coercive. For $v \in V$, we have

$$\begin{aligned}(\mathscr{B}(v), v)_{V^*,V} &= \int_G [\beta(v+g_1)(v+g_1-\lambda\mathbf{v}\cdot\nabla(v+g_1)) + \lambda\nabla(v+g_1)\cdot\nabla v]\\ &\quad - \int_G [\beta(v+g_1)(g_1 - \lambda\mathbf{v}\cdot\nabla g_1)] + \lambda\langle\gamma_\nu(g_2), \gamma_0(v)\rangle.\end{aligned}$$

Let $B: \mathbb{R} \to \mathbb{R}$ be the antiderivative of β with $B(0) = 0$. Then $B(s) \geqq 0$ for all s since β

is nondecreasing and $\beta(0)=0$. Also,

$$-\lambda\int_V \mathbf{v}\beta(v+g_1)\cdot\nabla(v+g_1)=-\lambda\int_G \mathbf{v}\cdot\nabla B(v+g_1)$$
$$=-\lambda\langle\gamma_\nu(\mathbf{v}),\ \gamma_0(B(v+g_1))\rangle+\lambda\int_G \operatorname{div}(\mathbf{v})B(v+g_1)$$

provided $\beta(v+g_1)\in L^\infty(G)$ (so that $B(v+g_1)\in H^1$). In this case, the rightmost side is given by

$$-\lambda\int_S \mathbf{v}\cdot\boldsymbol{\nu}B(v+g_1)+\lambda\int_G \operatorname{div}(\mathbf{v})B(v+g_1)\geqq-\lambda\int_{S_1}\mathbf{v}\cdot\boldsymbol{\nu}B(g_1)$$

since $\mathbf{v}\cdot\boldsymbol{\nu}\leqq 0$ on S_2 and div ($\mathbf{v}$), $B(v+g_1)\geqq 0$. This estimate now follows for all $v\in V$ by approximation.

With this estimate we have

$$(3.11)\qquad (\mathcal{B}(v),v)_{V^*,V}\geqq c\|v+g_1\|_{L^2}^2+\frac{\lambda}{2}\|v\|_V^2-K(\|\beta(v+g_1)\|_{L^2}+1),$$

where $\beta(s)s\geqq cs^2$ (since $\alpha=\beta^{-1}$ is Lipschitz) and K is independent of μ. From this estimate, it is easy to see that $(\mathcal{B}(v),v)_{V^*,V}/\|v\|_V\to\infty$ as $\|v\|_V\to\infty$; hence $\mathcal{B}$ is coercive.

We require one further estimate before we consider the case of non-Lipschitz β. Suppose u_1 and u_2 satisfy (3.10) with f replaced, respectively, with f_1 and f_2. Then

$$\int_G (f_1-f_2)v=\int_G[\beta(u_1)-\beta(u_2)][v-\lambda\mathbf{v}\cdot\nabla v]+\lambda\nabla(u_1-u_2)\cdot\nabla v$$

for all $v\in V$. We take $v=\sigma_\varepsilon^+(u_1-u_2)$, let $\varepsilon\downarrow 0$ and argue as in the previous lemma to get

$$(3.12)\qquad \|(f_1-f_2)^+\|_{L^1}\geqq\|(\beta(u_1)-\beta(u_2))^+\|_{L^1},$$

since $[\beta(u_1)-\beta(u_2)]\operatorname{sgn}_0^+(u_1-u_2)=(\beta(u_1)-\beta(u_2))^+$.

Theorem 1 is now proved in the case of Lipschitz $\beta=\alpha^{-1}$. Notice that we have *not* made use of the uniqueness of solutions to (3.4) yet; uniqueness in this case follows from (3.5), which we have obtained independently from the uniqueness proof. Our development in § 2, however, indicates that $\beta=\alpha^{-1}$ is not a function, since $\alpha'(s)=0$ for s belonging to the interval corresponding to the energy during a phase change. We suppose now, therefore, that β is not Lipschitz and set

$$\alpha_\mu(s)=\alpha(s)+\frac{s}{\mu},$$

and

$$\beta_\mu(s)=\alpha_\mu^{-1}(s).$$

Then β_μ is Lipschitz with Lipschitz constant $\leqq\mu$ and $|\beta_\mu(s)|\leqq|\beta_0(s)|$ for all s. Let u_μ denote the solution to (3.10) with β replaced by β_μ. Then

$$\|\beta_\mu(u_\mu)\|_{L^\infty}\leqq\max\{-K_1,K_2,\|f\|_{L^\infty}\}$$

uniformly in μ by our lemma, and from our calculation (3.11) we see that $\|u_\mu\|_V$ must also be uniformly bounded. Choose a sequence $\{\mu_k\}$ tending to $+\infty$ so that

$$u_{\mu_k} \rightharpoonup u \quad \text{in } V,$$
$$u_{\mu_k} \to u \quad \text{in } L^2,$$

and

$$\beta_{u_k}(u_{\mu_k}) \rightharpoonup w \quad \text{in } L^2.$$

To prove that $\alpha(w) = u$, let $w_k = \beta_{\mu_k}(u_{\mu_k})$. Then

$$\alpha(w_k) = \alpha_{\mu_k}(w_k) - \frac{(w_k)}{\mu_k} = u_{\mu_k} - \frac{w_k}{\mu_k},$$

which tends to u as $k \to \infty$ since $\|w_k\|_{L^\infty}$ is bounded. We have

$$\begin{aligned}\|\alpha(w) - u\|_{L^2} &\leqq \|\alpha(w) - \alpha(w_k)\|_{L^2} + \|\alpha(w_k) - u_{\mu_k}\|_{L^2} + \|u_{\mu_k} - u\|_{L^2} \\ &= \|\alpha(w) - \alpha(w_k)\|_{L^2} + 1/\mu_k \|w_k\|_{L^2} + \|u_{\mu_k} - u\|_{L^2}.\end{aligned}$$

The last two terms tend to zero as $k \to \infty$. Since α is nondecreasing and Lipschitz, there is a constant $c > 0$ so that

$$\|\alpha(w) - \alpha(w_k)\|_{L^2}^2 \leqq c \int (w - w_k)(\alpha(w) - \alpha(w_k)).$$

Since $w_k \rightharpoonup w$ and $\alpha(w_k) \to u$ in L^2, this term tends to zero as well. The function w is the solution to (3.4) since u is the solution to (3.10). We know that the solution w is unique (!); hence every subsequence $\{u_{\mu_k}\}$ converges to u weakly in V and every subsequence $\{\beta_{\mu_k}(u_{u_k})\}$ converges to w weakly in L^2. This convergence is enough to guarantee that estimate (3.12) and the result of the lemma remain valid upon passage to the limit; hence (3.5) and (3.6) follow, and the theorem is proved. □

COROLLARY. *The operator $\mathcal{A} \equiv \bar{\mathcal{A}}_2^{L^1 \times L^1}$ (the closure in $L^1 \times L^1$ of the graph of $\mathcal{A}_2$) is m-accretive.*

Proof. The closure of an accretive operator is accretive, so we need only prove the existence of solutions $w \in D(\mathcal{A})$ to

$$w + \lambda \mathcal{A}(w) \ni f$$

for any choice of $f \in L^1$ and $\lambda > 0$. Define

$$f_n(x) = \begin{cases} n, & f(x) > n, \\ f(x), & -n \leqq f(x) \leqq n, \\ -n, & f(x) < -n. \end{cases}$$

Then $f_n \in L^\infty$, $f_n \to f$ in L^1 and there is a solution $w_n \in D(\mathcal{A}_2)$ to (3.4) with f replaced by f_n. Since

$$\|w_n - w_m\|_{L^1} \leqq \|f_n - f_m\|_{L^1},$$

the sequence $\{w_n\}$ is Cauchy in L^1. If $w_n \to w$, then the pairs $[w_n, (f_n - w_n)/\lambda] \in \mathcal{A}_2$ converge in $L^1 \times L^1$ to $[w, (f - w)/\lambda] \in \mathcal{A}$, and we have our solution. □

4. Weak solutions. There is a unique integral solution $w \in C([0, T]; L^1)$ to the Cauchy problem (3.1) by Benilan's theorem. We shall now prove that w satisfies (3.2) for all appropriate testing functions φ.

THEOREM 2. *Let $w(t) \in C([0, T]; L^1)$ be the solution to* (3.1) *and assume $w_0 \in L^\infty \cap D(\mathscr{A})$. Then $w(t)$ satisfies*

$$\begin{aligned}\min\left\{\operatorname*{ess\,inf}_G(w_0), K_1\right\} + \int_0^t \operatorname*{ess\,inf}_G(g^-(s))\,ds \leqq w(t)\\ \leqq \max\left\{\operatorname*{ess\,sup}_G(w_0), K_2\right\}\\ + \int_0^t \operatorname*{ess\,sup}_G(g^+(s))\,ds \quad \text{a.e. in } G.\end{aligned} \tag{4.1}$$

By interpolation, $w(t)$ is continuous into L^p for $1 \leqq p < \infty$. If $g(t) \equiv g \in L^\infty$, then $\alpha(w) - g_1 \in L^2(0, T; V)$, and $w(t)$ satisfies (3.2) *for all $\varphi \in C^1([0, T]; L^2) \cap L^2((0, T); V)$. If $g(t) \in L^1((0, T); L^\infty)$, then $\alpha(w)$ may fail to belong to $L^2(0, T; H^1)$, but $w(t)$ satisfies*

$$\begin{aligned}\int_0^T \langle g, \varphi\rangle + \left\langle w, \frac{d\varphi}{dt} + \mathbf{v}\cdot\nabla\varphi\right\rangle + \langle \alpha(w), \Delta\varphi\rangle\,dt\\ = \langle w(T), \varphi(T)\rangle - \langle w(0), \varphi(0)\rangle + \int_0^T \langle \gamma_\nu(g_2), \nu_0(\varphi)\rangle + \langle \gamma_\nu(\nabla\varphi), \gamma_0(g_1)\rangle\end{aligned} \tag{4.2}$$

for all $\varphi \in C^1([0, T]; L^2) \cap L^2((0, T); V \cap H^2)$ with $\partial\varphi(t)/\partial\nu = 0$ on S_2.

Proof. We consider first the case $g(t) \equiv g \in L^\infty$. The operator $\mathscr{A}_g$ defined on $D(\mathscr{A})$ by

$$\mathscr{A}_g(w) = \mathscr{A}(w) - g$$

is m-accretive in $L^1 \times L^1$. The integral solution to (3.1) in this case is obtained using the Crandall–Liggett generation theorem for the operator $\mathscr{A}_g$ (cf. [1, Chap. 3]). Let $I: L^1 \to L^1$ be the identity and for $\lambda > 0$ let $J_\lambda \equiv (I + \lambda\mathscr{A}_g)^{-1}$ be the resolvent of $\mathscr{A}_g$. Since $\mathscr{A}_g$ is m-accretive, J_λ is a contraction on L^1. If $[t]$ denotes the greatest integer less than or equal to t,

$$w(t) = \lim_{n\to\infty} J_{1/n}^{[tn]}(w_0),$$

where convergence is uniform in L^1 for $0 \leqq t \leqq T$. We calculate

$$f = J_\lambda(w) \Leftrightarrow f + \lambda\mathscr{A}(f) - \lambda g \ni w \Leftrightarrow f + \lambda\mathscr{A}(f) \ni w + \lambda g.$$

By (3.6),

$$\min\left\{\operatorname*{ess\,inf}_G(w_0 + \lambda g), K_1\right\} \leqq J_\lambda(w_0) \leqq \max\left\{\operatorname*{ess\,sup}_G(w_0 + \lambda g), K_2\right\}$$

a.e. in G. We iterate this estimate and use

$$\max\left\{\operatorname*{ess\,sup}_G(w_0 + \lambda g), K_2\right\} \leqq \max\left\{\operatorname*{ess\,sup}_G(w_0), K_2\right\} + \lambda \operatorname*{ess\,sup}_G(g^+)$$

to get

$$\begin{aligned}\min\left\{\operatorname*{ess\,inf}_G(w_0), K_1\right\} + \lambda[tn] \operatorname*{ess\,inf}_G(g^-)\\ \leqq J_\lambda^{[tn]}(w_0) \leqq \max\left\{\operatorname*{ess\,sup}_G(w_0), K_2\right\} + \lambda[tn] \operatorname*{ess\,sup}_G(g^+).\end{aligned}$$

Let $\lambda = 1/n$ and let $n \to \infty$ to obtain (4.1).

Since $w_0 \in L^\infty$, $J_\lambda(w_0) \in D(\mathscr{A}_2)$, the domain of our original operator in L^2. Let us define

$$w_n(t) = J_{1/n}^{[tn]}(w_0) \in D(\mathscr{A}_2).$$

Then $w_n(t)$ converges uniformly in L^1 (and in L^2 by interpolation) to the solution $w(t)$. Notice that since α is Lipschitz, $\alpha(w_n(t))$ converges uniformly in L^2 to $\alpha(w(t))$. Moreover,

$$w_n\left(t+\frac{1}{n}\right) = J_{1/n}^{[tn]+1}(w_0) = J_{1/n}J_{1/n}^{[tn]}(w_0) = J_{1/n}(w_n(t)),$$

and consequently,

$$n\left(w_n\left(t+\frac{1}{n}\right) - w_n(t)\right) = -\left(\mathscr{A}_2\left(w_n\left(t+\frac{1}{n}\right)\right) - g\right). \tag{4.3}$$

The function $\alpha(w_n(t+1/n)) - g_1 \in V$, so

$$\begin{aligned}\left\langle \mathscr{A}_2\left(w_n\left(t+\frac{1}{n}\right)\right), \alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right\rangle \\ = \left\|\alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right\|_V^2 + \left\langle \nabla g_1, \nabla\left(\alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right)\right\rangle \\ - \left\langle \mathbf{v}w_n\left(t+\frac{1}{n}\right), \nabla\left(\alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right)\right\rangle \\ = -\left\langle n\left(w_n\left(t+\frac{1}{n}\right) - w_n(t)\right) - g, \alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right\rangle.\end{aligned}$$

It is a consequence of the Crandall-Liggett generation theorem that

$$n\left\|w_n\left(t+\frac{1}{n}\right) - w_n(t)\right\|_{L^1} \leqq \sqrt{2}\|\mathscr{A}_g w_0\|_{L^1}. \tag{4.4}$$

Since $g_1 \in L^\infty \cap H^1$, it follows from the uniform boundedness of $\alpha(w_n)$ that

$$\left\|\alpha\left(w_n\left(t+\frac{1}{n}\right)\right) - g_1\right\|_V^2 \leqq K(\|\mathscr{A}_g w_0\|_{L^1} + 1)$$

by an application of Hölder's inequality. The constant K is independent of n. We let $n \to \infty$. Then some (indeed, every) subsequence of $\alpha(w_n(t+1/n))$ converges weakly in H^1 to $\alpha(w(t))$, so $\alpha(w) \in L^\infty([0, T]; H^1)$ and $\alpha(w(t)) - g_1 \in V$ for almost every $t \in [0, T]$. We conclude that $\alpha(w_n(t+1/n)) - g_1$ converges weakly to $\alpha(w(t)) - g_1$ in $L^2([0, T]; V)$.

From (4.3) and the definition of $\mathscr{A}_2$, we have

$$\begin{aligned}\int_0^{T-1/n} n\left\langle w_n\left(t+\frac{1}{n}\right) - w_n(t), \varphi(t)\right\rangle dt \\ = \int_0^{T-1/n} \Bigg[-\langle \gamma_\nu, (\mathbf{g}_2), \gamma_0(\varphi(t))\rangle \\ + \left\langle \mathbf{v}w_n\left(t+\frac{1}{n}\right) - \nabla\alpha\left(w_n\left(t+\frac{1}{n}\right)\right), \nabla\varphi(t)\right\rangle + \langle g, \varphi(t)\rangle\Bigg] dt.\end{aligned} \tag{4.5}$$

We perform the standard change of variables on the right so that the difference quotients are on φ rather than w_n, take the limit as $n \to \infty$ and use the convergence properties of w_n, $\alpha(w_n)$ and the smoothness of φ to obtain (3.2).

If $g(t)$ is a step function, then we obtain the integral solution $w(t)$ to (3.1) by successively solving

$$\frac{dw(t)}{dt} + \mathscr{A}(w(t)) \ni g(t), \qquad T_j < t \leqq T_{j+1},$$

where g is constant on the interval (T_j, T_{j+1}) and $w(0) = w_0$. It is obvious that equation (4.1) continues to hold in this case. If the test function $\varphi(t)$ satisfies $\partial\varphi(t)/\partial\nu = 0$ on S_2 for all $t \in [0, T]$, then (4.5) is equivalent to

$$\int_0^{T-1/n} n\left\langle w_n\left(t+\frac{1}{n}\right) - w_n(t), \varphi(t)\right\rangle dt$$

$$= -\int_0^{T-1/n} \Bigg[\langle \gamma_\nu(g_2), \gamma_0(\varphi(t))\rangle + \langle \gamma_\nu(\nabla\varphi(t)), \gamma_0(g_1)\rangle$$

$$- \left\langle \mathbf{v}w_n\left(t+\frac{1}{n}\right), \nabla\varphi(t)\right\rangle - \left\langle \alpha\left(w_n\left(t+\frac{1}{n}\right)\right), \Delta\varphi(t)\right\rangle - \langle g, \varphi(t)\rangle\Bigg] dt.$$

Letting $n \to \infty$ and arguing as in the previous case, we obtain (4.2).

Finally, if $g \in L^1([0, T]; L^\infty)$ is arbitrary, then we obtain our results by approximating g by step functions and noting that the approximating solutions satisfy (4.1) and (4.2). These solutions converge uniformly in $L^1([0, T]; L^1)$ and are uniformly bounded in L^∞; hence they converge weakly in $L^2([0, T]; L^2)$ by the dominated convergence theorem. Under these conditions, (4.1) and (4.2) continue to hold for the solution to (3.1) □

5. Uniqueness of weak solutions. If the solution $w(t)$ to (3.2) is Lipschitz into L^1 and bounded into L^∞, then by interpolation it is Hölder continuous into L^p, $1 \leqq p < \infty$, with Hölder exponent $1/p$. A sufficient condition for $w(t)$ to be Lipschitz is to have $g(t) \in W^{1,1}([0, T]; L^1)$ and $w_0 \in D(\mathscr{A})$ (cf. [1, p. 131]; the relevant part of the theorem is true in *any* Banach space). We now prove the uniqueness of Hölder continuous solutions to (3.2).

THEOREM 3. *Let $T_0 \in \mathbb{R}$ be any point for which both w_1 and w_2 are Hölder continuous from $[0, T_0]$ into L^2 and assume that $w_1(0) = w_2(0)$ a.e. in G and both w_1 and w_2 satisfy (3.2) for all $\varphi \in C^1([0, T_0]; L^2) \cap L^2([0, T_0]; V)$. Then*

$$w_1(T_0) = w_2(T_0) \quad \textit{a.e. in } G.$$

Proof. For all $T \leqq T_0$ and $\varphi \in C^1([0, T]; L^2) \cap L^2([0, T]; V)$,

$$\langle w_1(T) - w_2(T), \varphi(T)\rangle$$

$$= \int_0^T \left\langle w_1 - w_2, \frac{d\varphi}{dt} + \mathbf{v}\cdot\nabla\varphi\right\rangle - \langle \nabla(\alpha(w_1) - \alpha(w_2)), \nabla\varphi\rangle$$

$$= \int_0^T 2K\langle w_1 - w_2, \varphi\rangle + \left\langle w_1 - w_2, \frac{d\varphi}{dt} + \mathbf{v}\cdot\nabla\varphi - 2K\varphi\right\rangle - \langle \nabla(\alpha(w_1) - \alpha(w_2)), \nabla\varphi\rangle,$$

where $K \geqq K(\mathbf{v})$, the constant defined in (3.9). This equation is equivalent to

$$\begin{aligned}&\langle w_1(T)-w_2(T), \varphi(T)\rangle\\&\quad=\int_0^T e^{2K(T-t)}\left[\left\langle w_1(t)-w_2(t), \frac{d\varphi(t)}{dt}+\mathbf{v}\cdot\nabla\varphi(t)-2K\varphi(t)\right\rangle\right.\\&\qquad\qquad\left.-\langle\nabla(\alpha(w_1)(t)-\alpha(w_2)(t)), \nabla\varphi(t)\rangle\right]dt.\end{aligned}$$

In fact, this second equation follows from the first by the usual variation of parameters formula for solutions of nonhomogeneous, first order ordinary differential equations. Theorem 3 follows by choosing an appropriate testing function $\varphi(t)$.

Let $c=\|\alpha\|_{\mathrm{Lip}}^{-1}$ and define the linear operator $\mathscr{L}$ by

$$D(\mathscr{L})=\{v\in V\,|\,\Delta v\in L^2, \langle\gamma_\nu(\nabla v), \gamma_0(\psi)\rangle=0 \quad \text{for all } \psi\in V\}$$

and

$$\mathscr{L}(v)=-2Kv+\mathbf{v}\cdot\nabla v+\Delta v/c.$$

Then $-\mathscr{L}$ is elliptic, and for all $v\in D(\mathscr{L})$ and $\psi\in V$,

$$\langle-\mathscr{L}v, \psi\rangle=\int_G \nabla v\cdot\nabla\psi/c-\mathbf{v}\cdot\nabla v\psi+2Kv\psi.$$

Notice that $-\mathscr{L}$ is V-coercive since

$$\begin{aligned}\langle-\mathscr{L}v, v\rangle&=\frac{\|v\|_V^2}{c}+\int_G 2K(v)^2-\frac{1}{2}\mathbf{v}\cdot\nabla(v^2)\\&=\frac{\|v\|_V^2}{c}+\int_G 2K(v)^2+\frac{1}{2}\operatorname{div}(\mathbf{v})(v^2)-\frac{1}{2}\int_S \mathbf{v}\cdot\nu(v^2)\\&\geqq\frac{\|v\|_V^2}{c}\end{aligned}$$

by the choice of boundary conditions on v.

Since $w_1, w_2, \alpha(w_1)$ and $\alpha(w_2)$ are Hölder continuous into L^2, the backward heat equation

$$\begin{aligned}&\frac{d\varphi(t)}{dt}+\mathscr{L}\varphi(t)=(w_1(t)-w_2(t))-c(\alpha(w_1(t))-\alpha(w_2(t))), \qquad 0\leqq t\leqq T,\\&\varphi(T)=\varphi_T,\end{aligned}$$

with $\varphi_T\in D(\mathscr{L})$, is well posed. In fact, the theory of analytic semigroups provides a strong solution $\varphi(t)\in C^1([0, T]; L^2)\cap D(\mathscr{L})$ for all $t\in[0, T]$ (cf. [9]). We observe that

$$\begin{aligned}\|\varphi(t)\|_V^2=c\Bigg[&\left\langle\frac{d\varphi(t)}{dt}+\mathbf{v}\cdot\nabla\varphi(t)+c(\alpha(w_1(t))-\alpha(w_2(t)))-(w_1(t)-w_2(t)), \varphi(t)\right\rangle\\&-2K\|\varphi(t)\|_{L^2}^2\Bigg].\end{aligned}$$

Since $\varphi\in C^1([0, T]; L^2)$ and $w_1, w_2, \alpha(w_1)$ and $\alpha(w_2)$ are continuous into L^2, we conclude, using Hölder's inequality, that $\|\varphi(\cdot)\|_V$ is uniformly bounded on $[0, T]$; hence $\varphi\in L^2([0, T]; V)$.

With this testing function, we have

$$\begin{aligned}
&\langle w_1(T)-w_2(T),\varphi_T\rangle\\
&=\int_0^T e^{2K(T-t)}\Big[\Big\langle w_1(t)-w_2(t)-c(\alpha(w_1(t))-\alpha(w_2(t))),\frac{d\varphi(t)}{dt}+\mathbf{v}\cdot\nabla\varphi(t)-2K\varphi(t)\Big\rangle\\
&\qquad\qquad+c\Big\langle\alpha(w_1(t))-\alpha(w_2(t)),\frac{d\varphi(t)}{dt}+\mathscr{L}(\varphi(t))\Big\rangle\Big]\\
&=\int_0^T e^{2K(T-t)}\Big[\Big\langle\frac{d\varphi(t)}{dt}+\frac{\Delta\varphi(t)}{c}+\mathbf{v}\cdot\nabla\varphi(t)-2K\varphi(t),\frac{d\varphi(t)}{dt}+\mathbf{v}\cdot\nabla\varphi(t)-2K\varphi(t)\Big\rangle\\
&\qquad\qquad+c\langle\alpha(w_1(t))-\alpha(w_2(t)),w_1(t)-w_2(t)-c(\alpha(w_1(t))-\alpha(w_2(t)))\rangle\Big]\,dt.
\end{aligned}$$

The last term on the right is nonnegative since $c=\|\alpha\|_{\mathrm{Lip}}^{-1}$. We conclude that

$$\begin{aligned}
c\langle w_1(T)-w_2(T),\varphi(T)\rangle&\geqq\int_0^T e^{2K(T-t)}\Big\langle\Delta\varphi(t),\frac{d\varphi(t)}{dt}+\mathbf{v}\cdot\nabla\varphi(t)-2K\varphi(t)\Big\rangle\,dt\\
&\geqq\int_0^T e^{2K(T-t)}\Big[\Big\langle\Delta\varphi(t),\frac{d\varphi(t)}{dt}\Big\rangle+K\|\varphi(t)\|_V^2\Big]\,dt
\end{aligned}$$

by (3.9) and the identity $\langle\Delta\varphi,-\varphi\rangle=\|\varphi\|_V^2$. If we formally integrate the first term by parts, we obtain the identity

$$\begin{aligned}
\int_0^T e^{2K(T-t)}\Big\langle\Delta\varphi(t),\frac{d\varphi(t)}{dt}\Big\rangle\,dt&=-\int_0^T e^{2K(T-t)}\frac{d}{dt}\|\varphi(t)\|_V^2/2\,dt\\
&=\frac{e^{2K(T)}}{2}\|\varphi(0)\|_V^2-\tfrac{1}{2}\|\varphi_T\|_V^2\\
&\quad+\int_0^T e^{2K(T-t)}(-K\|\varphi(t)\|_V^2)\,dt.
\end{aligned}$$

In fact, this formal calculation yields a valid identity as may be verified by considering Riemann sums of the (continuous) integrand. Hence, we may estimate

$$2c\langle w_1(T)-w_2(T),\varphi_T\rangle\geqq-\|\varphi_T\|_V^2$$

or

$$\langle w_1(T)-w_2(T),(-2c)\varphi_T\rangle\leqq\|\varphi_T\|_V^2$$

for *any* choice of $\varphi_T\in D(\mathscr{L})$. But $D(\mathscr{L})$ is a subspace of $L^2(G)$, so if we take $\varepsilon\varphi_T$ in place of φ_T, divide by ε and let $\varepsilon\to0$, we get

$$\langle w_1(T)-w_2(T),\varphi_T\rangle=0$$

for all $\varphi_T\in D(\mathscr{L})$. Finally, $D(\mathscr{L})$ is dense in L^2, and we conclude that

$$w_1(T)=w_2(T)\quad\text{a.e. in }G.\qquad\square$$

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.

[2] P. BENILAN, *Equations d'évolution dans un espace de Banach quelconque et applications*, thesis, Orsay, 1972.

[3] H. BREZIS AND W. STRAUSS, *Semi-linear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565-590.

[4] M. G. CRANDALL AND T. M. LIGGETT, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265-298.

[5] A. FASANO, M. PRIMICERIO AND L. RUBENSTEIN, *A model for heat conduction with a free boundary in a concentrated capacity*, J. Inst. Math. Appl., 26 (1980), pp. 327-347.

[6] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980.

[7] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non-lineares*, Dunod, Paris, 1969.

[8] L. RUBENSTEIN, *The Stefan problem: comments on its present state*, J. Inst. Math. Appl., 24 (1979), pp. 259-277.

[9] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

[10] ———, *Mathematical formulation of the Stefan problem*, Internat. J. Engrg. Sci., 20 (1982), pp. 909-912.

[11] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam-New York, 1979.

[12] A. VISINTIN, *General free boundary evolution problems in several space dimensions*, J. Math. Anal. Appl., 95 (1983), pp. 117-143.

SIAM J. MATH. ANAL.
Vol. 18, No. 6, November 1987

020

AN INVERSE PROBLEM FOR A NONLINEAR ELLIPTIC DIFFERENTIAL EQUATION [*]

MICHAEL PILANT[†] AND WILLIAM RUNDELL[†]

Abstract. In this paper we determine the form of an unknown forcing term f in the equation $\Delta u = \gamma(x, y, u) + f(u)$ from overprescribed boundary data on a domain in R^2 by reformulating the problem as a fixed point problem in an appropriate function space. Given suitable restrictions on the growth of f and the boundary data we recover a unique f.

Key words. undetermined coefficient, overposed boundary data, convergent iteration scheme

AMS(MOS) subject classifications. 35R30, 35J25

1. Introduction.

This paper considers the elliptic differential equation

$$-\Delta u = \gamma(x, y) + f(u) \tag{1.1}$$

defined on bounded open domain Ω of $\mathbf{R}^2$, with smooth boundary $\partial\Omega$. The function $\gamma(x, y)$ is known, but $f(u)$ as well as $u(x, y)$ has to be determined. We shall assume that $\partial\Omega$ can be decomposed into two connected components, $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, and on $\partial\Omega_1$ we impose the boundary condition

$$\alpha\frac{\partial u}{\partial\nu} + \beta u = g_1, \qquad (x, y) \in \partial\Omega_1 \tag{1.2}$$

while on $\partial\Omega_2$ we give *both* Dirichlet *and* Neumann boundary data

$$\begin{aligned} \frac{\partial u}{\partial\nu} &= g_2, \\ u &= \theta, \end{aligned} \qquad (x, y) \in \partial\Omega_2. \tag{1.3}$$

Here ν is the inner normal to the curve $\partial\Omega$.

By this overposing of the data on part of the boundary we hope to recover the pair of functions $\langle u(x, y), f(u)\rangle$. Our result is to show that conditions can be given on the boundary data and the known forcing function $\gamma(x, y)$ in order to obtain a unique function $f(u)$, provided $f(u)$ is, in a sense to be defined, "sufficiently small."

We are thus able to obtain the exact nature of a nonlinear forcing function in Poisson's equation that is subject to a small but unknown perturbation depending only on the function $u(x, y)$. The maximum allowable size of the perturbation will depend on the data and the region Ω. Our method will lead to a construction of the function $f(u)$.

Problems such as the above have received recent attention, and we mention some related work. First however we note that there is an inherent limitation on the form of any right-hand side of (1.1) that can be recovered by measurements made only on the boundary. If $-\Delta u = f(x, y)$, where f depends on x and y, then even in this case of a linear equation the function f cannot be determined uniquely no matter what the imposed boundary data. To see this, suppose (u_1, f_1) is a solution of $-\Delta u_1 = f_1$ where both u_1 and $\partial u_1/\partial\nu$ are prescribed on the boundary $\partial\Omega$. Then if ϕ is *any*

[*] Received by the editors July 14, 1986; accepted for publication October 22, 1986.

[†] Department of Mathematics, Texas A& M University, College Station, Texas 77843.

smooth function of compact support in Ω, set $u_2 = u_1 + \phi$ and $f_2 = f_1 + \Delta\phi$, so that u_1 and u_2 agree with all their derivatives on $\partial\Omega$. If the pair (u_1, f_1) is a solution, so also is (u_2, f_2). Thus one cannot expect uniqueness regardless of how many derivatives of the function u are imposed on the boundary. If, however, one restricts the allowed dependence of the forcing function f in advance then uniqueness results are possible. For example if Ω is a square and $f = f(x)$ only, then f can be recovered by overposing both Dirichlet and Neumann data on one side. If the function f has a linear dependence on u with proportionality coefficient depending only on *one* of the space variables, $f = a(x)u$ say, then again f or in this case $a(x)$ can be recovered by conditions of the form (1.2), (1.3) (see [4]). The authors in [7] considered the parabolic equation $u_t - \Delta u = f(u)$, and conditions were given on the initial–boundary data to guarantee the existence of a unique function f, locally in time, but with no restrictions on size. In fact $f(u)$ was restricted only by the requirements of uniqueness for the forwards or direct problem, that is the problem of finding u if the operator and the forcing function were known. The techniques of the present paper will have considerable overlap with those of [7], in that we shall set up a mapping from the space of admissible functions $f(u)$ to the overposed data using an evaluation of (1.1) on the arc $\partial\Omega_2$. We shall show that a solution of our problem is equivalent to this map having a fixed point, and accomplish this by showing it to be a contraction on a suitable function space.

We remark that the related inverse problem of finding the unknown function $f(u)$ in $u_t - \Delta u = f(u)$, or of finding the conductivity $a(u)$ in $u_t - \nabla.a(u)\nabla u = f(u)$ has an extensive literature but the techniques of these papers are different from the ones proposed here; cf. [1], [2], [3], [5], [9].

The plan of this paper is as follows. The next section describes our notation and gives the assumptions on the data. The third section constitutes the main body of the paper and contains the existence and uniqueness result.

2. Notation and assumptions.

Denote by $C^{k,\alpha}(D)$ the usual Schauder space of functions whose kth order partial derivatives are Hölder continuous of order α, $0 < \alpha \leq 1$ on the set D. The usual norm on this space we shall denote by $\|\cdots\|_{k,\alpha}$, the set D being usually understood by the context.

We shall assume that the boundary $\partial\Omega$ of $\Omega \subset R_2$ is the union of two simply connected arcs $\partial\Omega_1$ and $\partial\Omega_2$, each of which lie in $C^{2,1}$. It will be convenient to paramerize the segment $\partial\Omega_2$ on which both Dirichlet and Neumann data is prescribed, by the variable t, $0 \leq t \leq 1$. We denote by (x_0, y_0) and (x_1, y_1) the corresponding endpoints of the arc $\partial\Omega_2$.

Our assumptions on the data shall be for two purposes. First, to guarantee the existence of a sufficiently smooth solution u to the direct problem, that is a function $u(x, y; f)$ satisfying

$$-\Delta u = \gamma(x, y) + f(u) \tag{2.1}$$

for a given $f(u)$ and subject to the primary boundary conditions

$$\alpha(x, y)\frac{\partial u}{\partial\nu} + \beta(x, y)\, u = g_1(x, y), \qquad (x, y) \in \partial\Omega_1, \tag{2.2}$$

$$\frac{\partial u}{\partial\nu} = g_2(x, y), \qquad (x, y) \in \partial\Omega_2. \tag{2.3}$$

Second, to ensure that our mapping from f to the overposed data

$$u = \theta(x,y), \qquad (x,y) \in \partial\Omega_2 \tag{2.4}$$

is well defined we need to guarantee that $u(x,y;f)$ attains both its minimum and maximum on $\partial\Omega_2$, and is a monotone function of the parameter t on $\partial\Omega_2$.

We shall denote by Λ_0 the smallest eigenvalue for the problem $-\Delta\phi = \lambda\phi$ on Ω, where ϕ is subject to the homogeneous primary boundary conditions (2.2) and (2.3). Since Ω is bounded it follows that $\Lambda_0 > 0$ provided that β not identically zero on $\partial\Omega_1$.

We shall assume that the boundary data satisfies:

A1. *$\alpha(x,y)$ and $\beta(x,y)$ lie in $C^{2,1}(\partial\Omega_1)$. In addition, $\alpha \geq 0, \beta > 0$ on $\partial\Omega_1$, and $\alpha = 0$ in a neighborhood of (x_0,y_0) and (x_1,y_1).* (That is, we have Dirichlet boundary conditions in a neighborhood of the endpoints of $\partial\Omega_1$.)

A2. $g_1 \in C^{2,1}(\partial\Omega_1)$,
$g_2 \in C^{2,1}(\partial\Omega_2)$,
$\theta \in C^{3,1}(\partial\Omega_2)$.

A3. *For $(x,y) \in \Omega$ the function $\gamma(x,y) \in C^{1,1}(\Omega)$.*

For a given $E > 0$ denote by S_E the ball in $C^{(1,1)}$ centered at the origin, that is

$$S_E := \{ f \mid f(0) = 0, \quad \|f\|_{1,1} \leq E \}. \tag{2.5}$$

With some restrictions on the value of E this set will form the admissible class of functions f. By a solution pair (u,f) to (2.1) – (2.4) we shall mean that,

(a) $f \in S_E$.
(b) $u \in C^{2,1}(\bar{\Omega})$.
(c) (2.1) – (2.4) hold.

We note that conditions A1 – A3 are sufficient to guarantee the existence of a unique solution to the direct problem (2.1) - (2.3) for any $f \in S_E$, provided $E < \Lambda_0$. This ensures that we avoid any eigenvalues of the linearized problem.

If σ and ν denote unit vectors in the tangential and normal directions on the boundary $\partial\Omega$, then for some functions $a(s)$ and $b(s)$ of the arc length s, and with $a^2 + b^2 = 1$,

$$\frac{\partial}{\partial x} = a\frac{\partial}{\partial\sigma} + b\frac{\partial}{\partial\nu}, \qquad \frac{\partial}{\partial y} = b\frac{\partial}{\partial\sigma} - a\frac{\partial}{\partial\nu}.$$

Thus in this coordinate system the Laplacian has principal part $\frac{\partial^2}{\partial\sigma^2} + \frac{\partial^2}{\partial\nu^2}$.

Let D_ν and D_σ denote the normal and tangential components of the Laplacian on the boundary $\partial\Omega$, that is

$$D_\nu = \frac{\partial^2}{\partial\nu^2} + c(s)\frac{\partial}{\partial\nu}$$

and

$$D_\sigma = \frac{\partial^2}{\partial\sigma^2} + d(s)\frac{\partial}{\partial\sigma}$$

where $c(s)$ and $d(s)$ are coefficients that depend only on the boundary $\partial\Omega$. For example if Ω is the unit circle then $D_\nu = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}$ and $D_\sigma = \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}$. For $f \in S_E$, and $u(x,y;f)$ the corresponding solution of the direct problem, (2.1) - (2.3), define the mapping $\mathbf{T}_\theta[f]$ by

$$\mathbf{T}_\theta[f] = \tilde{f}(\theta) = -\tilde{\gamma}(t) + D_\sigma\theta + D_\nu\{u(x,y;f)\} \tag{2.6}$$

where $\tilde{v}$ denotes the restriction of a function v to the boundary $\partial\Omega_2$. Equation (2.6) is precisely the projection of (1.1) onto the boundary $\partial\Omega_2$.

We shall say that the mapping $\mathbf{T}_\theta[f]$ has a *theta fixed point* f in S_E if for some f in this space

$$\mathbf{T}_\theta[f] = f(\theta).$$

Let the functions $\psi(x,y)$, $\psi_+(x,y)$ and $\psi_-(x,y)$ satisfy

$$-\Delta\psi = \gamma(x,y), \tag{2.7}$$
$$-\Delta\psi_+ = \gamma(x,y) + E\psi_+, \qquad (x,y)\in\Omega \tag{2.8}$$
$$-\Delta\psi_- = \gamma(x,y) - E\psi_-, \tag{2.9}$$

and be subject to the primary boundary conditions (2.2), (2.3).

In order to develop the properties we need of the mapping $\mathbf{T}_\theta[\ ]$, we will require that,

A4. *The functions* $\psi_\pm(x,y)$ *satisfy,*
$g_1(0)/\beta(0) = \psi_\pm(\pm_0, y_0) \le \psi_\pm(x,y) \le \psi_\pm(x_1,y_1) = g_1(1)/\beta(1)$ *for all* $(x,y)\in\bar{\Omega}$.

The overposed boundary data $\theta(t) = u\,|_{\partial\Omega_2}$ will be assumed to satisfy,

A5. $\|D_\sigma\{\theta-\psi\}\|_{1,1} < C_2$, *for some constant* C_2 *to be defined later.*
In addition, $\theta(0) = g_1(0)/\beta(0)$, $\theta(1) = g_1(1)/\beta(1)$, *and* $\theta(t)$ *is an increasing function of the parameter* t *on* $\partial\Omega_2$. *Furthermore,* $\theta(0)\le\theta(t)\le\theta(1)$.

The monotonicity assumption on $\theta(t)$ is necessary to recover $\tilde{f}$ from $\tilde{f}(\theta(t))$ in $\mathbf{T}_\theta[f]$. By $G = G(x,y,\xi,\eta)$ we denote the Green's function for the Laplacian on Ω subject to the primary boundary conditions (2.2) and (2.3).

We define the following quantitities, for $f\in S_E$ and $\|u\|_{1,1} < \infty$

$$C(\Omega,u) := \sup_{f\neq 0}\frac{\|\iint_\Omega D_\nu G f(u)\|_{1,1}}{\|f\|_{1,1}} \tag{2.10}$$

and

$$C_4(\Omega) := \sup_{h\neq 0}\frac{\|\iint_\Omega D_\nu G h(\xi,\eta)\|_{1,1}}{\|h\|_{1,1}}. \tag{2.11}$$

We can make the following observations:

First, let K be defined by

$$K[h] = \iint_\Omega D_\nu G h(\xi,\eta)\,d\xi\,d\eta. \tag{2.12}$$

By standard elliptic estimates [6], we have

$$\|K[h]\|_{1,1} \le C\|h\|_{1,1} \tag{2.13}$$

where C is a constant depending only on the domain Ω. C_4 is therefore the smallest such constant.

Secondly, from inequality (2.13) we have

$$\| \iint D_\nu Gf(u) \|_{1,1} \leq C_4(\Omega)\|f(u)\|_{1,1} \leq C_4(\Omega)\|f\|_{1,1}\|u\|^2_{1,1}$$

which implies that

$$C(\Omega, u) \leq C_4(\Omega)\|u\|^2_{1,1} \tag{2.14}$$

and that (2.10) is therefore well defined.

Thirdly, $K[h] = \iint D_\nu Gh$ may have a nontrivial kernel. To see this, consider the following example.

Let Ω be the unit square, and let $u(x,y)$ satisfy (2.1) with $\gamma = \gamma(y)$ together with the primary boundary conditions

$$\frac{\partial u(1,y)}{\partial x} = 0, \qquad \frac{\partial u(0,y)}{\partial x} = 0,$$

$$u(x,0) = 0, \qquad u(x,1) = 1,$$

along with the overposed condition

$$u(0,y) = \theta(y).$$

For this case note that

$$u_{xx}(0,y) = \iint_\Omega D_\nu Gh(\eta)d\xi d\eta = 0 \tag{2.15}$$

for any $h \in Lip_1$. This leads to the following inequality (where $u_0 = u_0(y)$):

$$\begin{aligned} C(\Omega,u) = \sup_{f \neq 0} \frac{\|Kf(u)\|_{1,1}}{\|f\|_{1,1}} &= \sup_{f\neq 0}\left\{ \frac{\| Kf(u) - Kf(u_0) \|_{1,1}}{\| f(u) - f(u_0) \|_{1,1}} \cdot \frac{\| f(u) - f(u_0) \|_{1,1}}{\|f\|_{1,1}} \right\} \\ &\leq C_4(\Omega) \sup_{f\neq 0} \frac{\|f(u) - f(u_0)\|_{1,1}}{\|f\|_{1,1}}. \end{aligned} \tag{2.16}$$

We have, in this case, $C(\Omega, u_0) = 0$.

We shall require that the constant $C(\Omega, u)$ satisfy :

A6. *$C(\Omega,u)$ defined in (2.10), is a continuous function of u, for $\|u\|_{2,1} < \infty$.* Note that from the above, there exists domains and primary boundary conditions on this domain for which $\mathbf{T}_\theta[\ ]$ reduces to the trivial mapping

$$\mathbf{T}_\theta[f] = -\tilde{\gamma}(t) + \theta''(t) = f(\theta(t))$$

from (2.6) and (2.15), and thus gives a direct determination of the function f without any iteration. The above conditions essentially restrict the allowable class of unknown functions $f(u)$ and the primary boundary data to be sufficiently small perturbations of such problems. Here sufficiently small means that the operator defined by (2.13) have norm less than one in a suitable space.

3. Existence and uniqueness. We shall prove the following result:

THEOREM. *If assumptions* A1 - A6 *hold, then for E sufficiently small, there exists a unique solution pair* $\langle u, f\rangle$ *of* (2.1) - (2.4).

We shall need some lemmas, the first shows that the mapping $\mathbf{T}_\theta[\,]$ is well defined, while the second shows the inverse problem of finding the function $f(u)$ is equivalent to finding a theta fixed point of $\mathbf{T}_\theta[f]$.

LEMMA 1. *If $u(x,y;f)$ denotes the solution of the direct problem then it takes its minimum and maximum at the points (x_0,y_0) and (x_1,y_1) on the boundary $\partial\Omega$.*

Proof. From A4, A5 and the maximum principle, [6, Chap. 9] applied to the functions u, ψ_+ and ψ_- we have that

$$\theta(0) \le \psi_-(x,y) \le u(x,y) \le \psi_+(x,y) \le \theta(1) \qquad \text{for } (x,y) \in \bar{\Omega}.$$

LEMMA 2. *Given $f \in S$ then $\langle u(x,y), f\rangle$ is a solution of* (2.1) - (2.4) *if and only if f is a theta fixed point of* $\mathbf{T}_\theta[f]$.

Proof. Let $\tilde{u}(t;f)$ and $\tilde{\gamma}(t)$ denote the restrictions of these functions to $\partial\Omega_2$. If $\langle f, u\rangle$ is a solution to (2.1) - (2.4), then for $t \in \partial\Omega_2$

$$\begin{aligned} f(\theta) &= f(\tilde{u}(t;f)) \\ &= D_\nu\{u\} + D_\sigma\{\tilde{u}\} - \tilde{\gamma}(t) \\ &= D_\nu\{u\} + D_\sigma\{\theta\} - \tilde{\gamma}(t) \\ &= \mathbf{T}_\theta[f] \end{aligned}$$

so that f is a theta fixed point of $\mathbf{T}_\theta[\,]$.

Since

$$\mathbf{T}_\theta[f] = -\tilde{\gamma}(t) + D_\sigma\{\theta\} + D_\nu\{u\}$$

where u is a solution of the direct problem, by evaluating (2.1) on $\partial\Omega_2$ we obtain $D_\nu\{u\} + D_\sigma\{\tilde{u}\} = \gamma + f(\tilde{u})$ and hence $D_\nu\{u\} = \tilde{\gamma}(t) + f(\tilde{u}) - D_\sigma\{\tilde{u}\}$. Combining this with the above equation in the case that f is a theta fixed point of $\mathbf{T}_\theta[f]$,

$$\mathbf{T}_\theta[f] = f(\theta) = D_\sigma\{\theta\} - D_\sigma\{\tilde{u}\} + f(\tilde{u})$$

for $t \in \partial\Omega_2$.

Setting $\sigma(t) = \tilde{u}(t;f) - \theta(t)$ for $t \in \partial\Omega_2$, we have

$$\begin{aligned} D_\sigma\{\sigma(t)\} &= D_\sigma\{\tilde{u}(t;f) - \theta(t)\} \\ &= f(\tilde{u}(t;f)) - f(\theta(t)). \end{aligned}$$

From A3, since $f \in S_E$, it follows that $D_\sigma\{\sigma\} = \xi_f(t)\,\sigma$ for some Lip_1 function $\xi_f(t)$ satisfying $|\xi_f| \le E$. From assumption A5 it follows that $\sigma(0) = \sigma(1) = 0$. If we let Λ_1 be the lowest eigenvalue for the operator D_σ on $\partial\Omega_2$ acting on functions that vanish at the endpoints (x_0,y_0) and (x_1,y_1), then if E is restricted to be less than Λ_1 it follows by comparison arguments from Sturm–Liouville theory that $\sigma(t) = 0$. Hence u satisfies (2.4).

LEMMA 3. $\mathbf{T}_\theta[\,]$ *maps a ball into itself in* $C^{1,1}$.

Proof. By definition

$$\begin{aligned}\tilde{f}(\theta(t)) = \mathbf{T}_\theta[f](t) &= -\gamma + D_\sigma[\theta] + D_\nu u(x,y;f)\\ &= D_\sigma[\theta-\psi] + D_\nu \iint_\Omega Gf(u)\\ &= D_\sigma[\theta-\psi] + \iint_\Omega Kf(u).\end{aligned}$$

Assume that $f(0)=0$ and $\|f\|_{1,1} \le E$. With respect to the $\|\ \|_{1,1}$ norm, on $[0,T]$, we have

$$\|\iint Kf(u)\|_{1,1} \le C(\Omega,u)\|f\|_{1,1}.$$

The domain of $\tilde{f}$ is not $[0,T]$, but $[\theta(0),\theta(T)] = [0,\theta(T)]$. We have the bound, which results from a transformation of variables:

$$\|f\|_{1,1,\,[\theta(0),\theta(1)]} \le C(T;\theta)\,\|f\|_{1,1,\,[0,1]}$$

where we have indicated the intervals over which the norms are taken. This immediately implies

$$\begin{aligned}\|\tilde{f}\|_{1,1} &\le C(T;\theta)\,\|\mathbf{T}_\theta[f]\|_{1,1}\\ &\le C(T;\theta)\Big[\|D_\sigma(\theta-\psi)\|_{1,1} + C(\Omega,u)\|f\|_{1,1}\Big].\end{aligned}$$

If $C(\Omega,u_0)=0$ then if $\|u-u_0\|_{1,1}$ is sufficiently small, we can satisfy

$$C(T,\theta).C(\Omega,u) < 1/2.$$

If $\|D_\sigma(\theta-\psi)\|_{1,1}$ is sufficiently small then

$$C(T,\theta)\|D_\sigma(\theta-\psi)\|_{1,1} < E/2$$

and consequently

$$\|\tilde{f}\|_{1,1} < E$$

and the Lemma is proved.

LEMMA 4. *If* $\|D_\sigma(\theta-\psi)\|_{1,1}$ *is sufficiently small, then* $\mathbf{T}_\theta[\,]$ *is a contraction on* Lip_1.

Proof. Under the hypotheses, $\mathbf{T}_\theta[\,]$ maps a ball into itself in $C^{1,1}$. Consequently, $\|f'\|_{0,1}$ is uniformly bounded. We therefore have the following:

$$\begin{aligned}\|\tilde{f}_1 - \tilde{f}_2\|_{1,1} &= \|\,\mathbf{T}_\theta[f_1] - \mathbf{T}_\theta[f_2]\,\|_{1,1}\\ &\le C(T,\theta)\,\|\iint D_\nu G\big(f_1(u_1) - f_2(u_2)\big)\,\|_{1,1}\\ &\le C(T,\theta)\,\|\,K[f_1(u_1) - f_1(u_2)]\,\|_{1,1} + C(\Omega,u_2)\|f_1 - f_2\|_{1,1}\\ &\le C(T,\theta)\,C_4(\Omega)\,E\,\|u_1 - u_2\|_{1,1} + C(\Omega,u_2)\|f_1 - f_2\|_{1,1}.\end{aligned}$$

The inequality

$$\|u_1 - u_2\|_{1,1} \leq C(\Omega)\|f_1 - f_2\|_{1,1} \tag{3.1}$$

with $C(\Omega) \to 0$ as $\mathrm{Vol}(\Omega) \to 0$ follows easily from estimates on lower norms of u. Consequently, if $\|u_2 - u_0\|_{1,1}$ is sufficiently small, and if the region is sufficiently small we conclude that

$$\|\tilde{f}_1 - \tilde{f}_2\|_{1,1} \leq \delta\|f_1 - f_2\|_{1,1} \tag{3.2}$$

and the lemma is proved.

The crucial estimate is that

$$C(T,\theta).C(\Omega,u_2) < 1. \tag{3.3}$$

This is essentially the requirement that $C(\Omega, u) \ll 1$ for some ball about u_0. This is guaranteed by the presence of a nontrivial kernel for the mapping $K = D_\nu G$ and continuity.

We now compute an a priori bound on the iterates of $\mathbf{T}_\theta[f]$ by the following argument. With $f_0 = 0$, and using (2.7)

$$\begin{aligned} \| f_0 - \mathbf{T}_\theta[f_0] \|_{1,1} &= \| \mathbf{T}_\theta[f_0] \|_{1,1} \\ &= \| -\tilde{\gamma}(t) + D_\sigma\{\theta\} + D_\nu\{\psi\} \|_{1,1} \\ &= \| D_\sigma\{\theta - \psi\} \|_{1,1} \\ &\leq C_2. \end{aligned}$$

Define the sequence $\{f_n\}$ for $n \geq 1$ by

$$f_{n+1} = \mathbf{T}_\theta[f_n], \qquad f_0 = 0, \tag{3.4}$$

From (3.2) we have since $\delta < 1$, and $f_0 = 0$

$$\begin{aligned} \|f_{n+1}\|_{1,1} = \|f_{n+1} - f_0\|_{1,1} &\leq \|f_{n+1} - f_n\|_{1,1} + \|f_n - f_{n-1}\|_{1,1} + \cdots + \|f_1 - f_0\|_{1,1} \\ &\leq (\delta^n + \delta^{n-1} + \ldots + \delta + 1)\| \mathbf{T}_\theta[f_0] - f_0 \|_{1,1} \\ &\leq \frac{1}{1-\delta} \| \mathbf{T}_\theta[0] \|_{1,1} \\ &\leq \frac{C_2}{1-\delta} \quad \equiv R. \end{aligned} \tag{3.5}$$

We require that the constant C_2 (appearing in A5), satisfy $C_2 < E(1-\delta)$ where δ is defined in equation (3.2). Therefore all the iterates lie in a ball of radius $R < E$ in Lip_1.

Since $\mathbf{T}_\theta[\;]$ is a contraction on S_E, the sequence $\{f_n\}_{n=1}^{\infty}$ converges to a unique theta fixed point in $C^{1,1}$ with $\|f\|_{1,1} \leq R$. This completes the proof of the theorem.

Some remarks on the assumptions should be made.

First on the specification of boundary conditions and the assumption A4. In order for the functions $\psi_\pm$ to achieve their maximum on the boundary $\partial\Omega$ at the ends of the arc $\partial\Omega_2$, for a given forcing function $\gamma(x,y)$ and error bound E, we require that the contribution from the boundary conditions (2.2) – (2.3) overcome the tendency for the forcing functions to create internal maxima or minima.

Second, other than for clarity of exposition, there is no reason to restrict ourselves to the Laplacian in equation (2.1). We could replace this equation by $Lu = \gamma(x,y) + f(u)$ where L is any uniformly elliptic operator with smooth coefficients on Ω. It is probable that extension could be made to the case where the unperturbed problem is quasilinear, by for example taking $\gamma = \gamma(x,y,u,\nabla u)$. We have performed a numerical simulation of the algorithm defined by (3.3) including the case of γ depending on u and ∇u, and found it to be quite robust and rapidly convergent. Some of the numerical implementations for this fixed point method may be found in [8].

Finally the extension to R^n can easily be made. The section of the boundary where the overposed data is specified would still have to be a one dimensional arc on the surface $\partial\Omega$ and contain the entire range of values of $u(\vec{x};f)$.

REFERENCES

[1] J. R. CANNON, *Determination of a unknown heat source from overspecified boundary data,* SIAM J. Numer. Anal., 5 (1968), pp. 275–286.

[2] J. R. CANNON AND P. DUCHATEAU, *An inverse problem for a nonlinear diffusion equation,* SIAM J. Appl. Math., 39 (1980), pp. 272–289.

[3] J. R. CANNON AND P. DUCHATEAU, *Determination of unknown coefficients in parabolic operators from overspecified initial boundary data,* J. Heat Transfer, 100 (1978), pp. 503–507.

[4] J. R. CANNON AND W. RUNDELL, *The determination of a coefficient in an elliptic partial differential equation from overspecified boundary data,* submitted.

[5] P. DUCHATEAU AND W. RUNDELL, *Unicity in an inverse problem for an unknown reaction term in a reaction diffusion equation,* J. Differential Equations, 59, (1985), pp. 155–165.

[6] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations,* Springer–Verlag, Berlin, 1977.

[7] M. PILANT AND W. RUNDELL, *An inverse problem for a nonlinear parabolic equation,* Comm. PDE, 11 (1986), pp. 445–457.

[8] M. PILANT AND W. RUNDELL, *Undetermined coefficient problems for nonlinear elliptic and parabolic equations*, Internat. Ser. Numer. Math., 57 (1986), pp. 139–154.

[9] W. RUNDELL, *Determination of an unknown non-homogeneous term in a linear partial differential equation from overspecified boundary data,* Appl. Anal., 10 (1980), pp. 231–242.